The Dissertation Committee for Kaitlyn Elizabeth Johnson Certifies that this is the approved version of the following Dissertation:

An Integrated Approach to Model Cancer Cell Growth and Treatment Response with Multimodal Data Sources

Committee:

Amy Brock, Supervisor

Thomas E. Yankeelov

Mia Markey

Dean Bottino

An Integrated Approach to Model Cancer Cell Growth and Treatment Response with Multimodal Data Sources

by

Kaitlyn Elizabeth Johnson

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2020

Dedication

To Jojo, for gifting me, amongst many books,

"The Emperor of All Maladies", and along with it

encouraging me to pursue a path

that interests and fulfills me.

Acknowledgements

There are so many people who have supported me on my educational path, both directly and indirectly. If I am ever asked if I believe in luck, the answer will always be yes because I am so very lucky to have encountered each and every person who has inspired me, granted me opportunities, gifted me with resources, and encouraged me along this journey. First, I'd like to thank my advisor, Dr. Amy Brock, for welcoming me into her lab and giving me room to grow and pursue a new avenue of research that I had no experience in (I had never coded before starting graduate school). But above all, I thank Amy for being an excellent advisor to all of her students, for leading by example in the way that she runs both her lab and her family, for allowing graduate students to take ownership of their own independent projects, while also enabling and encouraging us to practice work-life balance and caring about us both as humans and as her graduate students. I'd also like to thank Dr. Thomas Yankeelov, my co-supervisor, for his support, encouragement, and enthusiasm throughout the years as I went from struggling to fit a line to a curve to attempting to contribute to the field of mathematical oncology.

I thank Dr. Dean Bottino for going above and beyond as my manager both during my time as an intern and for the years to follow. From meeting with me nearly daily as an intern, combing through the troubles I was having in my coding and mathematics with patience, diligence, and genuine interest, to answering my technical questions about completely unrelated problems over lengthy emails for years to come, and lastly for providing professional mentorship throughout my journey as a PhD student. My summer

internship in Boston was so pivotal in steering my entire graduate education in so many ways, and I have Dean to continuously thank for that. I also would like to thank Dr. Mia Markey, for encouraging me to pursue an industry internship, as well as providing unbiased external guidance as I struggled through my first year trying to figure out what was expected of me as a first-year graduate student, as well as providing insightful input into my work at committee meetings throughout my PhD.

I also have to give a dedicated and very sincere thank you to Lacy White, our graduate coordinator in the BME department. Since Lacy arrived, I have never once felt alone in this process. She has made it her mission to improve the lives of graduate students. Knowing that I have an advocate and friend who supports me, as well as all the other graduate students, has improved every step of my own graduate life.

As you will see in the coming chapters, many people contributed directly to this work in our both our lab and amongst collaborators. I'd specifically like to thank every graduate student in the Brock lab: Aziz Al'khafaji, Grant Howard, Hunter Joyce, Russ Durrett, Eric Brenner, Daylin Morgan, Andrea Gardner, and Tyler Jost. It is quite obvious that none of the work presented in this dissertation would have been possible without each of them contributing immensely, and I thank them all so much for letting me take their work and try and apply my own ideas and analyses to it. Thank you to the students of the Yankeelov lab as well: Ryan Woodall, Anum Syed, Kalina Slavkova, Chengyue We, Meghan Bloom, and Caleb Phillips for letting me bother them with questions and crash their weekly journal clubs. Thank you to Angela Jarrett, for mentoring me scientifically and always encouraging me. Thank you to all of our collaborators: Eduardo Sontag, James

Greene, Sui Huang, and Michael Strasser. Thank you to all my friends in the department, especially Alex Schonessen and Andrea Trementozzi for sitting through many Tupperware lunches, practice talks, and walks to Dunkin to discuss research ideas, professional opportunities, and the daily stresses of graduate school.

Outside of the research realm, I'd like to thank the many people in Austin who have been the fuel that has kept me going- all of my friends at Rogue Running for consistently setting examples of how to work hard to achieve huge big scary goals. Yaneli Rubio, Alexa Lund, Cate Barrett, Paul Rademacher, Austin Winn, Nicole Ledesma, Katie Watson, Rachel Baptista, Hattie Schunk, and Chris McClung: I don't know how I would've survived without weekly runs with all of you to remind me just how hard working and dedicated you all are in each of your personal, athletic, and professional endeavors, and motivating me to channel just a bit of that in my approach to my own research and professional objectives.

Lastly, to my family- Mom, Dad, Patrick, Brian, Jack, Grandma and Grandpa. For supporting me unconditionally and providing me with a home to come back to, and that will come with me wherever I go. You set no boundaries on what I believed I could achieve from the very beginning, and I am forever grateful for being able to attend Georgetown, where I learned to learn and build confidence in myself and my own decisions, even as they changed course throughout my education. I am so lucky to have had such supportive people with me every step of the way, and I couldn't have done any of this without their encouragement and support.

Abstract

An Integrated Approach to Model Cancer Cell Growth and Treatment Response with Multimodal Data Sources

Kaitlyn Elizabeth Johnson, PhD

The University of Texas at Austin, 2020

Supervisor: Amy Brock

Mathematical modeling and computational biology have been used to understand, describe, and predict critical behaviors of cancer progression. Recent technological advancements in the acquisition of single cell resolution data by high-throughput micrographic imaging and by single cell genomics now enable new analyses of cancer cells at the individual cell and cell population levels. This dissertation focuses on the development of math modeling frameworks capable of integrating and improving our utilization of these novel data types.

First, we investigate the relevance of deviations from the conventional exponential growth model via an ecological principle known as the Allee effect, in which cancer cells exhibit cooperative growth dynamics at low population densities relevant in tumor initiation and metastases. Using a large number of single cell resolution growth trajectories acquired at low cell densities, we apply a stochastic parameter estimation framework to

systematically evaluate the relevance of an Allee effect in a controlled experimental setting. Our findings reveal evidence for cooperative growth even in the presence of optimal space and nutrients, giving us motivation to consider Allee effects in making predictions regarding treatment response and tumor initiation.

The remainder of our work focuses on utilizing multimodal data sources to better understand the dynamics of resistance to chemotherapy. We utilize a mathematical model describing the effects of a treatment-induced resistance on a population of cancer cells and seek to utilize available snapshot and longitudinal data to identify the model parameters. Using lineage tracing technologies developed in the Brock lab, the transcriptomic data set is made actionable by developing a classifier capable of predicting whether a cell in a sample is sensitive or resistant to chemotherapy. We apply this to estimate the composition of the population at a few snapshots in time during treatment response and combine this with longitudinal data directly into our model calibration. The explicit incorporation of molecular level data with population-size dynamics data improves the identifiability and predictive power of the mathematical model. We intend this work to be exemplary of ways in which novel methods can improve the use of data to describe, evaluate, predict, and optimize cancer treatments.

Table of Contents

[1]Chapter 1: Introduction

PREFACE

This dissertation work is focused in two broad areas; intratumoral heterogeneity and mathematical oncology. It has recently been recognized that improving our understanding of the heterogeneity within a cancer population is critical to understanding disease progression and treatment response. Critically, mathematical modeling tools can be used to improve our understanding of the dynamics of heterogeneity, allowing us to uncover how distinct subpopulations of cancer cells might interact and/or transition between states, leading to broader changers in the observed behavior of the entire cancer cell population. In this research, we focus on utilizing experimental data at the population and cellular level into mathematical oncology frameowrks to improve our understanding of the underlying systems.

This introduction chapter is divided into three main parts. We begin by giving an overview of the biological questions driving the investigations to follow; specifically, the

[1] Note: Portions of this chapter are based on a book chapter in press to be published as:
Johnson, K.E., Brenner, E. & Brock, A. (2019). Implications of non-genetic heterogeneity in cancer drug resistance and malignant progression. "Phenotypic switching in biology and medicine". Elsevier Publishing. Book Chapter. In press.
Author contributions:
All authors contributed to conceptualization, writing and reviewing.
As well as based on a commentary originally published as:
Pan, J.* & Johnson, K.E.* (2019). "Hacking" our way across interdisciplinary boundaries. (2019). Cell Systems. 8(5):361-362. https://doi.org/10.1016/j.cels.2019.04.006
*= equal contribution
Author contributions:
Conceptualization: Josh Pan Writing- original draft, review, and editing: Josh Pan, Kaitlyn E. Johnson,
Funding: National Science Foundation's Quantitative Cell Biology Network (NSF MCB-1411898)

role of non-genetic heterogeneity in cancer growth and progression. We then focus on a brief overview of relevant topics in computational oncology. This is by no means exhaustive, but instead focuses only on a brief background of the methods used in this dissertation: bioinformatics in oncology, stochastic models to describe individual cancer cell behavior, and finally, ordinary differential equations for describing interactions and dynamics of cancer populations. Lastly, we close the introduction by describing the challenges, but also the power, of bringing together distinct fields of science to identify and tackle major scientific problems, and how the lessons learned from this experience shaped the future of the work in the remainder of this dissertation.

NON-GENETIC HETEROGENEITY IN CANCER

Non-genetic heterogeneity in cancer plays a critical role in disease progression and response to therapy. While variability in cellular phenotypes results from both gene expression noise and different stable phenotypic states, in this section we will focus on the latter, specifically the theory and evidence for non-genetic heterogeneity in stable phenotypic states in cancer. To elucidate the theory that allows for heterogeneous populations of cells that are independent of their genomic state, we incorporate the concept of the phenotypic landscape—in which cells reside in stable "attractor" states. In this framework, cells have the ability to transition to different states, and the probability of these transitions may be in part dependent on environmental conditions. Mathematical models allow us to build a simplified understanding of the heterogeneous states and the transition rates between states within a cancer cell population. Math models will be used

extensively in this dissertation as mathematical representations of hypotheses of the underlying heterogeneity of cell states in cancer.

After describing the theoretical basis for cell states as "attractors" in a phenotypic landscape, we will discuss ways that cell states are identified and measured experimentally to investigate non-genetic heterogeneity in various empirical settings in cancer. We will describe instances of non-genetic heterogeneity and phenotypic state switching defined by drug naïve cell states with functional relevance to cancer progression and drug response. To make progress in preventing the onset of chemoresistance, it is necessary to elucidate how drug exposure may directly induce cell state transitions between sensitive and resistant cell states. We discuss here the evidence that exposure to cytotoxic or targeted therapeutic treatments may cause cells to activate transcriptional or cell signaling programs that render them insensitive to treatment via a variety of different resistance mechanisms.

We highlight the diversity of operational "cell state" definitions that characterize non-genetic heterogeneity in cancer- from simple functional characterizations of phenotypes defined by growth rates, to high dimensional definitions defined by single cell RNA sequencing of the transcriptome. Both characterizations will be used in research works in this dissertation. Here we argue that no all-encompassing state-space of cancer cells needs to be defined, but rather that cell-states may be defined by the properties that are most relevant to the biological question of interest. We discuss the advantages and shortcomings of single cell RNA sequencing (scRNA-seq), as well as the contexts in which scRNA-seq may be most useful. Lastly, we propose using principles previously applied in the field of differentiation and development to integrate high dimensional scRNA-seq data

with functionally relevant cell states to better understand the role of non-genetic heterogeneity and phenotypic state switching in cancer. This work in differentiation provided the motivation for the project described in Chapter 4 of this dissertation.

Theory of cell states in a phenotypic landscape

Intratumoral heterogeneity is widely recognized as a critical factor in tumor progression, adaptation, and treatment response. Broadly speaking, intratumoral heterogeneity can be defined as the presence of distinct cellular phenotypes within a tumor cell population. The diversity of a cancer cell population can be examined at multiple spatial scales ranging from single nucleotide mutations in the genome to broad functional cellular behaviors such as growth rate or drug sensitivity. Numerous studies have demonstrated that increased heterogeneity is correlated with increased resistance to treatment and poorer patient prognosis (Fillmore and Kuperwasser, 2008b; Brock, Chang and Huang, 2009a; Pisco and Huang, 2015; Maley *et al.*, 2017). While the importance of understanding heterogeneity in cancer is well acknowledged by the field, the adoption of multiple different definitions of heterogeneity can cause complications, potentially preventing a common language between basic discovery and clinical measures (Maley *et al.*, 2017). Heterogeneity may manifest as clonal, or genetic, heterogeneity, in which distinct subclones of cells harbor genetic mutations that can confer phenotypic diversity to daughter cells; or non-genetic heterogeneity, which describes multistability in gene expression dynamics whereby one genome produces multiple stable or metastable phenotypic states. Observations of individual cells able to reversibly transition into different phenotypic states either spontaneously (Piyush B Gupta *et al.*, 2011; Thanos *et*

al., 2018) or due to an environmental stimulus (Pisco and Huang, 2015; Keisha N Hardeman *et al.*, 2017; S. Chen *et al.*, 2018) imply that a cell does not need to harbor a permanent genetic mutation to exhibit multiple cellular phenotypes.

The presence of distinct subpopulations of cells able to transition between cell states to form a heterogeneous cell population is commonly described using the analogy of a phenotypic landscape. The concept was first introduced by Waddington to model differentiation and development (Waddington, 1940). Stem cells occupy the top level of the landscape and as cells differentiate, they descend into valleys and assume stable discrete phenotypes represented by "basins"; these are defined by their characteristic gene expression profiles and the resulting phenotype (Fig 1.1).

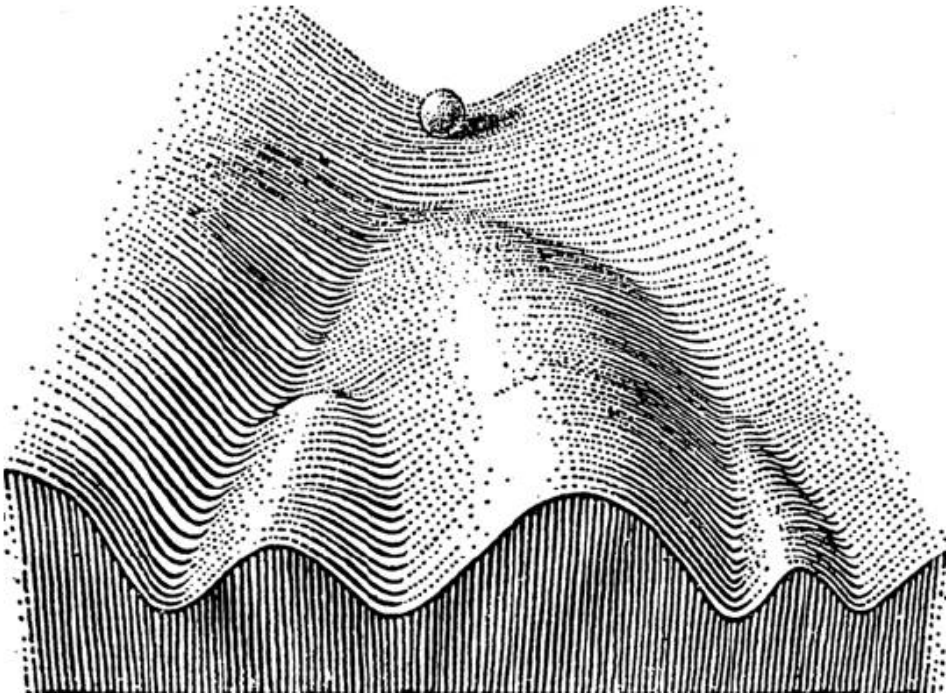

Figure 1.1: Waddington landscape concept of stem cell differentiation
Waddington first posited that stem cells traverse down buffered pathways. Up to a certain threshold neither external or internal perturbation affects the pathway, and transitions into an adjacent developmental pathway are rare.

In this framework, cells are more likely to equilibrate into stable states represented by a 'potential well' in the landscape, but reverse transitions back towards stemness are theoretically possible. The probability of transitioning between states is directly proportional to the energy barrier between states, i.e. the height of the basins. This framework has recently been extended to understanding cancer cell fates (Brock, Chang and Huang, 2009a; Huang, 2013; Paudel *et al.*, 2018). In cancer, a clear hierarchy of cell types is not generally believed to exist, but instead multiple metastable phenotypic states can coexist. This is consistent with the observations of non-genetic heterogeneity in cancer, in which the landscape represents all theoretically possible physiological cell states. A

population of cancer cells may spread out across these available phenotypes, and subpopulation compositions of cancer cell types represent a quasi-equilibrium of the landscape. The effect of a perturbation on this landscape can come in the form of a drug treatment, where in theory a treatment can have the effect of temporarily altering the topography of the landscape, resulting in temporary changes in phenotypic composition followed by a return to initial proportions. Alternatively, a perturbation can act to permanently alter the landscape by changing the relative depth of the wells that represent cell states, resulting in re-equilibration followed by stable changes in phenotypic composition of a cell population (Brock, Chang and Huang, 2009a; Zhou *et al.*, 2014; Li, Wennborg, Aurell, Dekel, J.-Z. Zou, *et al.*, 2016). These stable changes could be achieved through either permanent environmental stimuli, or mutations and epigenetic alterations in the cancer cells themselves that may make available cell states that were previously inaccessible or change the stability of existing states relative to one another.

While the phenotypic landscape concept may serve as an analogy rather than a true biological phenomenon to be exploited and tested, its utility lies in the framework it provides for understanding complex subpopulation dynamics. Heterogeneity in cancer cells not only encompasses the presence of distinct subpopulations in different frequencies within one tumor, but also that these subpopulations exhibit temporal variation. Under the landscape framework, one can model changing subpopulation compositions in cancer cell populations by defining cells with distinct cell states and tracking those cells as they grow, are killed by treatment, or transition into other states either stochastically or induced by a stimulus. Simple mathematical models can be developed under the guiding principle of

cell states as attractor states, and can be used to describe and predict how subpopulations of cancer cells change over time based on the topography of the landscape (Pisco and Huang, 2015). Studying the effect of perturbation through these modeling methods can provide evidence for the critical role of cell state transitions, rather than differential growth rates, as the driver of observed subpopulation levels in time (Piyush B. Gupta *et al.*, 2011; Pisco and Huang, 2015). These models can help predict the implications of different environmental stimuli and treatment strategies on the population composition over time. Models of phenotypic state switching can be employed to help us understand the dynamic processes related to cancer progression and treatment response by integrating theoretical models with experimental quantification of levels of relevant subpopulations over time.

Experimental evidence for non-genetic heterogeneity

What experimental findings support the functional importance of non-genetic heterogeneous cell states in cancer? To begin, in order to measure heterogeneity, one must first define the unit that is being measured. It turns out this is not quite so simple. While the field has sought to develop a universal definition of diversity, cell states, and heterogeneity (Maley *et al.*, 2017), we argue here that cell states should be defined not universally, but in the context of the relevant biological question. For example, this means if we are interested in examining heterogeneity in response to treatment, then heterogeneity in cell cycle phase may be less relevant, and may not need to be incorporated into the model of heterogeneity. In this overview we discuss heterogeneity defined at a breadth of levels—from molecular quantification of gene expression levels in a single cell to observations of diversity in cell morphology, functional behavior, and behavior related to proliferation

rates, metastatic potential, and responsiveness to therapies. A few key cell state definitions used as critical evidence for non-genetic heterogeneity in cancer are described in Table 1.1. Here, we only consider observations of stable changes in cellular phenotypes as cell states, as opposed to the quantification of cell states defined by rapid stochastic fluctuations in gene expression levels. In an analogy to thermodynamics, we are interested in the ability of cells to transition from one macrostate to another, rather than focusing on transitions at the level of microstates. Even at this higher level, the definition of a state can vary widely based on context. For example, cell states can be classified by their molecular characterizations (Piyush B. Gupta *et al.*, 2011), where cells are categorized as basal-like, luminal-like, and stem-like based on their surface markers that are commonly used to characterize these distinct cell types. Other times, cell states may be classified by their proliferative capacity (Paudel *et al.*, 2018), in which cells are assigned states based on doubling time, or based on sensitivity to drug (Howard *et al.*, 2018). In some studies, cell states may be more completely characterized by the entire gene expression state using single cell RNA sequencing and dimensionality reduction (Patel, 2014; Mojtahedi *et al.*, 2016; Stumpf *et al.*, 2017). While none of these methods of categorizing cell states can perfectly capture the most parsimonious definition and binning of cell states, a diverse set of experimental measures may further our understanding of the dynamic composition of cancer cell populations.

Cell State Definition	Functional Characterization	Molecular Characterization	Source
Epithelial/ Mesenchymal	Epithelial: cell-to-cell adhesion, less mobile, polygonal and cobble stone-like, apical-basal polarity Mesenchymal: lack of cell-to-cell adhesions, elongated and spindle-like invasive, mobile, front-back polarity, elevated resistance to apoptosis	Epithelial: high E-cadherin, low vimentin, cytokeratins Mesenchymal: high N-cadherin, high vimentin, production of ECM degrading enzymes, FSP1, desmin Transcription factors associated with transition to mesenchymal: Snail. Twist, Zeb, FOXC2, and Yap families	(Elosegui-artola *et al.*, 2017), (Wei *et al.*, 2015), (Yu *et al.*, 2015), (Ren *et al.*, 2016), (Ai *et al.*, 2014), (Chaffer *et al.*, 2016), (Ye and Weinberg, 2015), (Kalluri *et al.*, 2010), (Mani *et al.*, 2008), (Brabletz *et al.*, 2018)
Drug Sensitive/ Resistant	Sensitive: higher growth rate, quicker death in response to drug, Resistant: slower growth rate, slower/less death in presence and after drug exposure, mechanisms include: evading targeted treatment, multi-drug resistance pump, persistence, changes in morphology	Sensitive: high Ki67 (proliferation marker), mutations dependent (i.e. high expression of BRAF /Kras/GPX4 others Resistant: MDR1 expression high, low expression of Kras, BRAF, upregulated DNA repair	(Pisco and Huang, 2015) (P. Chen *et al.*, 2018), (Howard *et al.*, 2018), (Hangauer *et al.*, 2017)
OxPhos/ Glycolysis Metabolism	Ox Phos: oxidative phosphorylation as primary mechanism of ATP production, use oxygen, sensitive to Ox Phos inhibition Glycolysis: glucose used to produce ATP	OxPhos: up regulation of PCG1 and transcription factor MITF Glycolysis: up regulation of RAS-RAF-MEK-ERK signaling axis, glutamine transporter ASCT2	(Keisha N. Hardeman *et al.*, 2017). (Deberardinis *et al.*, 2008; Deberardinis and Chandel, 2016), (Heiden *et al.*, 2009), (Davies *et al.*, 2015), (Vazquez *et al.*, 2013)
Hypoxic/ Vascular	Hypoxic: de-oxygenated cell state Vascular: well-oxygenated and nutrient rich environment, characterized by presence of blood vessels in 3D tumor environment	Hypoxia: H& E high Vascular: CD31 high	(Syed *et al.*, 2019), (Sorace *et al.*, 2017), (Junttila and Sauvage, 2013), (Goel *et al.*, 2011)
Carcinomatous/ Sarcomatoid	Carcinomatous: round, less invasive, better cell-cell junctions Sarcomatoid: spindle-like, more aggressive, invasive/migration,	Carcinomatous: k7 positive, upregulated in cell-cell junction genes and epithelial related genes Sarcomatoid: k7 negative, upregulated EMT related genes (TWIST1, TGFB, ZEB1), upregulated stem cell genes,	(Thanos *et al.*, 2018), (Wang, Cui and Weng, 2012), (Miettinen *et al.*, 1999)
Basal/ luminal/ stem	Basal: Luminal: Stem: ability to generate other phenotypes, self-renewal	Basal: CD44 high, CD24 neg, EpCam neg Luminal: CD44 low, CD24 high, EpCam high Stem: CD44 high, CD24 neg, EpCam Lo	(Piyush B. Gupta *et al.*, 2011), (Fillmore and Kuperwasser, 2008b), (Shipitsin *et al.*, 2007)

Table 1.1: Examples of different categories of cancer cell state definitions used in the context of heterogeneous cancer cell populations in the literature, identified by function and molecular characterization.

Observations of drug-naïve cell states

There exists an abundance of evidence for the presence of distinct subpopulations of cells within a population, even in the absence of environmental stimuli that might drive phenotypic adaptation. The existence of phenotypic diversity does not necessitate the presence of non-genetic heterogeneity, as observed phenotypic composition could be obtained through differential growth rate of distinct subclones with different genomes that give rise to different phenotypes. Thus, this often leads to the question, are the observed proportions of subpopulations of cells maintained through differential proliferation rates of distinct subtypes or the interconversion between different cell states to maintain equilibrium proportions? While it is likely true that in cancer, both differential growth rates of distinct subclones and phenotypic plasticity both contribute to subpopulation composition, there exists an abundance of evidence for the idea that subpopulation proportions are maintained through phenotypic transitions between distinct cell states.

Recent studies have demonstrated that phenotypic transitions are the most likely mechanism for maintaining equilibrium proportions of cell states within a cancer cell line that might otherwise be considered homogenous (Piyush B. Gupta *et al.*, 2011). Researchers quantified the baseline proportions of distinct cell-states in SUM159 and SUM149 breast cancer cell lines using previously defined and characterized cell-surface markers corresponding to phenotypes of: stem-like (CD44 high, CD24 negative, EpCam low), basal (CD44 high, CD24 negative, EpCam neg) and luminal (CD44 low, CD24 high EpCam high) (Shipitsin *et al.*, 2007; Fillmore and Kuperwasser, 2008a). Using fluorescence activated cell sorting (FACS), the baseline proportions of cell phenotypes,

defined by the cell surface marker levels states above, were measured in each cell line (Piyush B. Gupta *et al.*, 2011). The cell-surface markers and FACS were used to isolate pure subpopulations, and the resulting proportions of cells in each cell state were sampled over time following isolation (Piyush B. Gupta *et al.*, 2011). The results revealed a rapid progression back toward the equilibrium proportions of the original cell line, and based on the short time it took to recapitulate initial proportions, cell-state transitions were more likely to achieve the observed proportions than differential proliferation rates, demonstrating the role of phenotypic state switching in maintaining the observed non-genetic heterogeneity in cancer cell lines (Piyush B. Gupta *et al.*, 2011).

In the phenotypic landscape model (Huang, 2011), cells may overcome the stability of their gene expression configuration to exit an attractor state in one direction, but the mechanisms by which this occurs are not well characterized. The question of mechanism of "escape" is addressed in blood cells by perturbing differentiated erythroid (red) and myeloid (white) blood cell lineages and observing the dynamic response to perturbation by quantifying cell state composition over time (Mojtahedi *et al.*, 2016). Stimulation with cytokines of erythroprotein and/or IL-3/GM-CSF trigger this transition. The resulting changes in subpopulation composition over time were measured at the single cell resolution using both FACS sorting on surface protein expression markers and single cell qPCR analysis of 19 genes. Transition was characterized by a transition period at day 3 in which cells exhibited a higher diversity in cell state space, consistent with the analogy of the landscape temporarily flattening. This was followed by coalescence into two distinct cell clusters, corresponding to committed erythroid (red) and myeloid (white) cell states. These

empirical data relating changes in phenotypic compositions over time provide the basis for a mathematical modeling framework to describe the probability of transitioning from one attractor state to another as functions of the relative depth of the well (Mojtahedi *et al.*, 2016). These modeling frameworks allow us to better understand how subpopulation dynamics are maintained through phenotypic plasticity (Piyush B. Gupta *et al.*, 2011; Mojtahedi *et al.*, 2016).

Since these landmark studies describing the dynamics of cell state transitions, additional work has demonstrated the important role of drug-naïve phenotypic state switching for different definitions of cell state. For example, in human liver cancer, heterogeneity in histopathology has demonstrated that sarcomatoid cholangiocarcinoma is characterized by distinct stable phenotypes classified as "sarcomatoid" and "carcinomatous" cells (Miettinen *et al.*, 1999; Wang *et al.*, 2016; Thanos *et al.*, 2018). The sarcomatoid cell type is characterized morphologically as spindle-shaped and functionally are known to be more invasive and motile, similar to mesenchymal cells. Transcriptional profiles of the two cell types indicates that sarcomatoid cells are down regulated in cell-to-cell junction related genes and are upregulated in invasion/migration related genes, epithelial-to-mesenchymal(EMT) related genes such as TWIST1, TGFB, and ZEB1, and stem cell genes, compared to the carcinomatous cells. While morphology, invasivity, motility, and molecular characterizations were used to "define" distinct subtypes of cells in this context, a single marker, keratin-7 expression, was used to isolate cell phenotypes via antibody staining (Thanos *et al.*, 2018), and revealed the presence of a single "mixed" cell type able to transition into either the k7 high or k7 low cell type. This work

demonstrated that phenotypic plasticity plays a significant role in maintaining heterogeneous subpopulations of cancer cells even in the absence of any environmental stimulus.

Additional instances of drug-naive phenotypic state switching have been described and characterized in cancer in a number of ways. For example, in the field of EMT, stochastic cell state transitions between epithelial and mesenchymal cells have been observed in the absence of any transition driving perturbation (Tian, Zhang and Xing, 2013). Additionally, cell states have been characterized functionally by the observed growth rate, with changes in the observed population growth rate over time and passage number explained due to changing subpopulations of cells in different phenotypic states (Paudel *et al.*, 2018). Thus, even in the absence of treatment or severe environmental pressures, cancer cells exhibit the capacity to overcome energy barriers between "basins" in the landscape to convert with some non-zero probability into alternative cell states and explore state-space. Although normal healthy cells likely exhibit these capabilities as well, for example in induced pluripotency, it is possible that cancer cells exist in a "flatter" landscape (Li, Wennborg, Aurell, Dekel, J. Zou, *et al.*, 2016; Mojtahedi *et al.*, 2016), resulting in non-genetic heterogeneity and the ability to evade environmental pressures.

Adaptive cell states that are induced in response to cancer treatment

The development of drug resistance in cancer is often explained by the presence of rare genetic mutations in the tumor population that allow for resistant subclones to expand in the presence of the selective pressure of treatment via Darwinian selection. However, recently there has been an increased interest in an alternative mechanism, in which the

treatment itself induces an altered, drug resistant or tolerant, phenotypic state (Pisco and Huang, 2015; Greene and Gevertz, 2017). The ability of cells within a population to transition from a drug-sensitive to a drug resistant state is demonstrated in a number of cancer types and in response to both targeted and cytotoxic therapies in cancer (Brock, Chang and Huang, 2009a; Zhou *et al.*, 2014; Pisco and Huang, 2015; Greene and Gevertz, 2017; Hardeman *et al.*, 2017). In this section, we will highlight the evidence for drug-induced phenotypic switching in cancer and overview the many ways in which these altered phenotypic states have been characterized in the context of drug resistance.

Targeted therapies are designed to specifically target and kill or inhibit the growth of cancer cells exhibiting a characteristic not present in high abundance on normal healthy cells. This could include anything from cell surface receptors often over-expressed in cancer to oncogene addictions in certain cancer types. Some examples of these targets for which known therapies have been developed include: Kras addicted pancreatic ductal adenocarcinoma (PDAC) (P. Chen *et al.*, 2018), HER2+ breast cancer (Hangauer *et al.*, 2017), EGFR positive non-small cell lung cancer, and BRAF mutated melanoma (Keisha N Hardeman *et al.*, 2017; Paudel *et al.*, 2018). The promise of targeted therapies is based on the assumption that cancer cells are dependent on activation of the specific target, and variants with pre-existing resistance to that target are rare. However, a number of studies of targeted drug treatments on "oncogene addicted" cancers have demonstrated that exposure to targeted treatment can enable cells to adapt to an alternative, potentially reversible, cell state. For example, in Kras addicted PDAC, Kras inhibition treatment results in the induction of a drug-tolerant cell state characterized by differences in cell

morphology, proliferative kinetics, and tumor-initiating capacity (P. Chen *et al.*, 2018). This drug-induced tolerant state is demonstrated to be reversible, resulting in no significant mutational or transcriptional changes but changes in gene expression related to cell signaling and focal adhesion pathways that cause the cells to have an increased dependence on adhesion for viability *in vitro* (P. Chen *et al.*, 2018). Similarly, in HER2+ breast cancer, lapatinib treatment induces a non-mutational "persister" cell state that is characterized by a transient dependency on GPX4 (Hangauer *et al.*, 2017). In this work, the ability to capitalize on non-genetic heterogeneity and phenotypic plasticity is demonstrated by attacking the drug-induced GPX4 dependent state with a GPX4 inhibitor, resulting in cell death *in vitro* that is not observed on the parental cell line alone with GPX4 inhibitor (Hangauer *et al.*, 2017).

In addition to targeting drug-induced states based on their molecular dependencies, recent work has also shed light on characterizing the drug-induced state via broader phenotypic changes. For example, BRAF mutated melanoma cells states have been defined in terms of their tumor metabolic phenotype (Keisha N Hardeman *et al.*, 2017). In this work, it is observed that cells fall along a spectrum of sensitivity to BRAF inhibition, with cell states resistant to BRAF inhibition characterized by a metabolic phenotype of oxidative phosphorylation instead of glycolysis. Thus, to overcome resistance to BRAF inhibition treatment, melanoma cells were treated with zalcitabine, a drug that suppresses normal oxidative phosphorylation and forces cells into glycolysis,. They found that cells could be resensitized to BRAF inhibition, thus demonstrating the reversibility of the drug resistance via driving the alternative metabolic phenotype (Keisha N Hardeman *et al.*, 2017). In this

context, knowing both the phenotype of the BRAF sensitivity and the metabolic phenotype allowed researchers to probe whether or not the phenotypic switching observed in response to drug was directly linked to an alternate phenotype. These types of relationships to characterize drug-induced phenotypes can be useful in developing targeted treatment strategies aimed at rationally modifying the landscape in favor of cell states with greater drug susceptibility.

The idea that cancer cells might evade attack from targeted treatment by utilizing alternative pathways for survival is quite rational, however, how does a specific multidrug resistant phenotype emerge in response to broad-based chemotherapeutic agents? Again, observed drug resistance is classically explained by selection of resistant mutant cancer cells, however recent work has demonstrated that a multi-drug resistant state, characterized by expression of the MDR1 drug-pumping family of genes, is directly induced via "Lamarckian" induction, following exposure to chemotherapy in HL60 leukemic cells (Pisco and Huang, 2015). Lamarckian induction is induction driven by epigenetic changes and transcriptional plasticity that can contribute to drug resistance in addition to "Darwinian" natural selection. This landmark paper not only demonstrated a drug-induced phenotypic state characterized by the functional ability to efflux drug, but also that this drug-induced phenotype was a result of cell-autonomous gene induction that was independent of fitness benefit, as the elevated levels of expression of MDR1 proteins were still observed even in the inhibition of the functional drug pumping mechanism. These findings introduced the idea that drug exposure can "instruct" a cell to switch between attractor states in a directed manner. Not only does this finding indicate that drug-induced

attractor states should be examined to identify novel targets of overcoming resistance, but also implies that drug treatment schedules must take resistance induction dynamics into consideration when developing optimal treatment strategies. This concept has been explored by a number of recent theoretical works (Gatenby, 1991; Gatenby *et al.*, 2009; Wood *et al.*, 2012; Greene and Gevertz, 2017). While understanding the dynamics of drug resistant phenotypes over time is important, it is equally important to integrate this knowledge into a mechanistic understanding of the biological process that drives drug-induced resistance. While the drug-induced resistant state has been characterized by a specific resistance mechanism, the MDR1 high state (Pisco and Huang, 2015), it is quite likely that multi-drug resistance is due to multiple resistance mechanisms. To understand the processes driving resistance induction, it is sometimes necessary to identify and quantify non-genetic heterogeneity through much higher throughput molecular and physical measurements.

BRIEF OVERVIEW OF RELEVANT TOPICS IN COMPUTATIONAL ONCOLOGY

In this section, we will give a brief literature review of the pivotal works in three areas of computational oncology that are contributed to and or utilized in this dissertation work: bioinformatics for "omics" data analysis, stochastic modeling of individual cell behavior, and finally, ordinary differential equations for modeling tumor dynamics. This review is by no means comprehensive, and we refer the reader to various reviews and commentary pieces in mathematical oncology for further reading (Gatenby and Maini, 2003; Byrne, 2010; Gallasch *et al.*, 2013; Enderling and Chaplain, 2014; Yankeelov *et al.*,

2016; Anderson and Maini, 2018). Mathematical and computational modeling are powerful tools to test biological hypothesis, confirm experimental observations, and simulate dynamics of complex systems. For this reason, they have the power to improve our understanding and decision-making process in oncology by enabling the simulation of a number of different scenarios, all of which would be infeasible or unethical to test in a laboratory or clinical setting.

Mathematical modeling translates qualitative hypothesis and observations, such as those described above regarding intratumoral heterogeneity, into quantitative models directly comparable with experimental data. In this dissertation, mathematical models will be used to reveal underlying compositions of heterogeneous populations, investigate the relevance of ecological principles in early-stage tumor growth, and to learn from multimodal datasets to develop mechanistic models capable of optimizing treatment regimens. To place this dissertation in context and provide background to readers, we will review key previous work in the field.

Bioinformatics for interpreting "omics" data sets

We begin with an overview of recent advancements in bioinformatics as it is applied to genomic and transcriptomic data sets in cancer. Recent technological advancements such as Next-Generation sequencing (Behjati and Tarpey, 2013) and scRNA-seq have enabled major advancements in the breadth and depth of our understanding of the genetic and expression signatures of individual and populations of cells. However, without advancements in the computational tools used to analyze these data sets, making sense of them, and in particular using them to answer a relevant biological

question, would be quite difficult with traditional statistical methods alone. Although advancements in informatics approaches are not major contribution of this dissertation work, they were an essential component for making use of the scRNA-seq data sets used in Chapter 4. For a more detailed overview of the field of bioinformatics, we refer the reader to comprehensive review papers in bioinformatics (Diniz and Canduri, 2017), and for a more comprehensive overview of the potential for machine learning to permeate medicine, we refer the reader to a recent book by researcher Eric Topol (Topol, 2019).

Advances in high-throughput sequencing now enable biological data to contain an unprecedented level of information, requiring novel approaches and methodologies to give biological meaning to the data generated. This new data has largely led to the development of the new fields of bioinformatics and computational biology which have an integrated interface with molecular biology. These fields truly began with the publication of the structure of DNA by Watson and Crick in 1953 (Crick and Watson, 1953), but were significantly advanced by computing power allowed for sequencing, annotation, processing and analysis of genomic data (Verli, 2014).

With these technologies came new computational methods for making sense of this data. For example, one critical step in analysis of biological sequences, whether they be derived from DNA or RNA, is alignment of sequences for comparison and quantification. Alignment methods such as Needleman-Wunsch and BLAST (Prosdocimi *et al.*, 2002) are two examples of local and global alignment programs, respectively, that enable the quantification of specific pieces of genomic information for use in downstream interpretation. For example, in scRNA-seq pipelines, alignment is used to map reverse-

transcribed cDNA (derived from RNA) to reference genomes where each individual sequence is compared to many possible genes. Alignment is also used in these pipelines to "cluster" pieces of cDNA with their unique molecular identifiers and their cell barcodes, both pieces of information which are used to quantify the abundance of a large number of individual genes in individual cells (Klein *et al.*, 2015). These advances and others, often referred to as preprocessing, are often followed by normalization, statistical methods used to allow for comparison across genes, cells, and samples (Diniz and Canduri, 2017). Advances in these techniques are critical for enabling downstream analysis via proper quantification of the quantities of interest, for example barcode abundance, mutational profiles, or gene expression levels.

The aforementioned methods provide a critical first step to getting raw sequencing read data into an interpretable format. Even after these processing steps however, there is still often a huge breadth of high dimensional, high-throughput data. Interpreting this data and recognizing features or patterns that might be relevant to biology, is just one place where machine learning methods can enter the biological arena. Machine learning algorithms can learn patterns in data for discovering structure in unlabeled data to simplify via dimensionality reduction or organize data via clustering methods (Kann *et al.*, 2019). These approaches can be used for a wide range of applications, such as visualization of clusters of phenotypes, predicting future expression, and identifying common mutational profiles, just to name a few. As these types of data become more common in both research and clinical settings, it is likely that machine learning algorithms will be applied to assist

in making these data types actionable to researchers, physicians, health care systems, and ultimately patients (Topol, 2019).

Stochastic models to describe individual cell birth, death, and interactions

Much as the field of cancer modeling began with descriptions of continuous and deterministic systems, the field of modeling chemical reactions also began this way and was extended to describe the time evolution of chemically reacting systems via discrete, probabilistic molecular events (Gillespie, 1977). Most relevant to the field of subcellular and cellular interactions are the contributions of Daniel T. Gillespie, who developed a foundational algorithm for stochastic simulation that is still used to this day (Gillespie, 1977, 2014). Mathematical oncologists have more recently realized the need to develop discrete models of cancer cells in order to account for the behavior of individual cells (Enderling and Chaplain, 2014) that drive behavior of metastatic spread and early-stage tumor growth. The first to develop a model that explicitly accounts for the behavior of individual cells was (Anderson *et al.*, 2000) where he modeled how individual cells could migrate beyond a margin of cancerous tissue that was visible by surgeons, predicting further penetration into healthy tissue than a continuum model would have predicted. Since this work, a number of other discrete models have been developed using a variety of techniques such as the Pott's model (Turner and Sherratt, 2002; Poplawski *et al.*, 2010), cellular automata (Rocha *et al.*, 2018), agent-based models (Kansal *et al.*, 2000; Zhang *et al.*, 2009; Araujo *et al.*, 2018), and multiscale models which combine continuum and discrete modeling in the relevant regimes (Ramis-Conde *et al.*, 2008; Zhang *et al.*, 2009). These stochastic models have been used to describe intracellular interactions (i.e. the

observed stochasticity in gene expression due to chemical reactions within cells), and to describe the individual interactions of cells with themselves and the environment.

Stochastic models of cell-level behavior have been successfully used to describe and explain observations of cancer cells, such as the non-constant time between cell-divisions and deaths (Stukalin *et al.*, 2013; S. X. Sun, 2015). A number of stochastic models to describe cancer cell birth and death events have been used to predict population dynamics. These models typically allow for the exploration of mechanistic hypotheses regarding the probabilities of birth and death events and how they depend on the presence of other species or environmental factors (Nowak, 2006; S. X. Sun, 2015). For example, (West *et al.*, 2016) developed a stochastic model of tumor growth that uses a Moran birth-death process, which describes how heterogeneity increases over time due to molecular mutations in independent cells (West *et al.*, 2016). Additionally, (West and Newton, 2018) also showed that stochastic models of individual cell-to-cell interactions describe a number of the most commonly observed continuous models of growth behavior, depending on the functional nature of the interactions between individual cells, demonstrating potential mechanistic underpinnings of observed phenomenological behavior. These works demonstrate the power that stochastic modeling and simulations has to recapitulate observed experimental behaviors of cancer cells in a wide variety of settings.

Stochastic models have been used extensively to simulate and ultimately explain expected behaviors for different scenarios. For example the observed behavior of the "go-or-grow" phenomena which leads to an emergence of a slowing of growth at low cell densities (a phenomena known as an Allee effect) (Böttger, Hatzikirou and Voss-böhme,

2015). However, many of these models are extremely computationally expensive, requiring stochastic simulation algorithms (Cao and Petzold, 2006) for each forward function evaluation. In the past, this had necessarily limited their capacity to be calibrated to experimental data. However, with novel technological advances now enabling more precise capturing of experimental data (for example single-cell resolution fluorescence activated cell sorting [FACS] measurements), the ability to calibrate these models has become more tractable using moment-closure approximations. The use of moment-closure approximations for parameter estimation from data was introduced in (Fröhlich *et al.*, 2016), and applied to reveal sources of heterogeneity in FACS sorting data and drug response data (Frohlich *et al.*, 2018; Loos *et al.*, 2018). This work is a key contribution to integrate stochastic modeling into data analysis. This method will be applied to single-cell resolution tumor cell growth data in Chapter 3 of this dissertation, where we investigate the mostly likely structure of the observed growth behavior of small, initiating populations of cancer cells.

Ordinary differential equations for describing the interactions and dynamics of the tumor and its components

The field of mathematical oncology dates back to the 1960s and emerged out of a practical problem- how best to dose newly developed chemotherapeutic agents. These models intended to describe the number of tumor cells as a function of time, in order to predict and compare the effects of treatment on cancer progression. However, describing the number of tumor cells in time is remarkably challenging, so instead researchers turned to a slightly more straightforward question of how tumor cell number changes in time,

resulting in differential equation models that describe the the possibilities of birth, death, and quiescence (Enderling and Chaplain, 2014). The first models were mostly descriptive, rather than mechanistic, and intended to reproduce the gross behavior of the tumor size over time (Anderson and Maini, 2018). Since then, the role of ordinary differential equations (ODEs) have evolved to both answer clinical questions regarding cancer growth and treatment as well as to improve our understanding of the underlying complexities of cancer biology.

The first major contribution to mathematical oncology came in the form of an ODE describing the total tumor cell number in time. In the 1960s, physician Howard Skipper performed a series of experiments in leukemia that demonstrated what is still largely used today- the log-kill hypothesis (Skipper, 1964). This model posits that the number of cells killed by a treatment is directly proportional to the number of cells present, i.e. a constant fraction of cells is killed with each treatment. This finding in turn led to the largely still pervasive idea that, if the goal is to seek a curative treatment, the maximum tolerated dose is the best chance at eradicating all of the tumor cells.

Building upon this work, physician Larry Norton later showed that a non-constant growth rate, in which only a fraction of the population of cells is in a proliferative state, explains the observed growth dynamics in solid tumors (such as breast cancer), in which a slowing of growth rate as the tumor gets larger is observed (Norton, 1988). The Norton-Simon model, as the name implies, merged these two ideas to propose that the fraction of tumor cells killed by a treatment is not just proportional to the number of cells present, but the number of proliferative cells present in the tumor, which in Norton's model is non-

constant (i.e. Gompertzian, in which the number of tumor cells over time gradually slows (Winsor, 1932; Norton, 1988). As a result of this finding, the first clinical trial designed based on a mathematical model of treatment response was tested (Citron *et al.*, 2003). The findings revealed that, as the model suggested, a dose-dense scheme is more effective than a conventional dosing regimen.

Additional examples of phenomenological differential equations used to make treatment decisions can also be found in modeling radiation therapy. In radiation therapy, two types of dynamics of cell death are observed to occur, those that cause near immediate cell death proportional to the dose, and cell death that is delayed and occurs as cells attempt to pass through the cell cycle and undergo mitotic catastrophe due to the DNA damaged induced via radiation (Brenner, 2008). In order to account for these observed dynamics in treatment response, a linear-quadratic model (Brenner, 2008) was proposed which is able to describe the observed changes in cell number via two parameters for the two rates of cell death. Both the linear-quadratic model in radio therapy, and the Norton-Simon model of chemotherapy response represent ways in which descriptive models, describing only population dynamics, have been able to have a significant clinical impact, improving how treatments are administered throughout the field.

However, ODEs have not been limited to phenomenological models describing population dynamics of tumors. Instead, ODEs have become critical for bridging the gaps between biological hypotheses of process at all different scales- from the intercellular level to describe gene network interactions (Rohrs, Makaryan and Finley, 2018) to models of heterogeneous subpopulations related to processes such as epithelial-to-mesenchymal

transitions, immunological responses (Jarrett, Bloom, *et al.*, 2018; Poleszczuk and Enderling, 2018), and drug-sensitivity states (Greene, Sanchez-Tapia and Sontag, 2018b; Greene, Gevertz and Sontag, 2019). The past few decades have seen an abundance of ODE models, built with biological hypothesis about the nature of component interactions, and able to be compared directly with experimental data. These models have a unique place in the field, as they represent one way of bridging the field of cancer systems biology, which seeks to make sense of individual interactions and components and their effects on the cell, tissue, or organ-level, with mathematical modeling. These mechanistic models can thus represent quantitatively different biological hypothesis and can be used when compared with experimental data to test these hypotheses and drive future experiments to validate those hypotheses.

There are a number of significant contributions of ODE-based models to the field of mathematical oncology at all different levels. Of particular relevance to this dissertation for understanding the role of tumor heterogeneity in cancer progression, are models of heterogeneous subpopulations defined by their drug sensitivity states. James Greene and Eduardo Sontag (Greene, Gevertz and Sontag, 2019) propose a model of drug-induced resistance which describes a population of cancer cells made up of resistant and sensitive cells. These two subpopulations grow according to logistic growth at independent rates, have different death rates consistent with the log-kill hypothesis described above, and are able to transition from one state to the other. This mathematical description of drug-induced resistance is described by a transition rate, proportional to the observed treatment, at which sensitive cells can transition into the resistant cell state. The structure of this model was

built based on a number of experimental observations describing the direct induction of a resistant phenotype in response to drug, as is described above (Pisco *et al.*, 2013; Fallahi-sichani *et al.*, 2017). This model not only provides a mathematical framework for testing this hypothesis of drug-induced resistance, but also has utility in optimizing treatment regimens. Just as the log-kill hypothesis indicated that a maximum-tolerated dose would be an optimal treatment strategy, the degree at which a treatment induces resistance changes optimal strategies for drug dosing. For example, Green and Sontag (Greene, Sanchez-Tapia and Sontag, 2018b) show that if a treatment does not induce resistance, then constant treatment improves overall tumor control, however if treatment-induced resistance is present, then pulsed treatment that includes a drug "holiday" allows for regrowth of sensitive cells and improves overall treatment response.

In conclusion, ODE models provide an excellent framework for investigating a number of questions in mathematical oncology and systems biology, and their utility often lies in their ability to be directly compared to experimental data. While PDEs and stochastic models require more extensive data collection to reproduce forward model simulations, the outputs of ODEs are typically single variable described over time, which we can often capture experimentally and clinically (albeit not always at the level of detail described by the mechanistic model). Because of their ability to extend to new mechanistic insight gained by biologists, as well as their ability to be calibrated and validated to data, ODEs are a promising path forward for improving our understanding of underlying cancer biology as well as answering practical question regarding treatment optimization. In this dissertation, we will attempt to bring together the insight from experimental biology to

build mechanistic models, calibrated to experimental data, to answer questions regarding underlying population composition dynamics, cooperative growth, and drug-resistance dynamics.

BRINGING TOGETHER MATHEMATICIANS, BIOLOGISTS, AND EVERYONE IN BETWEEN TO BREAK DOWN INTERDISCIPLINARY BOUNDARIES

This dissertation represents an exercise in integrated team science. The three main projects described here were unique in the ways in which each different discipline contributed, but what weaves through them all is the ways in which the expertise from experimental biologists, bioinformatics, technology developers, and computational modelers each contributed significantly to the works. Throughout each progress, we experience both barriers to entry in terms of entering into a collaborative project, as well as discovering the constant need for dialogue between all involved parties to break down these barriers and put the integrative goals into practical actions. In many ways, this dissertation mirrors to a greater degree my experiences at a quantitative cell modeling hackathon, which is described below and put into the context of the broader implications of the struggles and need for collaborate science at large (Pan and Johnson, 2019).

In ecological systems, evolutionary novelty is often found at the boundaries between disparate ecosystems—the so-called ‘‘edge effect.’’ In a similar fashion, conceptual breakthroughs in the natural sciences are often found at the boundaries between disparate disciplines. For instance, the modern synthesis in evolutionary genetics arose when statistical thinking was combined with Mendelian and Darwinian theories of

inheritance and speciation. The Human Genome Project combined efforts in mathematical modeling, molecular biology, and algorithm development to create our current understanding of our genetic information.

Scientific funding agencies recognize this core principle, and both the NIH and NSF have broadly promoted interdisciplinary research through their mission statements and funding efforts, such as the NIH Common Fund. However, in putting interdisciplinary science into action, scientists face several challenges: the risk associated with dabbling in the unknown with no guarantee of success; the uneasiness of thinking outside of a ‘‘comfort zone’’ that reflects decades of specialized training; and the communication challenges that arise between collaborators who speak different ‘‘languages.’’

As graduate students with interests in interdisciplinary science, we (Johnson & Pan) were aware of these obstacles when our programs advertised a ‘‘cell modeling hackathon,’’ which promised to bring together 30 mathematical modelers and biologists to Half- Moon Bay, CA in a three-day collaborative workshop modeled after similar events in Silicon Valley. While we may have initially been unsure of what to expect, we found that the cell modeling hackathon acted as a pilot study in addressing the challenges of creating interdisciplinary collaborations, and left participants with the experience, knowledge, and confidence to put these into action.

The challenges we face as interdisciplinary scientists

The first challenge in collaborative science that the hackathon successfully addressed was the high barrier to entry. The ‘‘activation energy’’ of exploring an interdisciplinary question can deter collaborations in different ways. For example, trainees

suffer an opportunity cost when exploring a new field before having established a specialty of their own. Faculty, on the other hand, must weigh their pre-existing commitments against spending the time to find a collaborator who is equally interested in their questions. The hackathon solved this problem in a few ways: first, the short three-day duration minimized the time commitment for attendance; and second, the organizers secured NSF funding (UCSF, 2014) to cover the cost of attendance for all participants. These two features lowered the typically high energy barrier of delving into interdisciplinary science for both trainees and professors alike.

The accessibility of the event led directly to the second critical component that made the event successful: the diversity of attendees. On the first day of the hackathon, all participants were given 60 seconds and one slide to introduce themselves and their interests. From the outset of these ‘‘lightning talks,’’ we were struck by the diversity in geography, research interests, fields of study, and career levels, with equal representation between professors and trainees. And just as particles in a highly entropic state can freely explore all possible states of a landscape, these energetic lightning talks lowered the barriers for interaction and allowed for novel collaborations that may not have occurred in a more structured educational setting.

This highly entropic state encouraged a third key aspect of participant behavior: leaving one’s comfort zone. Because no specific project was announced in the call for attendance, participants came in with few defined scientific expectations other than to learn something new. Following the buzz of the lightning talks, participants went on ‘‘speed-dating rounds’’ with potential collaborators to brainstorm hackable project ideas. The result

was that each participant found themselves in a previously unexplored state—from mathematics professors encountering new biological entities to model to experimentalists discovering how models can inform and advance their hypotheses. Teams began to form on the ‘‘edges’’ of common interests spanning theoretical and experimental disciplines. At the end of this period, we had small teams focusing on topics spanning biological networks to the biophysics of plant seed expansion to modeling cell motility.

Putting principles to practice

For the remainder of the three-day hackathon, groups focused on modeling their chosen biological questions. However, as with any new group effort, the initial ‘‘honeymoon’’ phase gave way quickly to the tension of overcoming the barriers that exist between disciplines. Critically, the three features of the hackathon that facilitated group formation—the low barrier entry, the diversity of the attendees, and the willingness to leave one’s comfort zone—allowed groups to overcome these challenges in a fluid way. While modelers and biologists initially faced a language barrier, this gave way to active learning between group members. The need to accomplish a common goal in a short amount of time meant that asking ‘‘stupid’’ questions was a necessity as opposed to a risk.

Furthermore, because all participants came in on an equal footing and with diverse expertise, participants found themselves serving as a student in one exchange and a teacher in the next. This is in stark contrast to more formal workshops where professors lecture to trainees. Abby Gerhold, assistant professor of cell biology at McGill University, found this inversion of the academic hierarchy rewarding. ‘‘Sometimes,’’ she said, ‘‘a student recently entered into a field makes a better teacher than someone who has been operating

in that sphere for many years, as they can remember what it was they did not know before entering.''

To us, this active learning across all disciplines and career levels was the key outcome that defined the hackathon's success. For modelers, the hackathon was a chance to think deeply with a biologist to contextualize their mathematical skills. As Wanda Strychalski, assistant professor of mathematics at Case Western Reserve University, put it, ''It's important for modelers to be tied to a specific problem and to actually help the biologist for the research to have scientific relevance.'' As students, we found ourselves explaining advanced mathematical approaches to professors, while we acquired insight into what models need to account for in complex biological systems to create new hypotheses. These kinds of exchanges became the hallmark of our next few days together—punctuated with more light-hearted moments including meals, early morning runs, and late-night beers.

During the capstone presentations on the last night of the hackathon, we were struck by the uncanny creativity and insight that groups deployed to model their biological phenomena. Although of course not all questions were answered in three days, several groups designed wet lab experiments to be performed once everyone got home. But beyond the tangible success of the specific collaborative projects, the capstone presentations left participants with a sense of potential and empowerment. They had gone through all the steps—overcoming the obstacles to forming collabo- rations and then actually struggling to work together—delegating expertise, constantly switching roles from teacher to learner—to produce something meaningful. This short, three-day experience became a

springboard for exploring the enormous landscape of possibilities that emerge when different disciplines come together, address the challenges they must face, and leave with the knowledge, power, and confidence to bridge that gap in the future.

Several of the lessons from the hackathon can be applied to interdisciplinary workshops at large. First, prioritizing a low cost of entry (in both money and time) and a diverse base of attendees can lead to a willingness to leave one's comfort zone that is essential for interdisciplinary research. Second, encouraging active learning between participants spanning career levels and expertise can help overcome communication barriers and unlock ''edge effects'' between disciplines. While the funding for this particular event was obtained through an experimental NSF grant, similar events have been included at the beginning or the end of specialized interdisciplinary conferences (examples include the Cold Spring Harbor Networks meeting (*Cold Spring Harbor Laboratory: Meetings & Courses Program*, 2019) as well as the CiViC users meeting (*CiViC: Clinical Interpretation of Variants in Cancer*, no date). Hackathon-style events have also been used by funding agencies, such as the Gordon Betty Moore Foundation (*Gordon and Betty Moore Foundation*, 2020) to generate new ideas for funding. Regard- less of the specific questions or format, we believe that following the basic principles addressed in the cell modeling hackathon will allow its success to be widely replicated in institutions worldwide and join others (Bauer *et al.*, 2018; Justman, 2018) in the call for focused collaborative efforts in our scientific community.

A preview of some of the interdisciplinary scientific efforts to follow

Each of the three body chapters represent examples of highly collaborative, integrated, scientific work, which dealt with and attempted to address some of the aforementioned challenges. In the second chapter, we present a project that is truly data-driven, and seeks to use drug response assays from multiple time points throughout treatment to uncover differences in subpopulation composition of sensitive and resistant cell states during treatment response. This work was initiated by Grant Howard, who performed all of the experimental work for the project and drove the conceptual question of how to use mathematical modeling to reveal the observed differences drug resistance over time. This necessitated much dialogue between all involved parties, and truly sought to answer a relevant question given the available data.

The project described in the third chapter instead focuses on answering a specific biological question- namely is it possible to identify from cell line data an ecological phenomenon known as the Allee effect in cancer cell growth dynamics. The Allee effect describes a situation in which, at low population densities, the per capita growth rate positively scales with population density, thought to be due to cooperative interactions. This biological question was of interest because of its vast implications in tumor initiation and metastasis, and its lack of experimental validation due in part to the difficulty of measuring cancer cells at very small population sizes. Taking advantage of novel technological advances, this project captures high-throughput, single cell resolution data, and used the full breadth of available data to calibrate to stochastic models of tumor growth. Here, stochastic models were necessary to one; make use of all of the data, but two; to deconvolve the difference between stochastic small-size effects and true cooperative Allee

effects. This work represents a case in which novel technological advancements in data collection enabled the application of more sophisticated mathematical methods to answer a relevant biological question.

Lastly, in the fourth chapter of this dissertation, we describe a project that was truly a collaborative effort between every single member of the lab. From the technology development enabling lineage-traced single cell RNA sequencing (Al'Khafaji, Deatherage and Brock, 2018; Al'Khafaji *et al.*, 2019), and the corresponding bioinformatics and data normalization performed by Russ Durrett, Eric Brenner, and Daylin Morgan, required to link the immense amount of raw data to useful gene-cell-lineage data sets to the longitudinal treatment response experiments acquired by Grant Howard in the same cell line. The two "separate" efforts were able to corroborate one another in that Grant's data only described gross population dynamics amenable to understanding total tumor cell number over time, whereas the lineage-traced scRNA-seq represented a few static snapshots of the underlying biology of the cell population. Given the depth of the scRNA-seq and breadth of the treatment response data, this project sought to make use of all of the available data to develop the most informed mathematical framework to understand treatment response dynamics. We demonstrate in this work the vast improvements made possible by positioning a mathematical model with the task of making sense and making use of all possible sources of data, and hope that this work reflects not only the power of collaborative science but an example in which a mathematical model can be of use to a variety of different disciplines and interests.

[2]Chapter 2: A multi-state model of chemoresistance to characterize phenotypic dynamics in breast cancer

PREFACE

In this chapter, a data-driven combined experimental-computational investigation is presented. The goal of this work was to use time-resolved dose-response assays following a drug treatment, to reveal how cancer cell populations were responding over time to a pulse treatment of chemotherapy. In order to do this, we developed a new method of analyzing serially acquired drug-sensitivity assays, and demonstrated that this framework could be used to identify the composition of mixed populations of cells via a separate validation experiment and testing of the mathematical framework. The findings in this investigation revealed that the drug resistance of a population changes dynamically following treatment, and inspired future investigations to be presented in subsequent chapters.

ABSTRACT

The development of resistance to chemotherapy is a major cause of treatment failure in breast cancer. While mathematical models describing the dynamics of resistant cancer cell subpopulations have been proposed, experimental validation has been difficult

[2] Note: This chapter is based on an article originally published as:
Howard, G.R.*, Johnson, K.E.*, Ayala, A.R., Yankeelov, T.E., & Brock, A. (2018). A multi-state model of chemoresistance to characterize phenotypic dynamics in breast cancer. Scientific Reports, (July), 1–11. https://doi.org/10.1038/s41598-018-30467-w.
*=equal contribution
Author contributions:
Conceptualization: Amy Brock, Grant R. Howard Investigation: Grant R. Howard, Kaitlyn E. Johnson, Areli R. Ayala, Data curation: Grant R. Howard, Kaitlyn E. Johnson, Areli R. Ayala, Formal analysis: Grant R. Howard, Kaitlyn E. Johnson, Amy Brock, Thomas E. Yankeelov, Writing- original draft: Grant R. Howard, Kaitlyn E. Johnson, Amy Brock, Writing- review and editing: Grant R. Howard, Kaitlyn E. Johnson, Amy Brock, Thomas E. Yankeelov, Supervision: Amy Brock Project Administration: Amy Brock Funding acquisition: Amy Brock.

due to the complex nature of resistance that limits the ability of a single phenotypic marker to sufficiently identify the drug resistant subpopulations. We address this problem with a coupled experimental/modeling approach to reveal the composition of drug resistant subpopulations changing in time following drug exposure. We calibrate time-resolved drug sensitivity assays to three mathematical models to interrogate the models' ability to capture drug response dynamics. The Akaike information criterion was employed to evaluate the three models, and it identified a multi-state model incorporating the role of population heterogeneity and cellular plasticity as the optimal model. To validate the model's ability to identify subpopulation composition, we mixed different proportions of wild-type MCF-7 and MCF-7/ADR resistant cells and evaluated the corresponding model output. Our blinded two-state model was able to estimate the proportions of cell types with an R-squared value of 0.857. To the best of our knowledge, this is the first work to combine experimental time-resolved drug sensitivity data with a mathematical model of resistance development.

INTRODUCTION

We aim to investigate how the therapeutic sensitivity of a breast cancer cell population changes over time following exposure to a pulse of chemotherapy. We hypothesize that intratumoral heterogeneity and cellular plasticity play a direct role in the progression of resistance. This hypothesis is based on previous work demonstrating that exposure to chemotherapy induces gene expression changes, metabolic state transitions, and increased drug resistance in subsets of cancer cells(Brock, Chang and Huang, 2009a;

Piyush B Gupta *et al.*, 2011; Saunders *et al.*, 2012; Basanta *et al.*, 2013; Lavi *et al.*, 2013; Pisco *et al.*, 2013; Brock, Krause and Ingber, 2015; Shajahan-Haq, Cheema and Clarke, 2015; Brock and Huang, 2017; Keisha N Hardeman *et al.*, 2017). We test this hypothesis of the direct role of the changing composition of subpopulations of differing drug resistance in the observed resistance response using mathematical modeling to estimate the relative frequencies of cells in different drug sensitivity states over time.

Approximately 30 percent of women diagnosed with early-stage breast cancer develop resistance and ultimately progress to metastatic breast cancer(Rivera and Gomez, 2010). Doxorubicin is a standard-of-care cytotoxic agent indicated for the treatment of breast cancer; however, the average time to develop resistance to doxorubicin is only 6 to 10 months (Rivera and Gomez, 2010). Thus, it is critical to develop a mathematical-experimental approach to describe and predict the conditions and dynamics associated with the onset of resistance *in vitro*, ultimately to improve the efficacy of clinical treatment regimens. We and others have demonstrated evidence of cellular plasticity and adaptability in response to treatment with chemotherapy (Brock, Chang and Huang, 2009a; Piyush B Gupta *et al.*, 2011; Saunders *et al.*, 2012; Pisco *et al.*, 2013; Brock, Krause and Ingber, 2015; Brock and Huang, 2017). For example, it has recently been revealed that melanoma cells exhibit heterogeneity in their metabolic state, with cells utilizing different amounts of oxidative phosphorylation and aerobic glycolysis (Keisha N Hardeman *et al.*, 2017). In this study of the role of metabolic usage in drug response, functional heterogeneity played a direct role in drug resistance as treating with a drug that inhibited aerobic glycolysis led to an increase in sensitivity to treatment (Keisha N Hardeman *et al.*, 2017). The ability of

individual cells to transition from a drug-sensitive to drug-resistant state has been observed in HL60 leukemia cells following chemotherapy exposure. Pisco *et al.* demonstrated that a subpopulation of cells increases expression of the ABC-transporter protein MDR1 in response to a chemotherapeutic pulse, leading to increased drug efflux and increased chemoresistance in those cells(Pisco *et al.*, 2013). These experimental results focus on specific drug resistance phenotypes that emerge in cell subpopulations following treatment. However, because of the vast complexity of resistance mechanisms, it is difficult to identify a single molecular marker of drug resistance that encompasses all drug resistant cells (Ibrahim-Hashim *et al.*, 2017; Wooten and Quaranta, 2017).

Mathematical descriptions of the dynamics of drug resistance may play a critical role in the development of strategies to combat drug resistance(Panetta, 1997; Mumenthaler *et al.*, 2013; Chisholm, Lorenzi and Clairambault, 2016; Enriquez-navas *et al.*, 2016; Yankeelov *et al.*, 2016; Wooten and Quaranta, 2017). Theoretical models have been proposed that incorporate heterogeneous subpopulations in predicting and optimizing treatment response(Foo and Michor, 2009; Wilkinson, 2009; Silva and Gatenby, 2010; Greene *et al.*, 2015; Mumenthaler *et al.*, 2015; Badri *et al.*, 2016; Harris *et al.*, 2016; Poleszczuk *et al.*, 2016; Hansen, Woods and Read, 2017; Matthew T. McKenna, Weis, Brock, *et al.*, 2018) however, these models have not been fully validated with experimental cell population data *in vitro* or *in vivo.* While approaches that incorporate the heterogeneity of resistant and sensitive subpopulations are promising, they remain largely theoretical in nature(Panetta, 1997). Strategies such as optimal control theory(Foo and Michor, 2009; Greene *et al.*, 2015)(treatment aimed at maintaining the optimal composition of cell

subpopulations), adaptive therapy(Gatenby *et al.*, 2009), and alternate metronomic dosing schemes(Foo and Michor, 2009; Montagna *et al.*, 2014) have rarely been implemented in patient care because of lack of experimental validation. Validation of the presence of the predicted subpopulations proposed in these models is essential for progressing from theoretical predictions to implementation.

Although resistance to chemotherapy is a major cause of failure in breast cancer, we do not currently have a mathematical model describing the development of resistance in the context of a dynamic heterogeneous cancer cell population. Conversely, experimental evidence concerning the variety of biological mechanisms of drug resistance is largely derived from static biological observations(Gottesman, 2002; Abuhammad and Zihlif, 2013). Many studies have relied on chemoresistant cell lines established by long-term exposure of cells to escalating doses of chemotherapeutic agent. In some cases, the chemotherapeutic is a required component of the cell culture media, to maintain resistant cell lines with a median lethal dose (LD50) up to 14 times higher than the original cell line(Abuhammad and Zihlif, 2013). Resistance observed in these cell lines may not be physiologically relevant to the clinical onset of chemoresistance, in which transient drug resistance may be induced in response to periodic treatment.

In this contribution, we calibrated experimental drug sensitivity data to multiple dynamic population models to test the hypothesis that there is a time-dependent population response to a chemotherapy treatment, and that this response is best described by models that incorporate heterogeneity and cellular plasticity. We combine the functional relevance of experimentally observed drug resistance data with various mathematical models to

reveal the dynamic proportions of cells in subpopulations defined by their degree of drug resistance. To validate that our modeling approach was able to identify the composition of a cell population, we applied the model to known mixtures of reference cell populations with different resistance. To the best of our knowledge, this is the first effort to temporally resolve the proportion of drug sensitive and resistant cells in an experimental population in response to transient drug exposure.

MATERIALS AND METHODS

Data acquisition

Cell Culture

MCF-7 human breast cancer cells were obtained from ATCC and maintained in MEM (Minimum Essential Media, Thermo Fischer) supplemented with 10% fetal bovine serum (Gibco) and 1% Penicillin-Streptomycin (Gibco). MCF-7/ADR human breast cancer cells were obtained from Robert Clarke(Vickers *et al.*, 1988) and maintained in MEM (Gibco) supplemented with 10% fetal bovine serum (Gibco), 1% Penicillin-Streptomycin (Gibco), and 500 nM doxorubicin (Sigma-Aldrich). A subline of the MCF-7 breast cancer cell line was engineered to constitutively express EGFP (enhanced green fluorescent protein) with a nuclear localization signal (EGFP-NLS). Genomic integration of the EGFP expression cassette was accomplished utilizing the Sleeping Beauty transposon system (Kowarz, Loescher and Marschalek, 2015). The EGFP-NLS sequence was ordered as a gBlock from IDT and cloned into the optimized sleeping beauty transfer vector pSBbi-Neo. pSBbi-Neo was a gift from Eric Kowarz (Addgene plasmid #60525)

(Kowarz, Loescher and Marschalek, 2015). To mediate genomic integration, this two-plasmid system consisting of the transfer vector containing the EGFP-NLS sequence and the pCMV(CAT)T7-SB100 plasmid containing the Sleeping Beauty transposase was co-transfected into the MCF-7 population utilizing Lipofectamine 2000. mCMV(CAT)T7-SB100 was a gift from Zsuzsanna Izsvak (Addgene plasmid # 34879)(Mátés *et al.*, 2009). GFP^+ cells were collected by fluorescence activated cell sorting. This MCF-7-EGFPNLS1 cell line is maintained in MEM (Gibco) supplemented with 10% fetal bovine serum and 200 μg/mL G418 (Caisson Labs).

Time resolved resistance measurement

MCF-7 cells were plated at 6600 cells/cm^3 and cultured for two days in growth media. The media was then exchanged for growth media containing 500 nM doxorubicin. After 24 hours, the doxorubicin media was removed and replaced with growth media to end the drug pulse. Cells were passaged and counted weekly and drug sensitivity assays were performed weekly, as described below. Cell number counts at each week were used to determine the average per capita growth rate per day of the recovering cell population (Figure 2.1a). The pulsed dosing of doxorubicin, followed by weekly drug sensitivity assays for 8 weeks, was repeated for a total of five independent MCF-7 cell populations in order to obtain multiple replicates at each time point assayed.

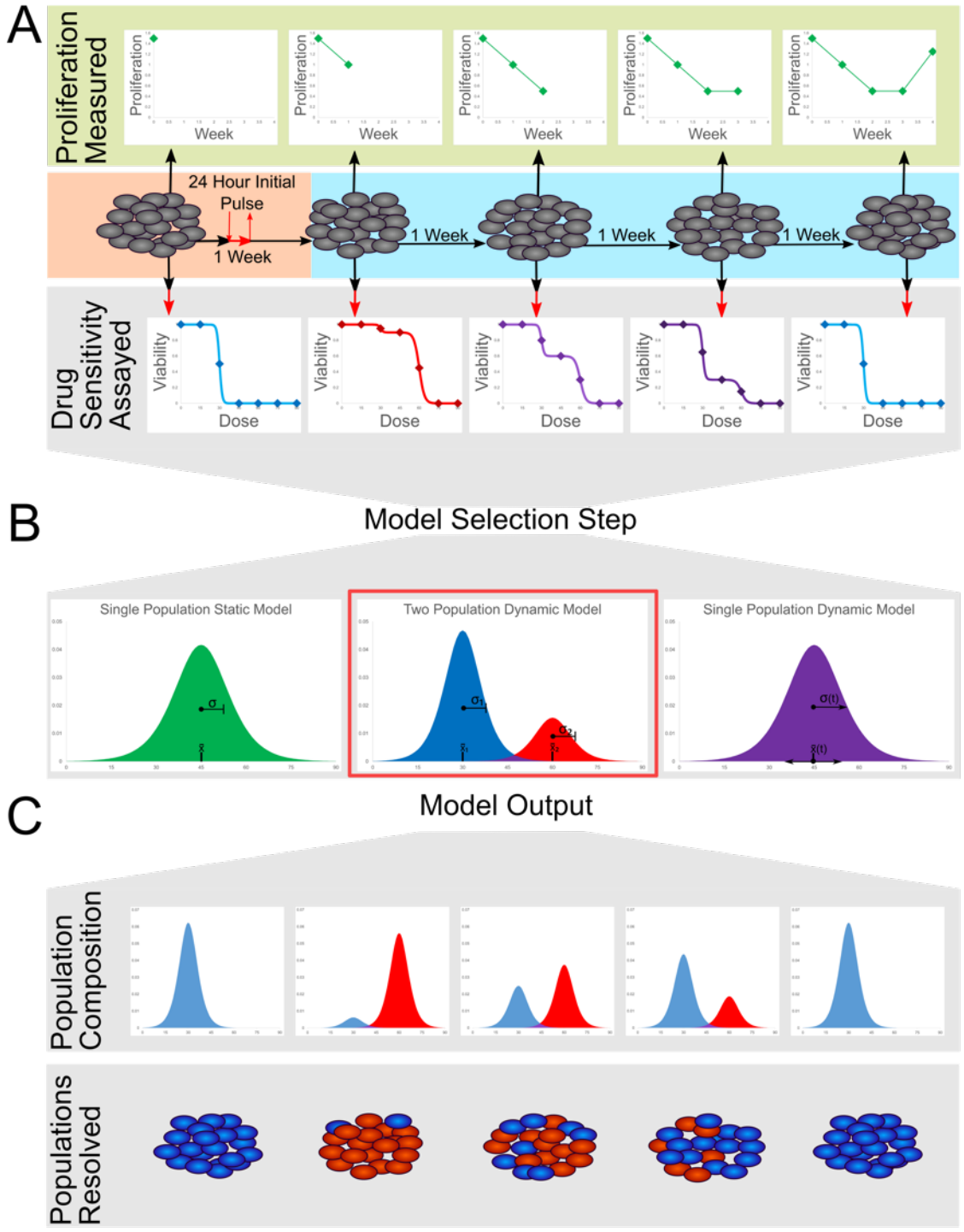

Figure 2.1: continued next page, Experimental and modeling workflow: a. MCF-7 breast cancer cells are treated with an initial pulse of doxorubicin (500 nM) for 24 hours.

After treatment, the instantaneous growth rate is measured at each week. However, the subpopulation composition of drug sensitive and resistant cells is not easily identifiable from any single biomarker, as is indicated by the gray cells. To quantify the changes in drug resistance as the population responds to treatment, a subset of cells are extracted each week and a drug sensitivity assay is performed. **b.** Using the combined data set containing a drug sensitivity assay at each time point, multiple mathematical models are tested to determine the optimal method for capturing the dynamic response of the cell population. Model selection statistics indicate that a multi-population model of at least two subpopulations is the optimal model. **c.** The dynamic two population model estimates the presence of two subpopulations with distinct LD50s and variances corresponding to a sensitive and resistant subpopulation. The model mandates that these states remain constant throughout drug response, with the changes in drug sensitivity of the whole population resulting from changes in the proportions of the areas under the curve of the sensitive versus the resistant population. The model reveals the composition of resistant and sensitive subpopulations at each time point, as is indicated schematically by the ability to identify the proportions of red and blue cells in the population at each week.

Weekly drug sensitivity assay

Each week, a subset (300,000) of the cells that were exposed to doxorubicin at the start of the experiment were plated into a 12-well plate in growth media. After two days of culture, media was exchanged for growth media containing doxorubicin at a range of concentrations (0, 4, 14, 24, 36, 48, 60, 72, 84, 96, 120, and 144 μM). Twenty-four hours after this dosing, the cells (including supernatant media) were collected via trypsinization, pelleted, and resuspended in 20 μL of media. Live and dead cells were identified with acridine orange and propidium iodide (ViaStain AOPI Staining Solution, Nexcelom Bioscience) and quantified with a Nexcelom Cellometer VBA. The ratios of live to dead cells were used to determine the viability at each concentration of doxorubicin (Figure 2.1a).

Cell mixtures for model validation

MCF-7-EGFPNLS1 and MCF-7/ADR cells were counted, mixed at desired ratios (1:0, 3:1, 1:1, 1:3, and 0:1), and plated in 12-well plates as described above (*Weekly drug sensitivity assay*). For each defined mixture, a sample of the untreated sample was counted in the Nexcelom Cellometer VBA to determine the measured percent of resistant cells, using the EGFP fluorescence of the MCF-7-EGFP-NLS1 cell line as a marker for the number of MCF-7-EGFP-NLS1 cells which are wild-type with respect to drug sensitivity. The measured percent of cells of each type was calculated by normalizing based on measurement of fluorescence in pure wild type (MCF-7-EGFP-NLS1) and MCF-7/ADR samples.

Data Analysis

Calibration of experimental data to multiple structural models of dose-response

The combined data set containing the drug sensitivity assays from each week were fit to three different models (Table 2.1). The combined data set consists of three variables, the time (in weeks) post-initial doxorubicin exposure, the concentration of doxorubicin applied at that time, and the corresponding cell viability. To perform an estimation of the parameters for all three models, a nonlinear, least-squares approach was implemented in MATLAB (Mathworks). Sigmoidal viability curves are often used to describe chemotherapy dose-response curves (Gardner, 2000) because they correspond to a population of cells that die at different doses that can be described with a unimodal distribution of cells versus dose as shown in in Figure 2.1b. The parameters of the sigmoidal dose response function are physically identifiable as the dose at which half the cells die (LD50) is the center parameter (c_{ss}) and the standard deviation of the unimodal distribution is inversely related to the slope of the sigmoidal dose-response curve (m_{ss}). The single static population model is the simplest of the models and ignores the temporal dependency of response. Here the drug response of the combined data set is described by a single homogenous cell population with a single (static) LD50 and slope. The single static population model equation is:

$$V_{singlestatic}(d) = \frac{V_{max}}{1+e^{(m_{ss}(d-c_{ss}))}} \quad (1)$$

where V is the proportion of cells viable at the dose, d, of doxorubicin in μM applied, c_{ss} is the LD50 of the population, m_{ss} is the slope at which the cells die due to increases in concentration, and V_{max} is the maximum viability of the cell population (as measured by

the assay in absence of drug). The V_{max} parameter is included to normalize for naturally occurring cell death independent of the effects of doxorubicin. The single static model represents the null hypothesis that the initial pulsed dose has no time-dependency in its effect on the cancer cell population.

Description	Model equation	Variables and parameters
Single static model	$V_{singlestatic}(d) = \frac{V_{max}}{1 + e^{(m_{ss}(d - c_{ss}))}}$	V = fraction of cells viable in population V_{max} = maximum viability of cell population (baseline viability at dose = 0 μM) d = dose of doxorubicin (μM) m_{ss} = slope of loss of viability as dose increases, $m_{ss} = \frac{1}{\sigma}$ c_{ss} = LD50 to describe all data assuming no change in time d = dose of doxorubicin (μM)
Single dynamic model	$V_{singledynamic}(d,t) = \frac{V_{max}}{1 + e^{(m_{sd}(t)(d - c_{sd}(t)))}}$	$m_{sd}(t)$= slope of loss of viability as dose increases, $m_{sd}(t) = \frac{1}{\sigma(t)}$ for each population $c_{sd}(t)$= LD50 to describe data at each time point
Two population dynamic model	$V_{twopop}(d,t) = Vmax(\frac{f_{sens}(t)}{1 + e^{(m_{sens}(d - c_{sens}))}} + \frac{1 - f_{sens}(t)}{1 + e^{(m_{res}(d - c_{res}))}})$	$m_{sens}(t)$= slope of loss of viability as dose increases, $m_{sens}(t) = \frac{1}{\sigma(sens)}$ $m_{res}(t)$= slope of loss of viability as dose increases, $m_{res}(t) = \frac{1}{\sigma(res)}$ c_{sens} = LD50 to describe sensitive population c_{res} = LD50 to describe resistant population

Table 2.1: Mathematical models to describe dynamic drug sensitivity data: We present the equations used for each of the three different structural models that were fit the time-resolved drug sensitivity assays. The column labeled, "Model equation" provides the functional form of the equation, with *t* representing a parameter that was fit to the data set at each time point measured. The column labeled, "Variables and parameters" describes the variables used in terms of their physical meaning and their relation to the time-resolved drug sensitivity assays.

The single dynamic population model incorporates a temporal dependency when fitting the combined data set. For each time point that drug sensitivity was assessed, the data is fit to an individual dose-response curve to generate LD50 and slope parameters. The

model describes the drug response as a single homogenous population whose drug tolerance can change in time. The single dynamic population model equation is:

$$V_{singledynamic}(d,t) = \frac{V_{max}}{1+e^{(m_{sd}(t)(d-c_{sd}(t)))}} \quad (2)$$

where the c_{sd} and m_{sd} (LD50 and slope, respectively) parameters pertaining to each week, leading to a 16-parameter model (slope and LD50 at each of the 8 weeks). This model is akin to individually fitting a dose response curve to each week that the drug sensitivity assays were performed.

Finally, the two-population dynamic model describes a cell population with two cell states that differ in drug sensitivity. The dynamics of the drug response are captured by the relative frequency of cells in each state at each time point. The two-state dynamic population model equation is:

$$V_{twopop}(d,t) = Vmax\left(\frac{f_{sens}(t)}{1+e^{(m_{sens}(d-c_{sens}))}} + \frac{1-f_{sens}(t)}{1+e^{(m_{res}(d-c_{res}))}}\right) \quad (3)$$

where each cell state is modeled as a subpopulation of cells whose LD50 is centered about a mean and slope (c_{sens}, m_{sens} and c_{res}, m_{res}, respectively) which remain constant over time and whose f_{sens} and f_{res} (1 - f_{sens}) parameters can vary to best capture the drug sensitivity assay at each week. To fit the two-state model to data from multiple time points, the parameters of the sensitive and resistant slope and LD50 were forced to be constant at all time points, and the sensitive and resistant fraction parameters were allowed to float at each week, leading to a 12-parameter model (4 fixed parameters, 8 time-dependent fraction parameters for each of the 8 weeks). Equation 3 describes two cell states with distinct LD50 values whose relative frequencies are able to change in time after initial chemotherapy

exposure, but whose LD50s remain constant. The overall cell population viability (measured) is modeled as a direct sum of the viability response in each subpopulation.

Statistical analysis and model selection

For all three models, the confidence intervals on the parameter estimates were constructed using the bootstrapping method of replacement (Efron, 1987), with 500 bootstrapped simulated data sets. For each model, the mean-squared error and the Akaike Information Criterion (Konishi and Kitagawa, 2008) were calculated for stand-alone model statistics. The Akaike Information Criterion estimator (i.e., the AIC value) is used for direct model comparison. The AIC value evaluates a model based on goodness of fit and penalizes for the complexity of the model using the number of parameters, with a lower AIC value indicating a better model. These evaluation criteria are used to determine the most appropriate model to describe the dynamic dose response data (Figure 3.1b, Table 2.2).

	Single static model	Single dynamic model	Two population model
Number of parameters	2	16 2 per time point	12 4 + 1 per time point
AIC value	-2004.4	-2015.6	-2126.4
Mean-squared error	0.057	0.040	0.024

Table 2.2: Model fit and model selection statistics indicate the two population model is the optimal model: The two population dynamic model has the lowest AIC value and the lowest mean-squared error, indicating that the this model is superior to the single dynamic population model and the single static population model. The number of parameters for the single dynamic model and the two population model vary by the number of time-resolved drug sensitivity assays examined. In this case, for the single dynamic model there are two model parameters of LD50 and slope at each time point examined. In the two population model, there are four model parameters of LD50 and slope for the sensitive and resistant populations respectively, which remain constant, and one additional parameter per time point is used to describe the proportion of cells in each subpopulation at the time the cells were assayed.

Model validation

To validate the modeling approach, we tested each models' ability to identify known mixtures of wild-type MCF-7 cells with MCF-7/ADR resistant cells. The same two population model and fitting algorithm described above (in the section labeled, *Calibration of experimental data to multiple structural models of dose-response***)** were used to fit the combined data set containing a mixture identification, concentration of doxorubicin, and corresponding cell viability. Therefore, in this validation step, instead of grouping the data by week post drug treatment, we grouped by mixture composition. We allowed each group of mixture replicates to be fit to their own fractional parameter and maintained that the LD50 and centers of the two populations remain constant as described previously. We evaluated the model output of relative frequencies against our measured relative frequencies of wild-type MCF-7 and MCF-7/ADR cells (see *Cell mixtures for model validation*). The bootstrapping method of replacement, again with 500 simulated data sets, was used to construct the 95% confidence intervals for each parameter estimate.

RESULTS

Cancer cell population exhibits time-dependent response to pulse treatment

To determine whether the resistance of the MCF-7 population changes in time after the pulse treatment of doxorubicin, we fit the combined data set to both the static and dynamic single population models, as shown in Figure 2.2a & b. In Figure 2.2a, the experimentally measured viability is shown alongside the single static population model curve, for both the untreated controls (black) and the pulse-treated cell populations

(purple). The LD50 to describe the resistance of these populations is 37.0 +/- 3.5 μM and 50.4 +/- 2.4 μM for the untreated and treated populations, respectively. Both slope and center parameters for the single static population model can be found in Table 2.4. In Figure 3.2b, we show the LD50 estimates, with the 95% confidence intervals, from the single dynamic model over all 8 weeks. The lower AIC value of the single dynamic model (-2015.6) compared to the single static model (-2004.4) indicates the time-resolved dose-response data are better described by the single dynamic model that allows the slope and LD50 value of the population to change at each time point (Table 2.2). The pattern in estimates of the LD50 values in the single dynamic model (Figure 2.2b) corroborate this statistical analysis, showing a significant peak in the drug resistance at intermediate time points, with an LD50 at week 2 at 67.2 +/- 10.0 μM, followed by a slow return towards baseline, reaching 46.4 +/- 5.1 μM at 8 weeks. The LD50 and slope parameter values for the single static population model for all time points can be found in Table S2.2 of the Supplementary Materials.

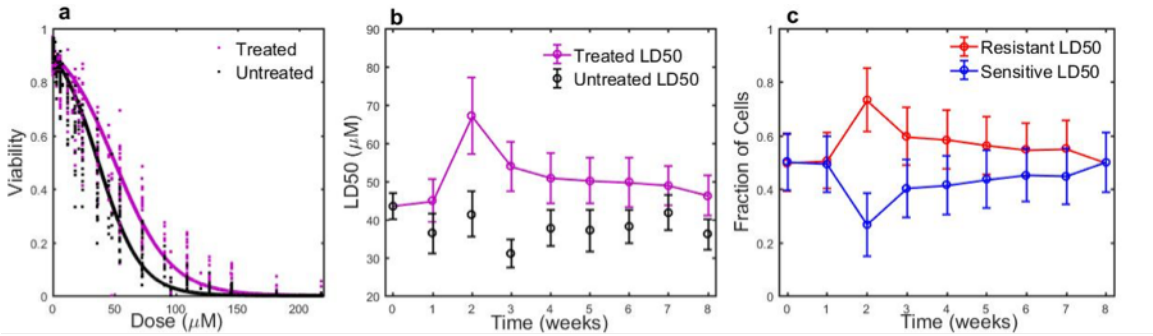

Figure 2.2: **Time-resolved drug sensitivity assays fit to multiple models: a.** The static single population model (solid lines in panel a) demonstrates that the average resistance of the 8 weeks of compounded drug sensitivity data significantly increases from an LD50 = 37.0 +/- 3.5 µM for the untreated to an LD50 = 50.4 +/-2.4 µM following exposure to pulse-treatment of doxorubicin **b.** The LD50 estimates resulting from analysis with the dynamic single population model indicates that the increase in resistance is time-dependent, peaking at 2 weeks after treatment with an LD50 = 67.2 +/- 10.0 µM, followed by a slow return towards baseline resistance levels, reaching an LD50 = 46.4 +/- 5.1 µM at 8 weeks. **c.** The two population dynamic model displays the model-estimated proportions of a population with LD50 = 79.7 +/- 6.5 µM (resistant) and a population with LD50 = 22.4 +/- 2.0 µM (sensitive). The relative proportions of these two populations change over time, yielding the observed LD50 for the overall population. In panels b and c, parameter fits at each time point are connected by a line for visual aid and the error bars represent the 95% confidence intervals on the parameter values using the bootstrapping method of replacement with n=500.

Incorporating heterogeneity via drug sensitivity states improves description of response

To determine whether the dynamic drug response could be explained by a model of two subpopulations, we fit the data to the two population dynamic model to determine the degree of drug sensitivity of the two subpopulations and the resulting sensitive and resistant fraction parameter estimates at each week, with their 95% confidence intervals (Figure 2.2c). The two-state dynamic model estimates the presence of a resistant subpopulation with an LD50 of 79.7 μM +/- 6.5 μM and a sensitive subpopulation with an LD50 of 22.4 +/- 2.0 μM (Figure 2.2c). The parameter values for the sensitive and resistant slope and center, and the fractional parameters, can be found in Table S3 of the Supplementary Materials. We again use model selection to demonstrate that the two population model is an improvement over the single dynamic population model for describing the dynamic drug response of the cell population. Support for selection of this model is indicated by the lower AIC value (-2126.4) of the two population dynamic model compared with the single dynamic model (-2015.6) (Table 2.2). The results of the model selection analysis for describing dose response curves with multiple conditions (in this case time points) reveals that the differences in dose-response can not be modeled well by allowing each condition to be fit to a single sigmoidal curve allowed to shift, but rather is improved by modeling a sum of sigmoids corresponding to a multi-modal distribution of cells versus lethal dose whose proportions of cells in each state can change for each condition. To illustrate the improvement in fit to the data, Figure 2.3a displays the two-state model curve overlaid on the dose-response data at 2 and 8 weeks, demonstrating the

ability of the "shoulders" in the curve to fit the initially steep loss of cell viability at lower doxorubicin concentrations, as well as the persisting cell viability at higher doxorubicin concentrations. Figure 2.3b indicates the model output of relative frequencies of resistant and sensitive subpopulations for the model fit and data shown in Figure 2.3a. The improvement in fit of the two-state model is additionally supported by the lower mean-squared error value for the two-state model than the single dynamic model (mean-squared errors of 0.024 and 0.040, respectively), despite it having less parameters than the single dynamic population model (Table 2.2).

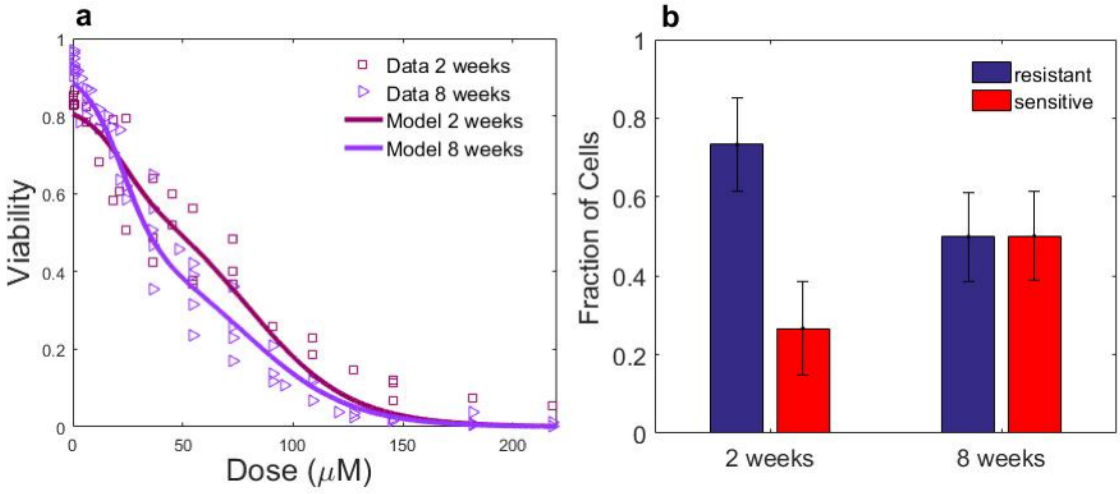

Figure 2.3: Example fit of drug sensitivity to the two-population model: a. Best fit of the two- population model to the dose response data at two and eight weeks post-treatment demonstrates the ability of the model to capture the differences in subpopulation levels at the different time points. **b**. Two population model output of parameter values of resistant and sensitive fractions at 2 and 8 weeks. The error bars represent the 95% confidence intervals on the parameter values using the bootstrapping method of replacement with n=500.

Subpopulation levels of resistant cells transiently increase

The key result for the treatment of doxorubicin on the MCF-7 cell line is that the proportion of cells in the resistant state is consistently higher than baseline from weeks 2 to 5 after the initial drug pulse, followed by return towards the initial resistant and sensitive subpopulation levels (Figure 2.2c). The measured per capita growth rate per day (births per cell per day minus deaths per cell per day) (Figure 2.4a) was used to estimate the total number of cells at each week. The number of total cells was combined with the fractional parameter estimates at each week (Figure 2.2c) to obtain estimates of the number of resistant and sensitive cells at each week in the treated population (Figure 2.4b,c). These estimates are purely empirical and make no assumptions about the mechanism at which the cells in each state reached the estimated number of cells at each time, allowing for the possibility of differential growth rates, drug sensitivities, and cell state transitions to determine the corresponding subpopulation levels. This differs from other published work which has assumed sensitive and resistant cells have different growth rates (Matthew T. McKenna, Weis, Brock, *et al.*, 2018) and distinct transition rates; here we do not attempt to define the mechanism by which the number of cells in each state was obtained.

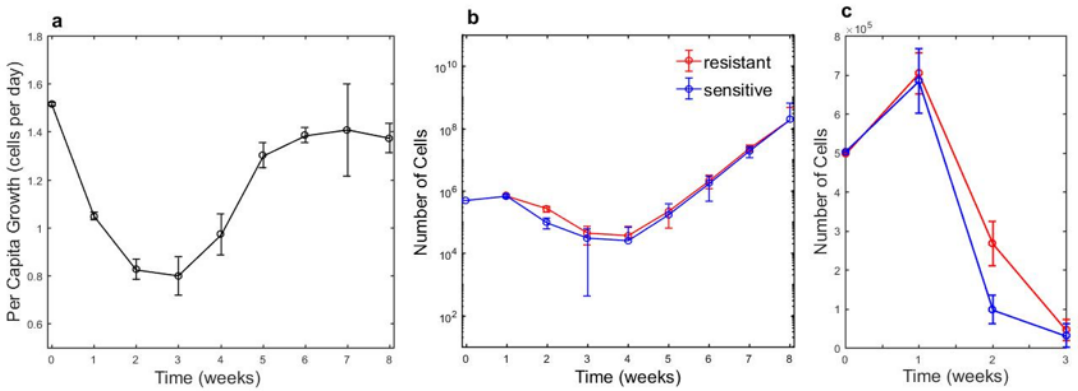

Figure 2.4: Data driven estimates of phenotypic dynamics: a. Per capita growth rate (number of births per cell per day minus number of deaths per cell per day) of the entire cell population following pulse treatment of doxorubicin. Error bars represent the 95% confidence intervals from three replicates with per capita growth rate measured at each week. **b.** Estimates of the number of resistant and sensitive cells over time are obtained by combining bulk population growth rate to estimate the total number of cells, with the estimates of subpopulation fractions from the two population model output. The estimates of the number of resistance and sensitive cells are purely empirical—cell numbers in each state can be obtained by a combination of differential growth rates, drug sensitivities, or cell state transitions and we do not attempt to identify which the means at which the cell numbers are obtained. Error bars are the compounded 95% confidence interval of the per capita growth rate measurement and the parameter estimation of resistant fraction from the two population model output using the bootstrapping method of replacement with an n=500. The number of cells is plotted in logarithmic scale. **c**. A closer look in numeric scale of weeks 1-3 displaying the higher number of resistant cells over this time interval.

Model validation confirms ability to reveal subpopulation composition defined by drug sensitivity

Without a molecular marker of drug resistance, the estimated changes in drug resistant and drug sensitive subpopulations from our two-state model are difficult to validate; that is, we did not know for certain that our model estimated parameters of resistant and sensitive fractions reflect the true subpopulation compositions. We validated the modeling approach by generating experimental reference standards consisting of mixtures of the wild-type MCF-7 cell line with its corresponding doxorubicin resistant cell line, the MCF-7/ADR. We evaluated the ability of the model to estimate the subpopulation composition of each reference mixture. In Figure 2.5a, the two population dynamic model fit for each mixture is overlaid on the dose-response data for the corresponding mixture. In Figure 3.5b, the measured percent of wild-type and MCF-7/ADR cells are plotted as the line of unity against the parameter estimations of the percent of resistant cells from the model output (Table 2.3). The measured proportions of wild type and ADR cells versus the model output have a coefficient of determination (R-squared) of 0.857 (Figure 2.5b).

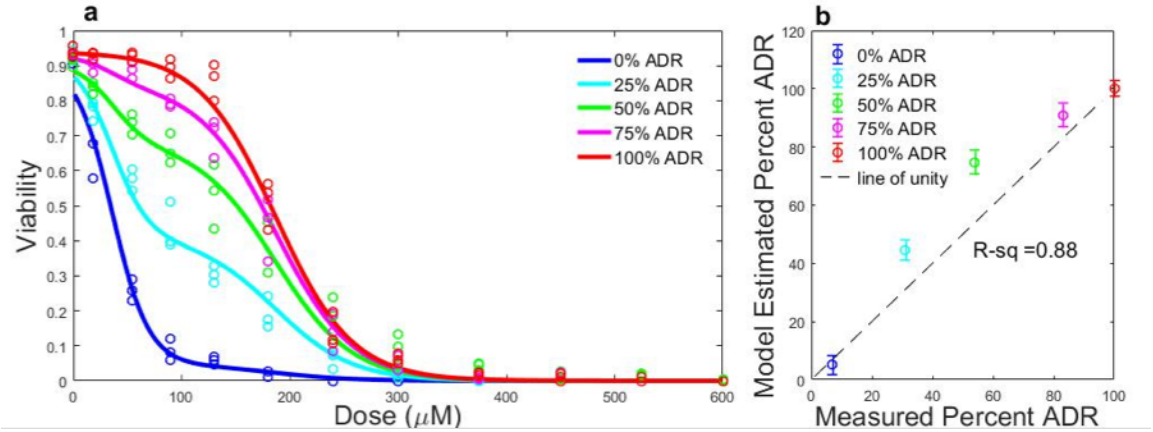

Figure 2.5: Model validation *via* identification of mixed populations: a. The two-population model fit for each mixture of MCF-7/ADR resistant cell line and wild-type MCF-7 cell line overlaid on the experimentally measured drug sensitivity for the corresponding mixture. **b**. Model-estimated percent of MCF-7/ADR cell line in mixture versus the measured MCF-7/ADR cell percentage using fluorescence cell counting. The coefficient of determination (R-squared) value of 0.857 indicates the efficacy of the fractional estimates from the two population model output.

	Measured	Model
$LD50_{ADR}$	187.5 +/- 3.8	186.7 +/-4.6
$slope_{ADR}$	0.034 +/- 0.0039	0.028 +/-0.0028
$LD50_{WT}$	37.1 +/- 3.3	35.95 +/-2.8
$slope_{WT}$	0.055 +/- 0.0077	0.060 +/-0.0076
$frac_{0\%ADR}$	0.07 +/-0.01	0.05+/- 0.03
$frac_{25\%ADR}$	0.30 +/-0.04	0.44 +/-0.04
$frac_{50\%ADR}$	0.54 +/-0.05	0.075 +/-0.04
$frac_{75\%ADR}$	0.83 +/-0.05	0.91 +/-0.04
$frac_{100\%ADR}$	1.0 +/- 0.02	1 +/- 0.0272

Table 2.3: Model validation demonstrates identifiability of subpopulation compositions: Measured versus two population model estimated parameters indicate the model's ability to reveal the known proportions of the MCF-7/ADR resistant cell line mixed with the wild-type MCF-7 cell line. "Measured" LD50 values and slopes indicate the LD50 and slope of the isolated pure MCF-7/ADR cells and pure wild-type MCF-7 cells fit to the single static population model. Measured resistant fractions are measured precisely using the Nexcelom fluorescence imaging counter of the number of green fluorescent wild-type MCF-7 compared to the total number of brightfield cells counted.

DISCUSSION

The key results of this combined experimental and modeling approach demonstrate time-dependency in the response to a clinically relevant pulse-dose chemotherapy treatment, and a modeling system that reveals estimates of the composition of a cancer cell population with functional heterogeneity in its chemoresistance. We have identified that a transient increase in drug resistance is observed following drug exposure and have shown that this can be best described through a model that captures the changing compositions of distinct subpopulations defined by drug sensitivity. These subpopulations were not

previously identified due to the complex nature of drug resistance mechanisms. The novelty in this modeling framework is that it is driven from drug sensitivity assays only, without imposing any assumed characteristics or isolated parameter measurements. A data-driven approach allows the model system to be applied to a variety of cell types and drug conditions; for example, in the model validation step, a mutant cell line with high resistance to doxorubicin is assessed. The key finding of this paper is to establish a foundation for describing observable resistance progression throughout time by revealing subpopulations that are not identifiable by molecular markers alone. Future work will need to systematically investigate the molecular and cellular mechanisms of these observed dynamics.

We acknowledge the many limitations of this study. The granularity of the modeling system was limited by technical constraints of the experiment dosing scheme. Bottlenecking of the population in the initial pulse-treatment limited the number of cells available at subsequent weeks for the corresponding drug sensitivity assay. While we were able to measure the response to 12 distinct doxorubicin concentrations each week, this was not sufficient to implement a multi-population model with more than two states, due to a lack of statistical power to significantly resolve differences in subpopulation compositions in time. We acknowledge that a model of only two subpopulations does not capture all likely relevant cell types but believe that this represents a useful simplification of resistance development. In the model validation phase, we were able to estimate the fractions of cells in each state with an R-squared value of 0.857 (Figure 3.5b). We were also able to estimate the LD50s of the two populations in mixture using the model within the 95% confidence

intervals of the isolated LD50s when fit to the static single population model alone (Table 2.3). It is possible that discrepancies in the fractional parameter estimates compared to the measured mixtures of the wild-type and resistant cell types may arise from biological interactions between the two that may increase the effective resistance of the cells when in contact with one another. A goal of future studies is to investigate the role of cell-cell interactions in drug resistance.

To ensure that the model parameters are uniquely identifiable and are not overfit to the experimental data sets, we tested the parameter identifiability of the model using simulated data with noise generated from the measured variability as a function of dose in the experimental data (Figure 2.6). We randomly generated 100 simulated data sets containing 5 different mixtures of resistant and sensitive cells to obtain a distribution for each of the parameter values, with the 95% confidence intervals around the true model parameter values shown in (Figure 2.7), demonstrating that the variability in the estimated LD50 (Figure 2.7a), slope (Figure 2.7b), and fraction parameters (Figure 2.7c) was reasonably small. We then addressed question of the proximity at which the fractional parameters could be distinguished from one another by generating simulated data sets with experimental noise as described, this time containing either mixtures of mostly low levels of resistant cells (Figure 2.7a) or of mostly high levels of resistant cells (Figure 2.7b). In Figure 2.8, the distribution of fractional parameter estimates is displayed as histograms, with each mixture labeled by color. We performed pairwise t-tests between each of the distributions of fractional parameter estimates to confirm that the distributions are statistically significantly different.

Experimental *in vitro* models of resistance to cytotoxic chemotherapy typically utilize resistant cell lines developed via continuous exposure to increased drug concentration (Ke *et al.*, 2011; Abuhammad and Zihlif, 2013). In most cases, drug resistant phenotypes are characterized by an end-point analysis following the stabilization of a resistant cell population. Previous work in the field has indicated that the resistant phenotype of doxorubicin-resistant MCF-7 arises by a multi-factorial process because of observable differences in morphology, gene expression, and DNA content between MCF-7 and MCF-7 resistant cell lines (Abuhammad and Zihlif, 2013). The MCF-7/ADR resistant cell line used in our model validation step has an LD50 value of 187.5 +/- 3.8 μM (Table 3.3) and is thus more than 5 times more resistant than the wild-type MCF-7 with an LD50 value of 37.5 +/- 3.8 μM (Table 2.3). The MCF-7/ADR we used typically show a larger proportion of spindle-shaped cells that grow in a more dispersed manner than the wild-type cells (Ke *et al.*, 2011). Other groups have developed resistant MCF-7 cell lines that are 14-fold more resistant to doxorubicin than the original MCF-7 cell line(Abuhammad and Zihlif, 2013). They report that the MCF-7 resistant cells are on average larger, contain multiple nuclei, and upregulate genes involving metabolism, drug efflux, and down regulate genes involving DNA repair(Abuhammad and Zihlif, 2013). While these experimental observations provide us with key observables to identify as markers of resistance, they do not address the dynamical changes associated with resistance as it develops, nor are they all encompassing. To our knowledge, previous studies involving resistant cell lines have not reported time-resolved measurements of drug resistance following a clinically relevant pulsed dose chemotherapy treatment.

Mathematical modeling of heterogeneity in cancer cell populations has been investigated *via* multiple structural models(Panetta, 1997; Foo and Michor, 2009; Silva and Gatenby, 2010; Mumenthaler *et al.*, 2013, 2015; Greene *et al.*, 2015; Badri *et al.*, 2016; Chisholm, Lorenzi and Clairambault, 2016; Enriquez-navas *et al.*, 2016; Harris *et al.*, 2016). In particular, many models have provided predictive capabilities of cell-line specific drug sensitivity (McKenna *et al.*, 2017), as well as insightful metrics for capturing the growth inhibitory capacity of different drugs(Hafner *et al.*, 2016; Harris *et al.*, 2016). Explorations into different therapy strategies such as optimal control theory have utilized the concept of resistant and sensitive cells within a tumor or cancer cell population (Foo and Michor, 2009; Silva and Gatenby, 2010; Mumenthaler *et al.*, 2013, 2015; Greene *et al.*, 2015; Badri *et al.*, 2016; Enriquez-navas *et al.*, 2016; Harris *et al.*, 2016; Hansen, Woods and Read, 2017). Many models of *in vitro* and *in vivo* cancer progression utilize compartmental ordinary differential equations and partial differential equations. In these models, oftentimes a number of key assumptions are made. For instance, in one model of a heterogeneous tumor, it is assumed that drug resistance is inversely related to proliferation rate (Gatenby *et al.*, 2009). Other models assume that all sensitive cells are susceptible to the chemotherapy, and do not account for the ability of initially sensitive cells to acquire drug resistance(Panetta, 1997). While these models can be extremely useful in capturing drug response and demonstrating the theoretical response to alternate treatment strategies under these sets of conditions, some of their predictions have yet to be fully validated experimentally due to technical limitations in identifying drug resistant subpopulation levels over time.

In this work, we reveal the dynamic changes in subpopulation composition in response to a pulse treatment of drug. In the future, time-resolved subpopulation relative frequencies can be used to develop a model that describes the relative stability of drug sensitivity states and how they change in response to chemotherapy exposure. Ultimately, the results of these experimentally guided models can be used to predict the effect, in terms of composition of resistant cells, of a specific dosing regimen on a cancer cell population over time. The goal of future studies is to use the proposed modeling framework to develop and experimentally validate optimal dosing regimens to be used to combat chemoresistance.

CONCLUSION

We present this work as one demonstration of the role of heterogeneity in the development of drug resistance. Our analysis indicates that the response to pulsed chemotherapy is time-dependent and that the two-population model identifies subpopulation compositions that change over time. The approach we describe here uncovers chemoresistant subpopulations in breast cancer cell lines and is generalizable to any system in which subpopulations may play a role in a dynamic measurable outcome.

CHAPTER 2 SUPPLEMENTARY FIGURES AND TABLES

Figures

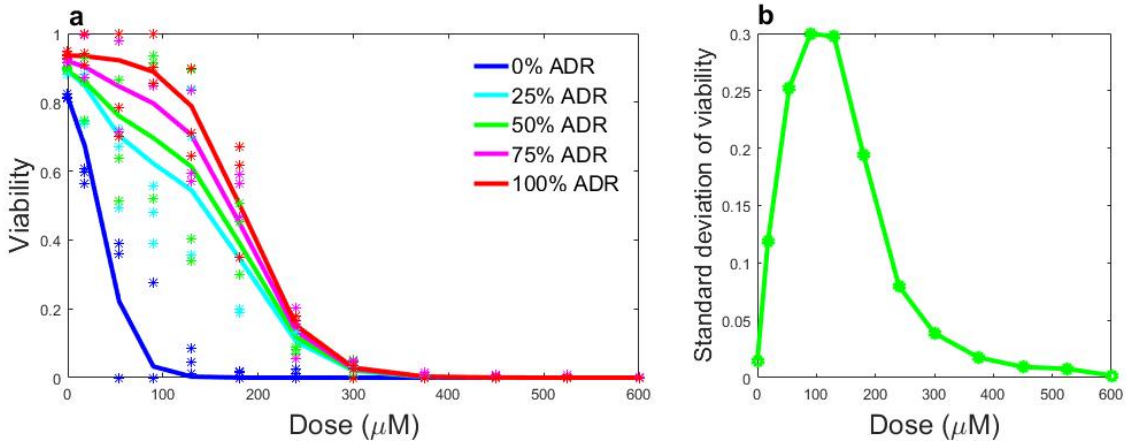

Figure 2.6: Two-population model and an example simulated data set with experimental noise. a. We inputted model parameters to simulate a resistant population with an LD50 of 185 μM and a sensitive population with an LD50 of 35 μM, and resistant fractions from 0-100 %, and simulated data with noise added according to experimentally distributed noise. **b.** This plot indicates the standard deviation in the measured cell viability of all mixtures of the naïve MCF-7 cells and the MCF-7/ADR resistant cells as a function of dose. The observed variability in viability measurements shown here were used to simulate data sets with experimentally observed noise. This plot indicates that the highest variability in cell survival response occurs at intermediate doses.

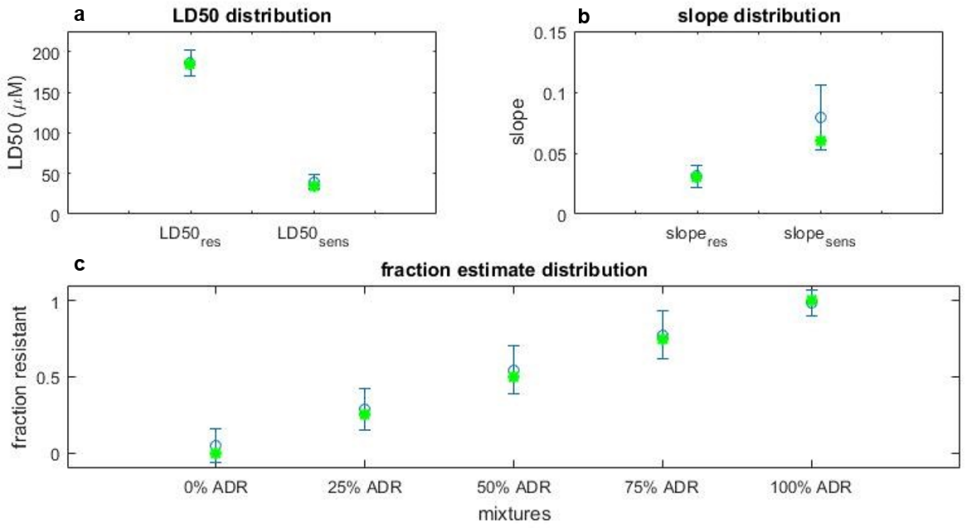

Figure 2.7: Parameter distributions for fitting of simulated data sets with experimental noise. a. We display the ninety-five percent confidence intervals around 100 fitted LD50 sensitive and resistant parameters for the simulated data sets in order to demonstrate the relative error in parameter identifiability of the LD50 values at the experimental noise observed. **b**. We display the ninety-five percent confidence intervals around 100 fitted sensitive and resistant slope parameters for the simulated data sets in order to demonstrate the relative error in parameter identifiability of the slope values at the experimental noise observed. We observe a higher relative error in the sensitive slope than in the resistant slope. **c**. We display the ninety-five percent confidence intervals around 100 fitted fraction estimate parameters for the simulated data sets in order to demonstrate the relative error in parameter identifiability of the fractions at the experimental noise observed. We observe the error to be fairly consistent across all of the mixtures.

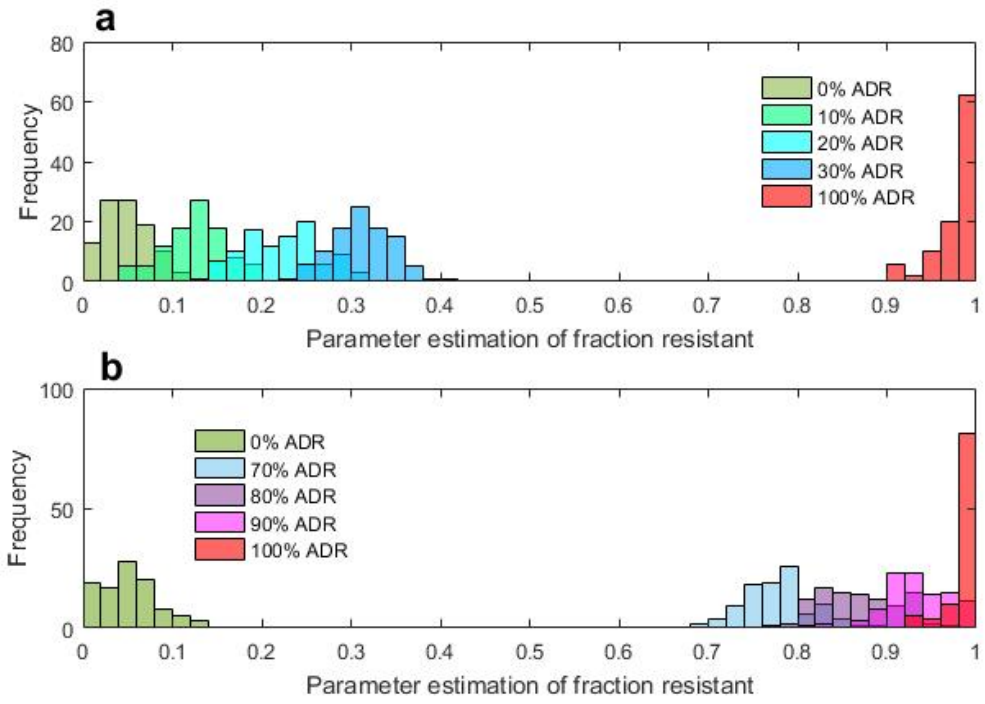

Figure 2.8: Parameter distributions around clustered input fraction parameters are identifiable within 10% with current experimental noise. a. Model parameters of clustered low resistant fractions were input at intervals of 10% from 0-30% resistant. One hundred simulated data sets with experimental noise were fit to obtain 100 model parameter estimates for each fractional parameter. We compared each pairwise set of low resistant fraction parameter distributions with a multiple comparison t-test in Matlab and found significant differences with a p-value of 9.92×10^{-8}. **b.** The same procedure was performed for high resistant fractions at intervals of 10% increase from 70-100% resistant. Again, the distributions of high resistant fraction parameter estimates were significantly different with a p-value of 9.92×10^{-8}.

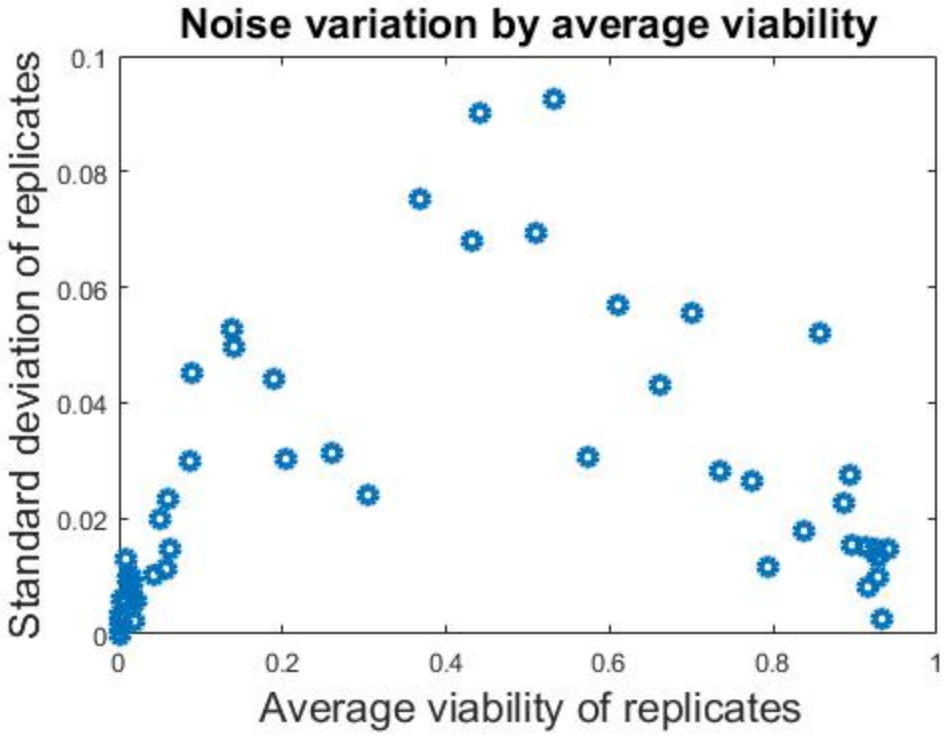

Figure 2.9 Data from the model validation experiment is used to estimate sources of error. The standard deviation among each set of replicate measurements is graphed against its average value, with each set of replicates represented as a single point.

Tables

	Treated			Untreated		
	lower bound	**value**	**upper bound**	**lower bound**	**value**	**upper bound**
LD50	48.0	50.4	52.4	33.5	37.0	40.5
slope	0.038	0.044	0.049	0.045	0.0577	0.071

Table 2.4 Parameter values for the single static population model (LD50 values are in units of µM doxorubicin.)

	Treated			Untreated		
	lower bound	**value**	**upper bound**	**lower bound**	**value**	**upper bound**
LD50 $_{t = 0wk}$	40.1	43.6	47.1	40.2	43.6	47.2
LD50 $_{t = 1wk}$	39.6	45.0	50.4	31.2	36.4	41.6
LD50 $_{t = 2wk}$	57.1	67.1	77.1	35.4	41.4	47.4
LD50 $_{t = 3wk}$	47.9	53.9	60.0	27.4	31.0	34.7
LD50 $_{t = 4wk}$	44.8	50.9	57.0	32.9	37.7	42.4
LD50 $_{t = 5wk}$	43.8	50.2	56.5	31.6	37.1	42.6
LD50 $_{t = 6wk}$	43.8	49.8	55.7	33.7	38.2	42.6
LD50 $_{t = 7wk}$	43.6	48.9	54.3	37.1	41.8	46.4
LD50 $_{t = 8wk}$	40.8	46.2	51.7	32.1	36.1	40.1
slope $_{t = 0wks}$	0.039	0.052	0.065	0.039	0.052	0.065
slope $_{t = 1wks}$	0.031	0.051	0.069	0.036	0.057	0.077
slope $_{t = 2wks}$	0.020	0.031	0.043	0.032	0.049	0.067
slope $_{t = 3wks}$	0.026	0.039	0.052	0.046	0.064	0.083
slope $_{t = 4wks}$	0.030	0.045	0.060	0.039	0.057	0.074
slope $_{t = 5wks}$	0.030	0.045	0.060	0.036	0.053	0.070
slope $_{t = 6wks}$	0.029	0.041	0.054	0.043	0.060	0.076
slope $_{t = 7wks}$	0.035	0.049	0.063	0.039	0.056	0.072
slope $_{t = 8wks}$	0.035	0.049	0.064	0.43	0.064	0.084

Table 2.5 Parameter values for the single dynamic model (LD50 values are in units of μM doxorubicin.)

	Treated			Untreated		
	lower bound	**value**	**upper bound**	**lower bound**	**value**	**upper bound**
$LD50_{res}$	73.2	79.7	86.3	73.2	79.7	86.3
$LD50_{sens}$	19.2	22.4	25.6	19.2	22.4	25.6
$slope_{res}$	0.037	0.043	0.050	0.037	0.043	0.050
$slope_{sens}$	0.073	0.115	0.158	0.03	0115	0.0158
$frac_{res\ t=0wk}$	0.379	0.497	0.626	0.379	0.497	0.626
$frac_{res\ t=1wk}$	0.406	0.507	0.608	0.280	0.381	0.483
$frac_{res\ t=2wks}$	0.597	0.733	0.869	0.356	0.464	0.572
$frac_{res\ t=3wks}$	0.500	0.597	0.693	0.174	0.270	0.367
$frac_{res\ t=4wks}$	0.473	0.585	0.697	0.286	0.392	0.500
$frac_{res\ t=5wks}$	0.460	0.563	0.667	0.279	0.381	0.483
$frac_{res\ t=6wks}$	0.448	0.548	0.648	0.310	0.406	0.502
$frac_{res\ t=7wks}$	0.449	0.551	0.653	0.350	0.464	0.578
$frac_{res\ t=8wk}$	0.395	0.499	0.603	0.256	0.361	0.465

Table 2.6 Parameter values for the two-population model (LD50 values are in units of µM doxorubicin.)

Additional Analysis/Discussion

Discussion of Intrinsic Stochastic Biological Variability and Sources of Error

While we often look at variation in measured data and mentally summarize it as "measurement error", the controls present in our model validation experiment allow us to decompose this variation into several sources of error. Error in the viability measurement performed using the Nexcelom Cellometer is approximately 0.5%, estimated using replicate measurements drawn from a single pool of cells. The variation in viability is larger, at an average of 2.2% across all samples, but is also non-uniformly distributed, as shown in Figure 2.9.

Part of this variation can be explained as the result of differences in the proportion of cells of each subline plated into a given well; this variation has a standard deviation of approximately 1.7%. In Figure 2.9, however, the standard deviation ranges as high as 9%,

indicating that an additional source of variation is present, and appears to be dependent on the population viability for that set of replicate measurements. We believe that this is best described as an actual variability in the response that replicate populations will display when given the same stimulus, due to downstream effects of stochastic variation in the early response. Among other things, we are aware that confluence has a large effect on cell survival – as a result, random fluctuations in early survival could snowball into these substantial differences over the duration of the drug perturbation. As each cell dies, it increases the probability that other nearby cells will die. (We speculate that this is mediated by the loss of pro-survival signals which the cells exchange.) Any fluctuation of survival in the early stages of the response is then propagated and amplified, magnifying the fluctuations into the pattern of variation seen here in supplemental Figure 2.9. This theory is consistent with the non-uniform distribution of variability; doses that are high or low enough to have more deterministic effects show minimal variation, more comparable to the error known to be present from variation in the initial seeding proportions and technical error in the assay while in doses where the probability of death is closer to 50%, a marginal change in the probability of survival is more likely to influence the outcome.

[3]Chapter 3: Stochastic parameter estimation to reveal an Allee effect in tumor growth

PREFACE

This work provides an illustrative example of one way in which we can utilize technological advances in data acquisition to improve our mechanistic understanding of biological phenomena. In this instance, we were able to acquire data at a precision level of single-cell resolution cell counting capabilities via the Incucyte and were able to seed cells at precise low initial cell numbers using Fluorescence Activated Cell Sorting. Combining these two advances in a high throughput manner enabled for data collection in which we could achieve very high time resolution data for a large number of replicates- thus giving us the ability to capture the variability in growth dynamics. We harnessed this capability by combining it with mathematical modeling; namely the ability to derive explicitly, from a mechanistic model of tumor cell growth dynamics, the expected mean and variability in the observed data for the stochastic growth process being studied. This work represents one neat way in which new technological advances can be met with existing mathematical frameworks for improved mechanistic understanding. Potential extensions of this work span both further biological investigation of cooperative interactions that give rise to observed Allee effects, but also more broadly the further development of analytical tools

[3] Note: This chapter is based on an article originally published as:
Johnson, K.E., Howard, G.R., Mo, W., Strasser, M.K. Lima, E. A. B. F., Hunag, S., & Brock, A. (2019). Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an Allee effect. PLoS Biol 17(8):e3000399. https://doi.org/10.1371/journal.pbio.3000399
Author contributions:
Conceptualization: Sui Huang, Amy Brock,.Data curation: Kaitlyn E. Johnson, Grant Howard, William Mo, Formal analysis: Kaitlyn E. Johnson., Funding acquisition: Sui Huang, Amy Brock, Investigation: Kaitlyn E. Johnson, Methodology: Kaitlyn E. Johnson, Grant Howard, Michael K. Strasser, Ernesto A. B. F. Lima, Sui Huang, Amy Brock, Project administration: Amy Brock, Supervision: Amy Brock, Visualization: Kaitlyn E. Johnson, Writing – original draft: Kaitlyn E. Johnson, Amy Brock, Writing – review & editing: Michael K. Strasser, Sui Huang, Amy Brock.

in which we can utilize the full distribution of data for model fitting- replacing fitting the average of data set to fitting for a full distribution. A number of mathematical techniques, including the one presented in this work, offer ideas as to how this can be done for new sets of problems.

ABSTRACT

Most models of cancer cell population expansion assume exponential growth kinetics at low cell densities, with deviations to account for observed slowing of growth rate only at higher densities due to limited resources such as space and nutrients. However, recent preclinical and clinical observations of tumor initiation or recurrence indicate the presence of tumor growth kinetics in which growth rates scale positively with cell numbers. These observations are analogous to the cooperative behavior of species in an ecosystem described by the ecological principle of the Allee effect. In preclinical and clinical models, however, tumor growth data are limited by the lower limit of detection (i.e., a measurable lesion) and confounding variables, such as tumor microenvironment, and immune responses may cause and mask deviations from exponential growth models. In this work, we present alternative growth models to investigate the presence of an Allee effect in cancer cells seeded at low cell densities in a controlled in vitro setting. We propose a stochastic modeling framework to disentangle expected deviations due to small population size stochastic effects from cooperative growth and use the moment approach for stochastic parameter estimation to calibrate the observed growth trajectories. We validate the framework on simulated data and apply this approach to longitudinal cell proliferation data of BT-474 luminal B breast cancer cells. We find that cell population growth kinetics are

best described by a model structure that considers the Allee effect, in that the birth rate of tumor cells increases with cell number in the regime of small population size. This indicates a potentially critical role of cooperative behavior among tumor cells at low cell densities with relevance to early stage growth patterns of emerging and relapsed tumors.

INTRODUCTION

The classical formulation of tumor growth models often begins with the assumption that early stage tumor growth dynamics are driven by cell-autonomous proliferation, manifested as an exponential increase in cell number. The exponential growth model describes a growth rate that is proportional to the number of cells present and is often captured by a single growth rate constant at this stage. However, current imaging technologies have a lower limit of detection of about 1 million cells on a typical CT scan (Kobayashi *et al.*, 2017), and thus measurements of the growth dynamics of very small tumor cell populations are not typically captured in the clinical setting (Kobayashi *et al.*, 2017). Recent findings in preclinical mouse models (Panigrahy *et al.*, 2012) and from clinical outcomes following tumor resection (Neufeld *et al.*, 2017) reveal that tumor growth at low tumor cell densities does not match the expectation of exponential growth. In addition, observations of in vitro cell growth have long recognized that very low seeding density may have a detrimental effect on population fitness. These findings give rise to an intriguing possibility: does tumor cell growth deviate from the model of exponential growth at low tumor cell densities? In this study, we ask whether early stage tumor growth kinetics exhibits a behavior analogous to a principle in ecology known as the Allee effect,

in which the fitness of a population, measured by the per capita growth rate, scales with population size at low population sizes. In ecology, the Allee effect arises due to cooperative growth, such as cooperative predation, feeding, and mating systems(Courchamp, Berec and Gascoigne, 2008). In tumors, there exists an abundance of evidence for subclonal interactions among cells, e.g., with specific subpopulations releasing signaling molecules critical to the growth of other subsets of cells(Scheel *et al.*, 2011; Cleary *et al.*, 2014; Marusyk *et al.*, 2014; Archetti, Ferraro and Christofori, 2015; Kumar *et al.*, 2018). Thus, it is quite intuitive that cancer cell growth may exhibit cooperative interactions analogous to the cooperation among species in an ecosystem.

The ability to describe and predict tumor growth is essential to developing strategies to eradicate cancer cell populations (Yankeelov *et al.*, 2016; Matthew T. McKenna, Weis, Brock, *et al.*, 2018). Understanding tumor growth kinetics at low cell numbers is of clinical importance because they govern tumor initiation, treatment response, and recurrence. In ecology, the Allee effect has informed strategies for the control of invasive species (Cloonan *et al.*, 2008) and has been used to predict how an introduced species might take hold in a new environment (Courchamp, Berec and Gascoigne, 2008). Applying ecological principles to control tumor growth is a growing interest (Gatenby, 1991; Mcgregor, Axelrod and Axelrod, 2008; Chen and Pienta, 2011; Basanta *et al.*, 2013; Korolev, Xavier and Gore, 2014; Amend and Pienta, 2015; Amend *et al.*, 2016; Han *et al.*, 2016; Axelrod and Pienta, 2018; Kaznatcheev *et al.*, 2019; Kimmel *et al.*, 2019). A better understanding of the factors that govern tumor cell growth at early stages could help to

improve predictions of initial growth, relapse, and metastasis, as well as guide therapeutic strategies borrowed from ecological principles to control tumor progression.

Although exponential growth is a common initial assumption used to develop more complex models of tumor progression, few models strictly interested in characterizing tumor growth prescribe a fixed birth and death rate over time and population size. Many modifications have been made to account for a changing growth rate as the population grows. The widely used Gompertzian model is a phenomenological model that introduces a growth rate that decays exponentially with time (Winsor, 1932; Benzekry *et al.*, 2014; Pacheco, 2016). Similarly, the logistic growth model exhibits a modification that introduces a population size dependency, slowing the growth rate as the population size approaches carrying capacity (Benzekry *et al.*, 2014; Pacheco, 2016). This can be explained mechanistically as a result of competition over finite space and nutrients. Further refinements to these models have used a statistical mechanics framework to explicitly introduce a correlation function, reducing the number of accessible growth states of individual cells as the population size increases, leading to slowing of growth (West and Newton, 2018). Other mechanistic models have used a replicator system of equations for competing species, with population-dependent fitness based on a payoff matrix (Gerlee and Altrock, 2017). Stochastic models of tumor growth have also been used to describe tumor growth based on a Moran birth–death process, a stochastic model that describes how heterogeneity increases over time due to molecular mutations in individual cells (West *et al.*, 2016). This stochastic modeling framework leads to a population-dependent fitness landscape that exhibits nonconstant tumor growth rates; specifically, tumor growth rates

that slow in the later stages of development (West *et al.*, 2016). Although all of these models take into account decreasing growth rates as the population grows(Winsor, 1932; Speer *et al.*, 1984; Norton, 1988; Benzekry *et al.*, 2014; Pacheco, 2016; West *et al.*, 2016; West and Newton, 2018), none explicitly investigate the opposing effect of an increase in growth rate with population size at low population densities.

Recent evidence of deviations from exponential growth at early stages of tumor growth have been observed in glioblastoma, in which patient brain tumors were resected and monitored over time for relapse (Neufeld *et al.*, 2017). These studies of relapsed tumor growth revealed that the observed growth rate at the clinically detectable stages of tumor growth failed to match models of simple logistic growth and instead were better described by a weak Allee effect model. In the weak version of the Allee effect, populations grow at a much slower rate at very low tumor cell numbers but continue to grow for any initial population size. By contrast, a strong Allee effect describes a population that becomes extinct below a threshold initial population size. In ecology, both strong and weak Allee effects are observed (Courchamp, Berec and Gascoigne, 2008). Although the observation of a weak Allee effect in glioblastoma recurrence is certainly provocative, it is limited by the fact that the earliest stages of tumor growth from low cell densities cannot be easily detected in vivo and thus the critical measurements at the relevant regime cannot be captured with current imaging technologies. Numerous studies have investigated the manifestation of the Allee effect in ecology (Duncan *et al.*, 2014; Wittmann, Gabriel and Metzler, 2014; Vieira, Ribeiro and Souto, 2015; Bose *et al.*, 2017) and a few have posed theoretical implications and possible mechanisms of cooperative kinetics of the Allee effect

in cancer growth (Rodriguez-brenes, Komarova and Wodarz, 2013; Böttger, Hatzikirou and Voss-böhme, 2015; Greene *et al.*, 2016; Konstorum, Hillen and Lowengrub, 2016; Sewalt *et al.*, 2016). However, none have performed an in-depth quantitative analysis of cancer cell proliferation kinetics captured in the low cell density regime. The explicit investigation of the Allee effect in describing tumor growth dynamics at low population sizes is the main contribution of this paper.

In this study, we investigate the behavior of various structurally distinct models of tumor growth representing alternative hypotheses of growth dynamics that consider the Allee effect. We present a framework for the analysis of cancer cell growth at low cell densities in a controlled in vitro setting in which cells are subject to optimal growth conditions with sufficient nutrients and space. Monitoring growth in vitro allows for studying the effects of cell number on growth in the absence of confounding factors, such as the immune system interactions and tissue microenvironmental factors, in order to test explicitly the dependence of growth dynamics on cell density. We take advantage of recent technological advances that allow for the seeding of a precise small initial cell number and the ability to measure cell number at single-cell resolution and at high-temporal resolution. This enables capturing of accurate growth kinetics in the low cell-density regime in which the Allee effect is most relevant and cannot be studied in vivo. Because we focus our examination in the low cell-density regime exclusively (<200 cells in a 1 mm^3 well), our modeling analysis excludes additional terms that describe the slowing of cancer cell population growth at higher densities in which competition for limited resources and space becomes relevant. We examine the average behavior of 3 models of increasing complexity:

the exponential growth model, a strong Allee model, and an extended Allee model that can be either strong or weak.

At the small population size of interest in this study, the inherent stochasticity of the birth–death problem leads to a nonzero probability of extinction, even for a model of constant, net-positive birth rate minus death rate. This phenomena, in which the average population behavior appears to have a reduced growth rate because some trajectories become extinct and have a growth rate of zero, is known in ecology as demographic stochasticity (Courchamp, Berec and Gascoigne, 2008). In order to disentangle these stochastic effects of small population sizes that decrease observed average growth rate from true cooperative effects, we develop 7 stochastic models whose average behavior follows one of the deterministic models. In this framework, each stochastic model represents a different hypothesis of the mechanism underlying the growth kinetics. For each stochastic model, we perform a parameter estimation using the method of moments (Fröhlich *et al.*, 2016) and use model selection to identify the model most likely model to describe the growth data (Fröhlich *et al.*, 2016). This is performed for both a simulated data set and the in vitro BT-474 breast cancer cell line data to test the hypotheses that our framework reveals an alternative tumor growth model that incorporates an Allee effect.

MATERIALS AND METHODS

Data processing and aanlysis

All mathematical modeling and analysis was performed in MATLAB. Code and data for all analysis is available on Github at: https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

Cell culture and low cell density seeding

The human breast cancer cell line BT-474 was used throughout this study. BT-474 is a standard cell line from ATCC. Cell lines were maintained and studied in Dulbecco's Modified Eagle Medium (DMEM, Thermo Fischer Montreal, Canada) supplemented with insulin (Gibco Gaithersburg, MD) and 10% fetal bovine serum (Gibco) and 1% Penicillin-Streptomycin (Gibco Gaithersburg, MD Gaithersburg, MD). A subline of the BT-474 breast cancer cell line was engineered to constitutively express enhanced green fluorescent protein (EGFP) with a nuclear localization signal (NLS). Genomic integration of the EGFP expression cassette was accomplished through the Sleeping Beauty transposon system (Kowarz, Loescher and Marschalek, 2015). The EGFP-NLS sequence was obtained as a gBlock from IDT and cloned into the optimized Sleeping Beauty transfer vector psDBbi-Neo. pSBbie-Neo was a gift from Eric Kowarz (Addgene plasmid #60525) (Kowarz, Loescher and Marschalek, 2015). To mediate genomic integration, this two-plasmid system consisting of the transfer vector containing the EGFP-NLS expression cassette and the pCMV(CAT)T7-SB100 plasmid containing the Sleeping Beauty transposase was co-transfected into a BT-474 cell population using Lipofectamine 2000. mCMV(CAT)T7-SB100 was a gift from Zsuzsanna Izsvak (Addgene plasmid # 34879) (Mátés *et al.*, 2009). GFP+ cells were collected by fluorescence activated cell sorting. BT-474-EGFPNLS1 cells

are maintained in DMEM (Gibco Gaithersburg, MD) supplemented with insulin (Sigma Life Science St. Louis, MO), 10% fetal bovine serum (Fisher), and 200 μg/mL G418 (Caisson Labs Smithfield, UT). Cells were grown in precoated culture dishes at 37 °C in a humidified, 5% CO_2, 95 air atmosphere. Cells were seeded into the center 60 wells of a 96-well plate (Trueline Saint-Anne-de-Bellevue, Quebec, Canada) at precise initial cell numbers using fluorescence activated cell sorting (BD Fusion Franklin Lakes, NJ) plate sorting at single-cell precision. Plates were kept in the Incucyte Zoom, a combined incubator and time-lapsed microscope. Initial cell seeding numbers were verified by eye at 4× magnification using an image taken within 4 hours from the FACS seeding. Low cell density cultures were allowed to grow in media for 7 days and were subsequently fed fresh media every 2 to 3 days for up to 2 weeks.

Time-lapse imaging

Time-lapse recordings of the cell cultures were performed using the whole-well imaging feature in the Incucyte Zoom (Essen Biosceince Ann Arbor, MI). Cells were maintained in the Incucyte at 37°C in humidified 5% CO_2 atmosphere. Phase contrast and green-channel images were collected every 4 hours for up to 2 weeks.

Image analysis

Recorded green-channel images were analyzed using the built-in analysis program in the Incucyte Zoom (Essen Bioscience Ann Arbor, MI) software analysis package. The true initial cell number of each well was confirmed by eye from the images at 4× magnification, and cell-number trajectories were binned accordingly. For each 96-well

plate, an image processing definition was optimized using the built-in software and confirmed by eye to account for background fluorescence and local bubbles. Wells whose cells died off or did not exhibit any growth and wells without of focus images were removed from analysis.

RESULTS

The deterministic strong and weak Allee effect models

This work studies stochastic growth models because of the inherent stochasticity of cell growth processes in the regime of small cell numbers that is the focus of our work. However, the structure of the functional forms of the kinetic equations describing the cell-number changes can be understood within a deterministic framework, thus providing a link to the historical and most widely implemented tumor growth models. The deterministic models involving the Allee effect are chosen because they have previously shown to be useful for applied ecologists working in regimes in which the Allee effect is relevant (Courchamp, Berec and Gascoigne, 2008).

At the core of our modeling effort are the following 3 deterministic phenomenological models of increasing complexity that describe cell population growth kinetics. The first model represents the null model of tumor growth (Pacheco, 2016) where the growth rate ($\frac{\partial N}{\partial t}$) is proportional to the number of cells present, N, and a single growth rate constant, g, resulting in the classical exponential growth model:

$$\frac{dN(t)}{dt} = gN(t) \quad (1)$$

$$N(t) = N_0 e^{gt} \quad (2)$$

This model describes N cells that exhibit cell autonomous proliferation (Fig 1A) and a constant per capita growth rate ($\frac{\frac{dN}{dt}}{N}$) given by the growth rate constant g over time and cell number (Fig 3.1B) for initial cell numbers $N(t = 0) = 3$, 8, and 16 cells displayed in (Fig 3.1A, 3.1B, and 3.1C). In the remainder of the manuscript we denote the initial cell number, $N(t = 0)$ by N_0. The normalized growth rate ($log(\frac{N(t)}{N_0})$ is constant for each initial condition, with all growth curves falling on a line of equal slope (Fig 3.1C). Eq. 1 and 2 represent the well-known exponential growth model and the simplest of the tumor growth models analyzed.

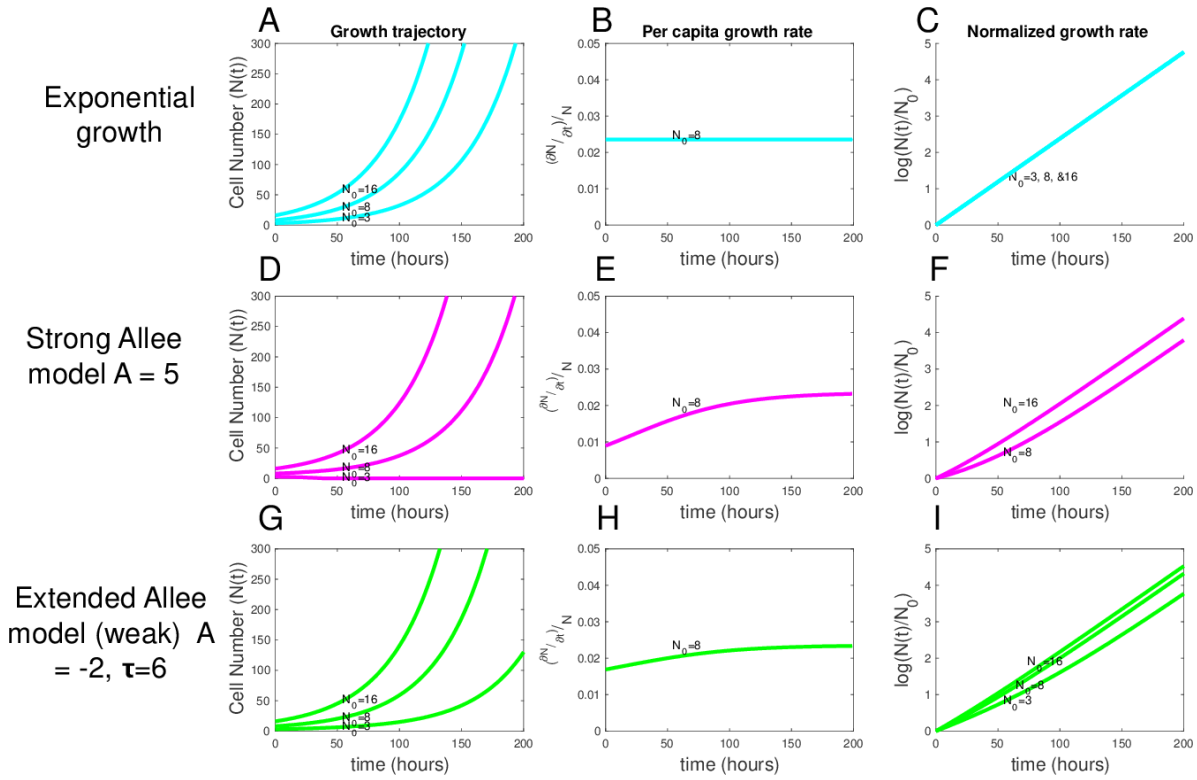

Fig 3.1: Average behavior of exponential, strong, and weak Allee models for different initial conditions. (A, D, and G) Deterministic growth curves of the exponential growth model (blue), the strong Allee model (pink), and the weak Allee model (green), respectively, shown for N_0 = 3, 8, and 16 for all models. (B, E, and H) Per capita growth rates demonstrate that growth rate increases in time with cell number for both Allee models. (C, F, and I) For normalized cell numbers, a clear difference is observed in the slopes depending on the initial cell number for both Allee models. Data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

Most departures from the exponential growth model of cancer cells (Eq. 1 and 2) describe cancer cell growth in which the growth rate is proportional to the number of cells present but with modifications that account for slowing of growth over time and/or with increasing population size. For example, in the classical formulation of the logistic growth model (Pacheco, 2016), the growth rate is characterized by a growth rate constant *g*

modulated by an additional term to describe the slowing of growth rate as the population approaches carrying capacity (*K*):

$$\frac{dN(t)}{dt} = g(1 - \frac{N(t)}{K})N(t) \quad (3)$$

The logistic growth model describes cells in 2 regimes: when $N << K$ the $\frac{N}{K}$ term is negligible and the cells essentially exhibit exponential growth, and when *N* is near *K*, the net growth rate ($\frac{dN}{dt}$) slows towards zero as *N* approaches *K* and the $1 - \frac{N}{K}$ term approaches zero.

We present the logistic growth model (Eq. 3) to demonstrate that the second model equation (Eq. 4), the strong Allee model, is analogous to the logistic growth model, except that the dependency on *N* in this model occurs in the opposite regime—introducing an Allee effect term of $1 - \frac{A}{N}$ that lowers the observed growth rate at small *N* near the Allee threshold *A*:

$$\frac{dN(t)}{dt} = g(1 - \frac{A}{N(t)})N(t) \quad (4)$$

This model describes *N* cells whose net growth rate exists in 2 distinct regimes: when *N* is less than the Allee threshold (*A*), the Allee effect term $1 - \frac{A}{N}$ in Eq. 4 becomes negative and the net growth rate ($\frac{dN}{dt}$) becomes negative, predicting the population will ultimately go extinct (Fig 1D; $N_0 = 3$). When $N(t)$ is near *A* but larger than *A*, the net growth rate is slowed by a factor of $1 - \frac{A}{N}$ (Eq. 4) but remains positive, resulting in a growth rate that scales with cell number, as can be seen for the per capita growth rate over time for $N_0 = 8$ (Fig 1E).

When $N(t)$ is much larger than A, the Allee effect term $(1 - \frac{A}{N})$ becomes negligible and the cell population begins to behave like in the exponential growth model (Fig 1D and 1E). This behavior in which a population is predicted to go extinct below a critical threshold (here A) describes a strong Allee effect. The expected scaling of the normalized growth rate ($log(\frac{N(t)}{N_0})$) demonstrates the expected differences in net growth rate based on initial seeding number (N_0) for a strong Allee model (Fig 1F). As expected, only initial conditions greater than the Allee threshold of $A = 5$ (corresponding to $N_0 = 8$ and 16) result in a net positive growth rate. This model is able to explain the threshold-like behavior observed in preclinical studies of engrafted tumors in mouse (Panigrahy *et al.*, 2012), where, below a threshold number of inoculated cells, tumors never form. To account for weak Allee effect behavior, in which the growth rate is always greater than zero for any N_0, we introduce the third deterministic model, the extended Allee model:

$$\frac{dN(t)}{dt} = g(1 - \frac{A+\tau}{N(t)+\tau})N(t) \quad (5)$$

This model is similar to the strong Allee model (Eq. 4) but introduces an additional parameter τ that allows the model to exhibit either a strong Allee effect when A is positive, or a weak Allee effect when $\tau > |A|$ and $A < 0$. When weak Allee conditions hold, at low N the $(\frac{A+\tau}{N+\tau})$ term always remains less than 1 but greater than zero, keeping the net growth rate positive but resulting in a growth rate that approaches zero as N decreases. Fig 3.1G, 3.1H, and 3.1I display the behavior of the extended Allee model with parameters that produce a weak Allee effect. The extended Allee model explains potential weak Allee effects, such as those observed in glioblastoma resection (Neufeld *et al.*, 2017). See Table

3.3 for a complete description of each of the 3 deterministic models, their parameters, and their behaviors.

Extension to stochastic growth models

Given that the growth kinetics are measured here in very small cell populations, where the expected variability of individual cell behavior with respect to division "birth" and "death" events (which jointly determine net growth rate $\frac{dN}{dt}$) is high, this scenario can give rise to apparent growth kinetics that deviate significantly from the average population behavior. In order to detect slowing of growth that is not due to stochastic small population effects that result in reduced average observed growth, a stochastic modeling framework was implemented to test the relevance of the Allee effect models presented above. Stochastic models are often derived from microscopic models that describes density-dependent birth and death rates. However, in this approach, we chose the expressions for our stochastic models in order to recapitulate a first-order moment (mean) that matched the corresponding deterministic ordinary differential equation (ODE) of the Allee effect. Birth and death rates were thus chosen not directly from a first-principles derivation of a microscopic model but based on reasonable hypotheses consistent with the Allee effect model behavior.

We developed 7 stochastic models whose expected mean cell number in time $\langle n(t) \rangle$ are equivalent to that predicted by the deterministic models described above (Eq. 1, 4, and 5). For each deterministic model structure, the total number of cells $n(t)$ is modeled by the ODEs in Eq. 1, 4, and 5 above. The time evolution of $n(t)$ for the stochastic models are defined by the following birth and death events (Fig 3.2A):

$$\text{Event 1:birth} C \rightarrow 2\text{C}(\text{reaction rate: } r_{\text{birth}}(n))$$

$$\text{Event 2:death} C \rightarrow \emptyset(\text{reaction rate: } r_{\text{death}}(n))$$

Where $r_{\text{birth}}(n)$ and $r_{\text{death}}(n)$ describe the rate at which the events occur, which may depend on the number of cells, n, present or be constant. The probability of an event i happening in an infinitesimal time step Δt is given generally by the product of the rate of the event, the state of the population, and the time step:

$$P_{event} = r_{event}(n) n \Delta t \tag{6}$$

For our purposes in modeling tumor growth, we limit the possible events to birth or death events, and in all cases, the probability of an event occurring is a function of n, because it is a first-order reaction described by the schematic in (Fig 2A).

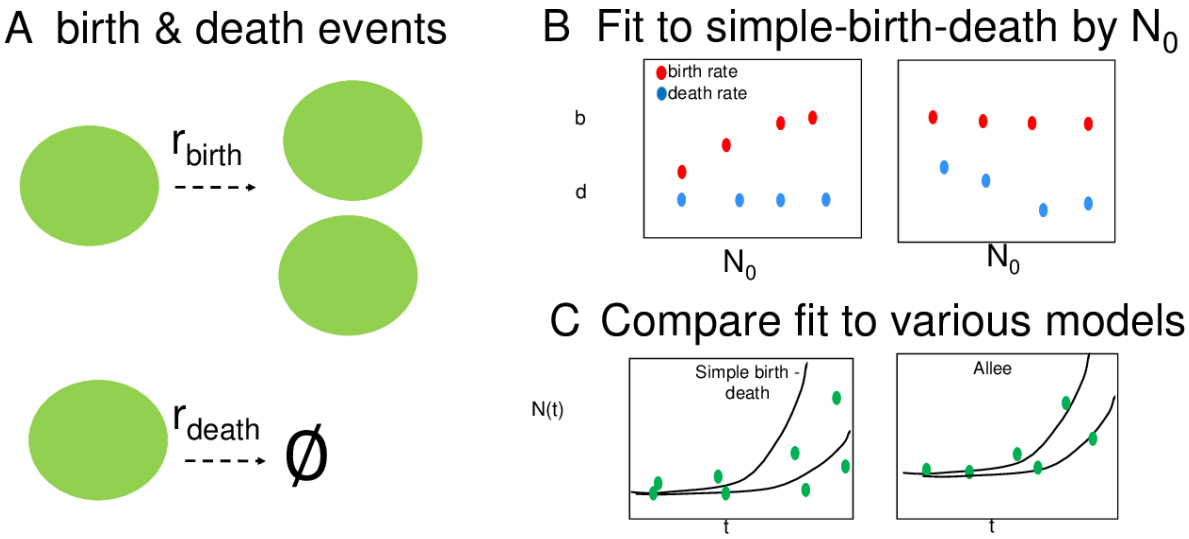

Fig 3.2: A stochastic model of tumor growth and expected outputs if Allee effect is present. (A) Schematic illustration of generalized stochastic framework in which a cell can either give birth or die at a rate given by r_{birth} or r_{death} respectively. (B) Schematic of the expected results from fitting of the simplest birth–death model with each data set grouped by initial cell number (N_0), where, if an Allee effect is present, we expect to observe that either the birth rate constant (b) (red) or death rate constant (d) (blue) change with the initial condition. (C) Schematic of the expected outcomes of fitting full data set to the simple birth–death model (left) and a model incorporating an Allee effect (right).

For each model presented, the generalized framework described above holds and only the birth and death rates (r_{birth} and r_{death}) differ for each model based on the hypothesis about the dependency of the birth or death rate on cell number. To give an illustrative example of the components of the stochastic model, we explicitly state the reaction rates and the resulting birth and death probabilities for the simple birth–death model. To remain concise, for the remaining 6 Allee models described, we just present the birth and death probabilities for each model.

In the simple birth–death model, the birth rate and death rates are independent of cell number, n. They are described by rate constants, denoted b and d:

$$r_{birth} = b \tag{7}$$

$$r_{death} = d \tag{8}$$

And birth and death probabilities in time step Δt of

$$P_{birth} = bn\Delta t \tag{9}$$

$$P_{death} = dn\Delta t \tag{10}$$

The average behavior in this model corresponds to the exponential growth model (Eq. 1 and 2), where the growth rate constant g is equivalent to the birth rate constant (b) minus the death rate constant (d) ($g = b - d$). For the remaining stochastic models we introduce birth and/or death rates that are functions of n, corresponding to the hypotheses that the birth and/or death rates are not constant and instead depend on the population size, n. Similar to the way that stochastic growth models have modified the growth rate with a term that decreases growth rate in proportion to increasing n (i.e., in the work by Sun and colleagues (S. Sun, 2015), where the division rate k is defined as $k = k_0 - \gamma n$), we prescribe modifications to birth and death rates that decrease birth rates proportional to the reciprocal of n. We note that this nonlinear dependency is prescribed as such in order to achieve the desired slowing of growth rates at low n that the Allee effect models produce (Eq. 4 and 5). The first stochastic Allee model is the strong Allee on birth model.

$$P_{birth} = (b - \frac{A}{n}(b - d))n\Delta t \tag{11}$$

$$P_{death} = dn\Delta t. \tag{12}$$

This model hypothesizes that the birth probability is lowered by a factor proportional to the growth rate ($b - d$) and the reciprocal of n, and thus for large n, the $\frac{A}{n}$ term in Eq. 12 is

negligible, but at small n the birth probability is significantly decreased by the Allee term, resulting in a lower birth probability and observed slower net growth.

Alternatively, we can hypothesize that the Allee effect acts to increases the death probability at n near A, resulting in the strong Allee on death model probabilities of

$$P_{birth} = bn\Delta t; \tag{13}$$

$$P_{death} = (d + \frac{A}{n}(b - d))n\Delta t. \tag{14}$$

And lastly, we present a model that assumes that the Allee effect term acts equally on both decreasing the birth probability and increasing the death probability for n near A, resulting in

$$P_{birth} = (b - \frac{A}{n}\frac{(b-d)}{2})n\Delta t; \tag{15}$$

$$P_{death} = (d + \frac{A}{n}\frac{(b-d)}{2})n\Delta t. \tag{16}$$

For simplicity, this model assumes that the Allee term acts equally, with half of its effect decreasing the birth rate and half increasing the death rate at n near A. Of course, there could be an infinite number of ways of distributing the Allee threshold onto the birth and death probabilities, and this could have been introduced with an additional fractional parameter. However, for simplicity, we only consider equal partitioning of the Allee effect on both birth and death rates.

The last family of stochastic models corresponds to the extended Allee model (Eq. 5). Again, this model introduces birth and death rate dependencies on n. By the same arguments described for the strong Allee effect model, the extended Allee effect model can

manifest itself either on the birth probability only, the death probability only, or the birth and death probabilities equally, leading to the following birth and death probabilities.

If the Allee effect acts on birth only:

$$P_{birth} = (b - (b - d)\frac{A+\tau}{n+\tau})n\Delta t; \tag{17}$$

$$P_{death} = dn\Delta t. \tag{18}$$

If the Allee effect acts on death only:

$$P_{birth} = bn\Delta t; \tag{19}$$

$$P_{death} = (d + (b - d)\frac{A+\tau}{n+\tau})n\Delta t. \tag{20}$$

If the Allee effect term acts on birth and death equally:

$$P_{birth} = (b - \frac{(b-d)}{2}\frac{A+\tau}{n+\tau})n\Delta t; \tag{21}$$

$$P_{death} = (d + \frac{(b-d)}{2}\frac{A+\tau}{n+\tau})n\Delta t. \tag{22}$$

A complete description of each of the above 7 stochastic models grouped by the corresponding deterministic model and their assumptions of birth or death mechanism is displayed in Table 3.1.

Exponential model family	Strong Allee model family	Extended Allee model family
Mean cell-number change expressed as a deterministic ODE of each family of stochastic models		
$\frac{dN(t)}{dt} = gN(t)$	$\frac{dN(t)}{dt} = g(1 - \frac{A}{N(t)})N(t)$	$\frac{dN(t)}{dt} = g(1 - \frac{A+\tau}{N(t)+\tau})N(t)$
Probabilities of birth and death events to describe stochastic models within each family		
	Allee effect on birth rate	**Allee effect on birth rate**
	$P_{birth} = \Delta t(bN - (b-d)A)$ $P_{death} = \Delta t(dN)$	$P_{birth} = \Delta t(bN - (b-d)N(\frac{A+\tau}{N+\tau}))$ $P_{death} = \Delta t(dN)$
	Allee effect on death	**Allee effect on death**
$P_{birth} = \Delta t(bN)$ $P_{death} = \Delta t(dN)$	$P_{birth} = \Delta t(bN)$ $P_{death} = \Delta t(dN + (b-d)A)$	$P_{birth} = \Delta t(bN)$ $P_{death} = \Delta t(dN + (b-d)N(\frac{A+\tau}{N+\tau}))$
	Allee effect on birth and death	**Allee effect on birth and death**
	$P_{birth} = \Delta t(bN - (\frac{(b-d)}{2})A)$ $P_{death} = \Delta t(dN + (\frac{(b-d)}{2})A)$	$P_{birth} = \Delta t(bN - N(\frac{b-d}{2})(\frac{A+\tau}{N+\tau}))$ $P_{death} = \Delta t(dN + N(\frac{b-d}{2})(\frac{A+\tau}{N+\tau}))$

Table 3.1: Stochastic growth model families whose average behavior correspond to one of the deterministic growth models. For the Allee model families, within each family, the Allee effect can alter birth, death, or both probabilities, representing distinct mechanistic hypotheses. **Abbreviation:** ODE, ordinary differential equation

To simulate growth trajectories of the stochastic models, we use the Gillespie algorithm (Text 3.1) (Gillespie, 1977, 2014). The above models are used to test the relevance of the Allee effect in cancer cell population growth. The conventional exponential growth model (Eq. 1 and 2) assumes that growth rate (birth rate minus death rate) is constant and independent of initial condition. To test the validity of this assumption in an exploratory analysis, we first fit each group of trajectories individually for each initial cell number, N_0. If an Allee effect is present in the data, a systematic increase in the fitted birth rate constant, *b*, or decrease in death rate constant, *d*, with increasing initial cell number would be expected (shown schematically in Fig 3.2B). We next investigated the relevance of the 7 stochastic models by fitting the simulated cell-number trajectories from all initial conditions to each stochastic model described above (Eqs. 9–22) to determine which model structure best describes the observed growth dynamics (shown schematically in Fig 23.C).

Parameter estimation and model selection framework

The parameters of stochastic processes are often inferred using approximate Bayesian computation(Beaumont, Zhang and Balding, 2002), which require exhaustive stochastic simulations in order to minimize the differences between simulation and experimental data for each parameter set searched. These algorithms require a high number of simulating runs, making them computationally expensive and instantiating issues of nonconvergence and model selection (Robert *et al.*, 2011). To render inference on the stochastic process feasible, we apply the moment-closure approximation method described in Frohlich and colleagues (Fröhlich *et al.*, 2016) to fit the 7 proposed stochastic growth models to experimentally measured growth curves (Fig 3.3).

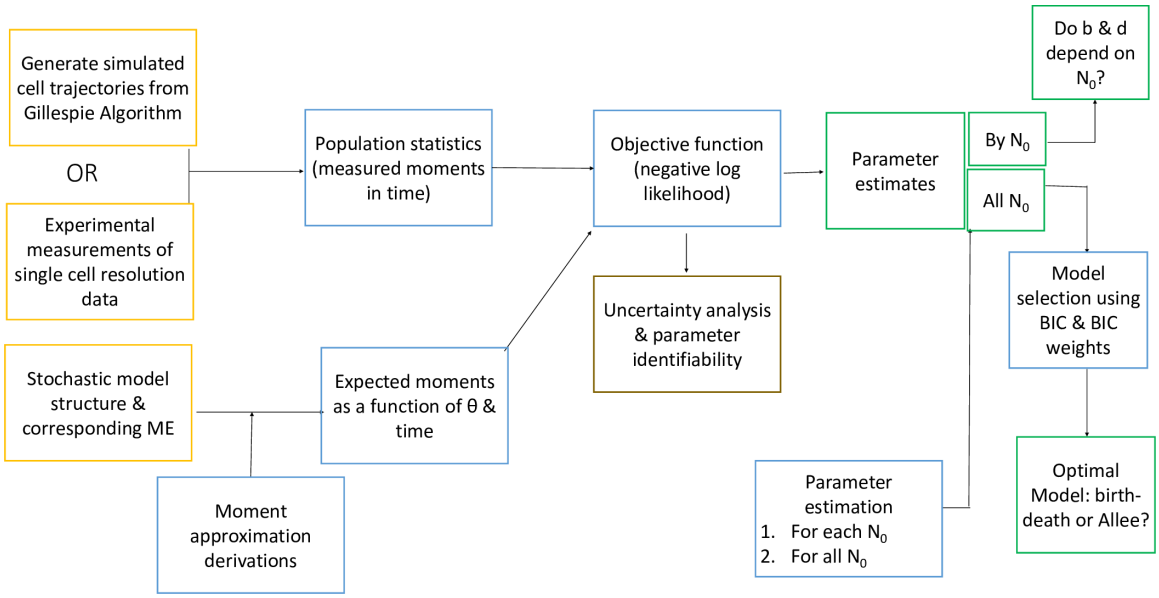

Fig 3.3: Moment-closure approximation approach for stochastic parameter estimation. Framework for moment-closure approximation approach to derive moments from the ME of a stochastic process and how model expected moments are fit to stochastic data. BIC, Bayesian Information Criterion, ME, master equation

The master equation of a stochastic process

The master equation (ME) describes the change in the probability distribution that the system has (in this case number of cells, $\boldsymbol{n}$) as a function of time. From the ME, the time derivative of the moments, or expected values of $\boldsymbol{n}$, $\boldsymbol{n_2}$,…$\boldsymbol{n_m}$ can be derived. In this framework, we developed stochastic models so that the derivative of the first-order moment corresponds to one of the deterministic models presented (Eq. 2, 4, and 5). The ME describes the probability of their being $\boldsymbol{n}$ cells at time $\boldsymbol{t}$ as a sum of probabilities of a birth, death, or neither event occurring given there are $\boldsymbol{n} - 1$, $\boldsymbol{n} + 1$, and $\boldsymbol{n}$ cells at time $\boldsymbol{t}$, respectively:

$$\frac{dp_n(t)}{dt} = r_{birth}(n-1)p_{n-1}(t) - [r_{death}(n) + r_{birth}(n)]p_n(t) + r_{death}(n+1)p_{n+1}(t), \quad (23)$$

where r_{birth} and r_{death} are functions of the parameters b, d, A and/or τ for each of the 7 stochastic structural models (Table 3.4).

Derivations of moment-closure approximations from the master equation

From longitudinal data of cell number over time ($\boldsymbol{N(t)}$), with sufficient replicates, we expect to be able to measure the mean and variance in cell number over time. We want to be able to directly compare the mean and variance in the experimental longitudinal data to the model expected mean and variance in time as a function of the model parameters. We therefore want to derive the first and second moments from the ME. From the ME of each stochastic model (S3.2 Table), the time derivative of the first and second moments were derived according to the procedure outlined in (Houchmandzadeh, 2009)(Text 3.2). Using the definition of variance, the ODEs of the mean and variance for each model can be written in terms of the lower order moments ($\langle \boldsymbol{n} \rangle$ and $\langle \boldsymbol{n^2} \rangle$), where $\langle \dots \rangle$ denotes the expectation value of the moment; Table 3.1). Within each family of models (exponential, strong Allee, and extended Allee; Eq. 2, 4, and 5) the stochastic forms (Eq. 9–22) share the same mean ODE corresponding to their deterministic model family but differ in their variance based on whether the Allee effect alters the birth, death, or both event terms. The time evolution of the variance can be used to properly identify individual rate parameters such as the birth and death rates because the variance in time is proportional not just to the

net growth (birth minus death rates) but to the sum of the birth and death rates, as shown in S3.1 Fig.

For each stochastic model, we confirmed that the mean and variance of a simulated data set of 5,000 trajectories with known parameters matched the derived model mean and variance described in Table 3.2. (See Text 3.3 Text, Figs 3.11-3.17.)

Model	Mean $\langle \boldsymbol{n}(\boldsymbol{t}) \rangle$	Variance $\langle \boldsymbol{\Sigma}(\boldsymbol{t}) \rangle$
Simple birth-death	$\frac{d\langle n \rangle}{dt} = \langle n \rangle (b - d)$	$\frac{d\langle \Sigma_{ii} \rangle}{dt} = 2(b - d)\langle \Sigma_{ii} \rangle + (b + d)\langle n \rangle$
Strong Allee on birth	$\frac{d\langle n \rangle}{dt} = \langle n \rangle (b - d)(1 - \frac{A}{\langle n \rangle})$	$\frac{d\langle \Sigma_{ii} \rangle}{dt} = 2\langle n^2 \rangle (b - d) - 2\langle n \rangle (b - d)A + (b + d)\langle n \rangle - (b - d)A - 2\langle n \rangle (b - d)(\langle n \rangle - A)$
Strong Allee on death	$\frac{d\langle n \rangle}{dt} = \langle n \rangle (b - d)(1 - \frac{A}{\langle n \rangle})$	$\frac{d\langle \Sigma_{ii} \rangle}{dt} = 2\langle n^2 \rangle (b - d) - 2\langle n \rangle (b - d)A + (b + d)\langle n \rangle + (b - d)A - 2\langle n \rangle (b - d)(\langle n \rangle - A)$
Strong Allee on both	$\frac{d\langle n \rangle}{dt} = \langle n \rangle (b - d)(1 - \frac{A}{\langle n \rangle})$	$\frac{d\langle \Sigma_{ii} \rangle}{dt} = 2\langle n^2 \rangle (b - d) - 2\langle n \rangle (b - d)A + (b + d)\langle n \rangle - 2\langle n \rangle (b - d)(\langle n \rangle - A)$
Extended Allee on birth	$\frac{d\langle n \rangle}{dt} = \langle n \rangle (b - d)(1 - \frac{A + \tau}{\langle n \rangle + \tau})$	$\frac{d\langle \Sigma_{ii} \rangle}{dt} = 2\langle n^2 \rangle (b - d) + \langle n \rangle (b + d) - 2\langle n^2 \rangle (b - d)(\frac{A + \tau}{\langle n \rangle + \tau}) - \langle n \rangle (b - d)(\frac{A + \tau}{\langle n \rangle + \tau}) - 2\langle n^2 \rangle (b - d)(1 - \frac{A + \tau}{\langle n \rangle + \tau})$

Table 3.2. Differential equations of the moment-closure approximations of the mean and variance for each stochastic model obtained from the ME using the moment approach. Abbreviation: ME, master equation

Maximum likelihood and Bayesian parameter estimation

To infer the parameters of the stochastic models, a maximum likelihood parameter estimation approach was employed using derivations from Frohlich and

colleagues (Fröhlich *et al.*, 2016). The likelihood function assumes that the measured mean and variance of the data at each time point t_k is normally distributed around the model predicted first moment (mean cell number ($\mu_i(t_k, \theta)$) and mean variance in cell number ($\Sigma_{ii}(t_k, \theta)$) with standard deviations for each distribution of mean cell number and variance in cell number given by $\sigma_{\mu_{i,k}}(\theta)$ and $\sigma_{\Sigma_{ii,k}}(\theta)$respectively. These standard deviations in the first moment and variance, $\sigma_\mu(\theta)$ and $\sigma_\Sigma(\theta)$, are functions of the parameters θ and were derived by Frohlich and colleagues (Fröhlich *et al.*, 2016). The likelihood function (Eq. 24) and its corresponding negative log likelihood (Eq. 25) are the following:

$$L(\theta) = \prod_{i,k} \frac{1}{\sqrt{2\pi\sigma^2_{\mu_{i,k}}(\theta)}} exp\left(-\frac{1}{2}\left(\frac{\mu_i(t_k,\theta)-\hat{u}_{i,k}}{\sigma_{\mu_{i,k}}(\theta)}\right)^2\right) \times \prod_{i,k} \frac{1}{\sqrt{2\pi\sigma^2_{\Sigma_{ii,k}}(\theta)}} exp\left(-\right.$$

$$\left.\frac{1}{2}\left(\frac{\Sigma_{ii,k}(t,\theta)-\widehat{\Sigma}_{ii,k}}{\sigma_{\Sigma_{ii,k}}(\theta)}\right)^2\right); \quad (24)$$

$$NLL(\theta) = \frac{1}{2}\sum_{k,i}\left(log\ 2\,\pi\sigma^2_{\mu_{i,k}}(\theta) + \left(\frac{\mu_i(t_k,\theta)-\hat{u}_{i,k}}{\sigma_{\mu_{i,k}}(\theta)}\right)^2\right) + \frac{1}{2}\sum_{k,i}\left(log\ 2\pi\sigma^2_{\Sigma_{ii,k}}(\theta) + \right.$$

$$\left.\left(\frac{\Sigma_{ii,k}(t,\theta)-\widehat{\Sigma}_{ii,k}}{\sigma_{\Sigma_{ii,k}}(\theta)}\right)^2\right). \quad (25)$$

These weigh equally the likelihood of the measured mean and variance of the data from each trajectory over all time points measured. To perform maximum likelihood parameter estimation, we used the fminsearch function in MATLAB to minimize the NLL(θ) (Eq. 25). For this optimization, non-negative parameters (rate constants *b* and *d*) were log-transformed while parameters allowed to be negative (extended Allee model Eq. 17–22

parameters A and τ) were normalized between 0 and 1 over a domain of reasonable values of A and τ. We used the log of the slope of the mean cell number in time as an initial guess for the growth rate ($b - d$), a death rate of $d = 0.0005$ cells/hour, $A = 1$ or -1 and $\tau = 2$ were used in order to make a conservative initial guess.

Uncertainty analysis and parameter identifiability

A key benefit of the moment approach for stochastic parameter estimation is that deriving a system of coupled differential equations enables the use of already established methods for parameter identifiability and uncertainty analysis. To evaluate structural identifiability from each model, a differential algebra approach (Meshkat, Sullivant and Eisenberg, 2015; Brouwer *et al.*, 2017) was used to reveal identifiable combinations of parameters in terms of the output we were able to measure in time—in this case, both the mean and the variance of the cell-number trajectories in time. (See Text 3.4 for an example of this approach applied to the birth–death model). This analysis revealed that the parameters in all 7 models are uniquely identifiable.

To ensure that the predicted mean and variance of the models exhibited distinguishable differences from each other qualitatively, we investigated some illustrative cases of the expected mean and variance for a simple birth–death model, a strong Allee model, and a weak case of the extended Allee model (Figure 3.18A and 3.18B). Likewise, to ensure the different forms of the stochastic models within each broader class of deterministic models were distinguishable by the expected differences in their variance (Table 3.1), we display the solutions of the expected mean and variance for strong and weak Allee effects on both birth, death, and equally on both (Figure 3.19A-D). This gave

us confidence that the candidate models were theoretically distinguishable using the mean and variance of the data collected. To evaluate whether these model parameters were practically identifiable and quantify the corresponding uncertainty on these model parameters, the profile likelihood method was used as described in (Raue *et al.*, 2009). The profile likelihood method evaluates the ability to uniquely identify each parameter individually by profiling one parameter at a time, fixing it to a range of values, and fitting for the rest of the parameters at each fixed value. The resulting curvature of likelihood is used to evaluate the uncertainty on the parameter and determine confidence intervals.

Modeling framework is able to distiniguish between different growth models from simulated stochastic trajectories

The parameter estimation and model selection framework were verified by applying the calibration scheme to simulated data from a model of intermediate complexity—the strong Allee effect on birth (Eq. 11 and 12). Using the Gillespie algorithm(Gillespie, 1977, 2014), we generated 5,000 simulated trajectories from initial conditions of $N_0 = 3$, 5, and 10 from the strong Allee effect on birth model. In order to most closely simulate the expected experimental data, the stochastic trajectories were sampled every 4 hours corresponding to the time intervals used in the experimental measurements of cell growth, and the mean and variance were calculated at each time point. A constant random noise term was added to the measurements of mean and variance in time in order to generate trajectories that resembled experimental measurements of mean and variance and to simulate the additive noise of the experimental system (Fig 3.4A and 3.4B). The

simulated data were fit to the 7 candidate models (Eq. 9–22, Table 3.2) representing the range of biological hypotheses, with model complexities ranging from 2 to 4 parameters. To identify the most likely underlying model structure from each of the candidate stochastic models, the Bayesian Information Criterion (BIC) was used for model selection (Raftery, 1999; Wagenmakers and Farrell, 2004; Loos *et al.*, 2018) (See Text 3.5). As expected, the strong Allee effect on birth had the lowest BIC value (Fig 3.4A), revealing that the underlying model structure was correctly identified. The BIC weighting analysis (Wagenmakers and Farrell, 2004) (Text 3.5) revealed strong evidence in favor of the strong Allee effect on birth (Fig 3.4B), indicating the ability of the BIC value to distinguish between overly simple models with 2 parameters and overly complex models with 4 parameters (Fig 3.4C). In order to ensure the method was not overweighing goodness of fit, the data were down-sampled from the true data collection interval of every 4 hours to every 36 hours to demonstrate that down-sampling changed the magnitudes of the BIC values but did not affect the order of the BIC values of each model relative to one another (Figure 3.20). The chosen model provided a good fit to the mean and variance in the data (Fig 3.5B and 3.5C), and the parameter search displayed the expected convergence of accepted parameter values (Fig 3.5D). Profile likelihoods on parameter distributions demonstrated that each of the parameters were practically identifiable and parameter estimates fell close to the true parameters (Fig 3.5E, 3.5F and 3.5G). The true parameter values of $b = 0.00238$, $d = 0.0005$, and $A = 2$ fell within the confidence intervals of the profile likelihood analysis of the fitted parameters of [0.02340–0.02425] for b, [0.00461–0.00563] for d, and [1.853–2.026] for A. This confirms that the calibration approach selects

the appropriate underlying model structure from a set of hypotheses and properly identifies the parameters.

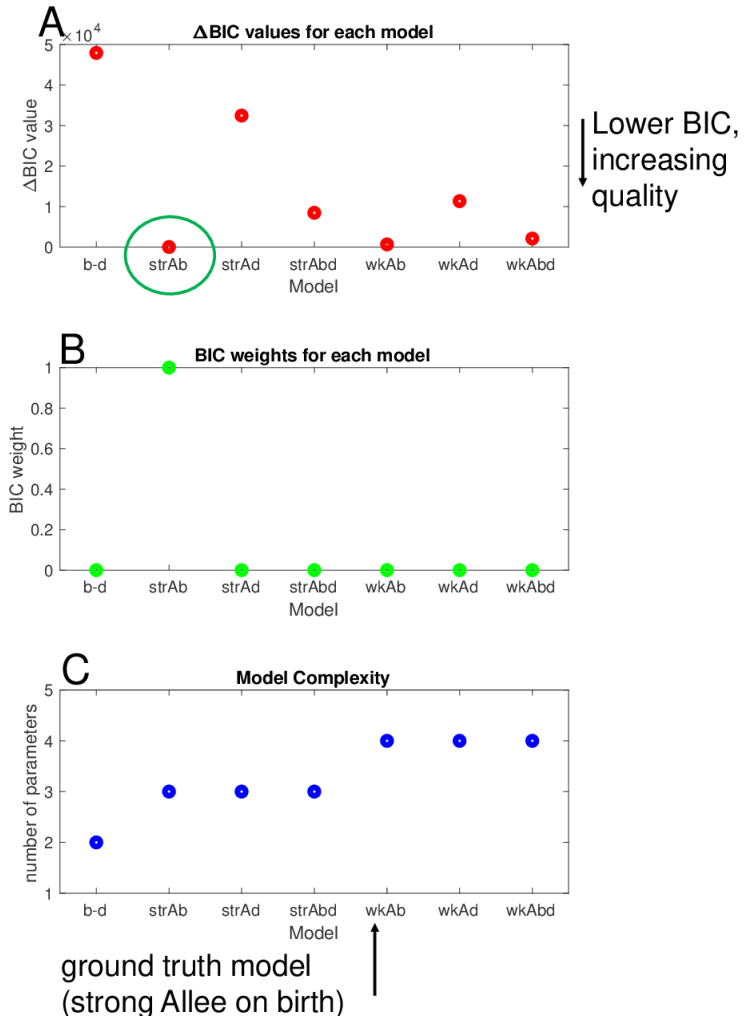

Fig 3.4: continued on next page, Model selection based on the BIC identifies the ground truth model in simulated data. (A) ΔBIC values ($BIC_i - BIC_{min}$) are plotted for the fit of the simulated data set to each of the 7 models from left-to-right: simple birth–death model ($b - d$), strong Allee model on birth (strAb), strong Allee model on death (strAd), strong Allee model on birth and death (strAbd), weak extended Allee model on birth (wkAb), weak extended Allee model on death (wkAd), and weak extended Alee model on birth and death (wkAbd) compared with the highest quality, minimum BIC value model: the strong Allee model on birth. (B) BIC weighting reveals strong evidence to choose the strong Allee model over the other candidate models. (C) Number of parameters of each

model as a measure of relative complexity of the model. The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.
Abbreviations: BIC, Bayesian Information Criterion

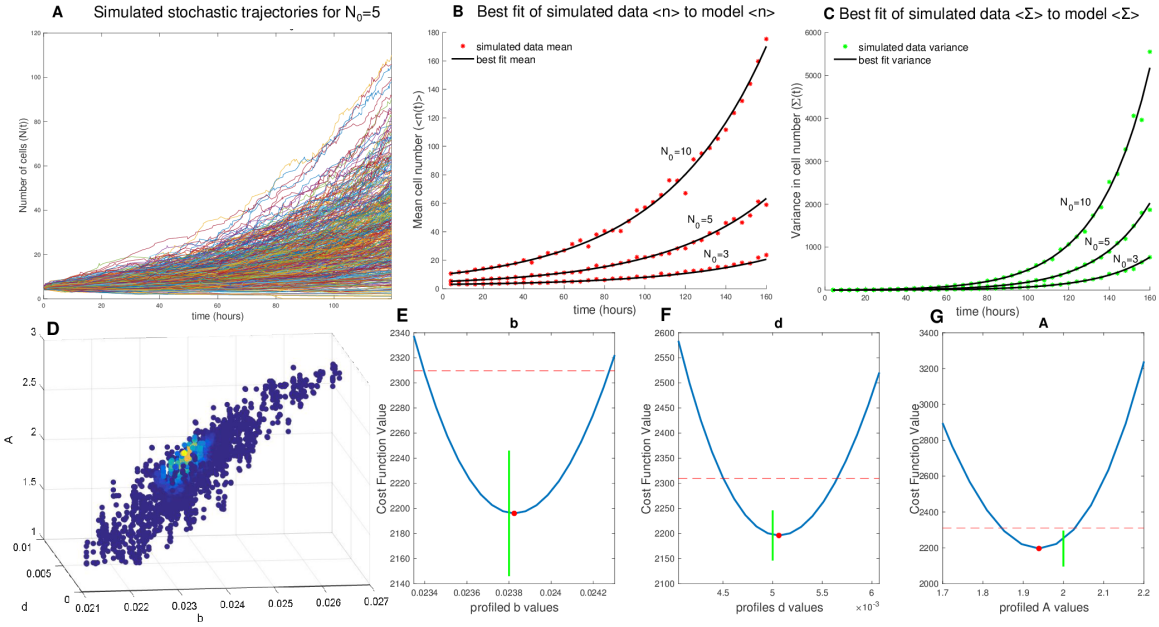

Fig 3.5: Fit to mean and variance from simulated stochastic data set. (A) Example of stochastic growth model output from 5,000 simulated cell-number trajectories with initial condition of $N_0 = 5$ and a birth rate of $b = 0.0238$, death rate of $d = 0.005$, and an Allee threshold $A = 2$, revealing the expected variability in growth dynamics apparent at low initial numbers. (B) From the simulated stochastic trajectories, we sample time uniformly and measure the mean cell number at each time point for $N_0 = 3$, 5, and 10. (C) Again from the simulated stochastic trajectories, we sample time uniformly and measure the variance in cell number at each time point for $N_0 = 3$, 5, and 10 (D) Display of parameter space searched, with parameter sets of b, d, and A colored by likelihood, indicating the framework converges on the true parameters. (E) Profile likelihood analysis of birth rate parameter estimate (red dot) of $b = 0.0238$ [0.02340–0.02425] with true $b = 0.0238$ (green line). (F) Profile likelihood analysis of death rate parameter estimate (red dot) $d = 0.0051$ [0.00461–0.00563] with true $d = 0.005$ (green line). (G) Profile likelihood analysis of Allee threshold parameter estimate (red dot) of $A = 1.9393$ [1.853–2.026] with true $A = 2$ (green line). The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

Experimental measurement reveals scaling of growth rate with initial cell number

Next, we investigated whether the growth of cancer cells in vitro is governed by alternative growth models other than the exponential growth model commonly used to describe tumor cell growth well below carrying capacity. BT-474 breast cancer cells were seeded at a precise initial cell number ranging from 1 to 20 cells per well of a 96-well plate, and time-lapse microscopy images were collected every 4 hours for replicate wells at each initial condition (20–50 wells per condition; see Cell culture and low cell density seeding). Example images are shown in Fig 3.6A, 3.6B, and 3.6C. Cell number as function of time was measured for a total of 328 hours (just under 2 weeks) and cell-number counts in time were determined using digital image processing for each individual well imaged (see Time-lapse imaging and image analysis).

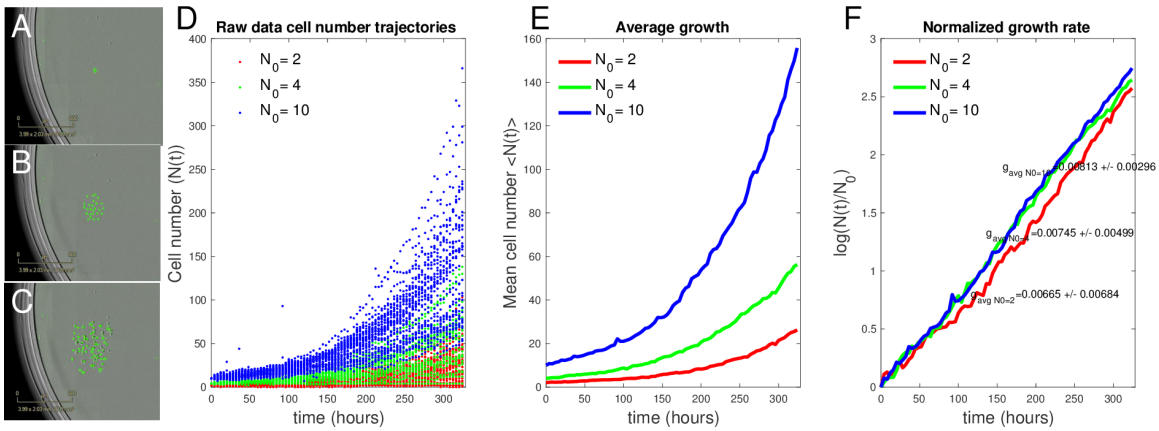

Figure 3.6: BT-474 cancer cells in culture exhibit growth rate scaling with initial cell density. (A, B, and C) Representative images from day 1 (A), day 6 (B), and day 14 (C) of BT-474 GFP labeled cells proliferating in vitro. (D) Individual cell-number trajectories for different $N_0 = 2$, 4, and 10. (E) Average cell number every 4 hours from each trajectory of $N_0 = 2$, 4, and 10. (F) Cell number in time normalized by initial cell number in log scale reveals scaling of growth rate by initial cell number, with $g = 0.00665 \pm 0.00684$, 0.00745 ± 0.00499, and 0.00813 ± 0.00296 for $N_0 = 2$, 4, and 10, respectively.The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

The true initial cell number (N_0) sorted into each well was confirmed by eye from the initial image, and wells were binned according to the observed initial cell number. Cell-number trajectories of wells with initial cell numbers of 2, 4, and 10 cells are shown in Fig 3.6D in red, green, and blue, respectively. As a preliminary analysis of this data, we fitted each well individually to the exponential growth model (Eq. 1 and 2) to obtain a distribution of growth-rate constants at each initial condition. The mean growth rates for N_0 = 2, 4, and 10, respectively, were g = 0.00665 ± 0.00684, 0.00745 ± 0.00499, and 0.00813 ± 0.00296. Fig 6 displays the average cell-number trajectory (Fig 3.6E) and the normalized growth rate ($log(\frac{N(t)}{N_0}$); Fig 3.6F) for the measured data at each time point. These results indicate clear deviations from the simple exponential growth model in which the normalized growth rate ($log(\frac{N(t)}{N_0})$) is expected to be identical for all initial conditions (see Fig 3.1C). Instead, growth behavior resembled the characteristic scaling of normalized cell numbers by initial cell number that is observed for both Allee effect models (Fig 3.1F and 3.1I). The scaling of average growth rate with initial cell number had been observed by Neufeld and colleagues (Neufeld *et al.*, 2017) in their in vitro studies of cell culture, providing us with the motivation to further investigate whether an Allee model better describes BT-474 breast cancer cell growth.

To ensure that the observed differences in growth rate at low cell densities were significantly different from what is observed at normal cell culture seeding densities, we sorted N_0 = 512 and N_0 = 1,024 cells and captured 30 growth trajectories from each initial condition. The mean and standard deviation of the growth rates were not significantly

different from one another and also significantly higher than the observed low cell density growth rates, with $g = 0.0112 \pm 0.00062$ and $g = 0.0115 \pm 0.00074$ for $N_0 = 512$ and $N_0 = 1{,}024$, respectively (Figure 3.21). The absence of density-dependent growth rates at these higher initial cell numbers may explain why the Allee effect hasn't been described using standard cell culture seeding densities.

Fit of experimental data to stochastic growth models reveals Allee effect

The variability in the observed cell-number trajectories for a single initial condition is reflected in the experimental measurements of BT-474 cells growing at low initial cell densities (Fig 3.6D). This variability in cell growth dynamics is expected due to the inherent stochasticity of the birth and death processes, which is apparent at the small population sizes measured in this study (Fig 3.5A). Because stochasticity is more apparent and can be observed experimentally at the low cell numbers (Fig 3.6D), such dynamics are appropriately modeled by a stochastic rather than a deterministic process. In order to determine whether the preliminary observations of growth-rate scaling with the initial cell number could be described by alternative models of cell population growth that consider the Allee effect, the experimental data of BT-474 growth trajectories shown in Fig 3.6D for initial cell numbers of 2, 4, and 10 were calibrated to the 7 stochastic models using the stochastic modeling framework presented above (Fig 3.3).

Fitting each initial condition separately to the simple birth-death model reveals net growth rate increases with initial cell number

To determine whether birth and/or death rates depend on the initial cell number, we first fit the data for initial cell number of $N_0 = 2$, 4, and 10, grouped by initial condition N_0

individually, to the stochastic simple birth and death model (Eq. 9 and 10) using the workflow described in Fig 3.3. The results of the fitting to the mean and variance in time to the simple birth–death model for each initial condition are shown in Fig 3.7 (3.7A, 3.7B, and 3.7C for the mean and 3.7D, 3.7E, and 3.7F for the variance). Each data set of a single initial condition N_0 revealed identifiable birth and death rate parameters via profile likelihood analysis (see Figure 3.22). Birth and death rate maximum likelihood parameter estimates are shown in Fig 3.7G, with confidence intervals obtained from the profile likelihood analysis (Figure 3.22). Parameter estimates for birth rates by initial cell number are b_2 = 0.00793 [0.00785–0.00794], b_4 = 0.00945 [0.0093–0.0096], and b_{10} = 0.0113 [0.0112–0.0114], and for death rates, the parameter estimates are $d_2 = 6.67 \times 10^{-18}$ [−0.0005 to 0.0001], d_4 = 0.0011 [0.0008–0.0013], and d_{10} = 0.00286 [0.0025–0.0028] for N_0= 2, 4, and 10, respectively. The trend suggests a slight increase in net growth rate (birth rate minus death rate) with initial cell number, as is consistent with the preliminary growth rate analysis by initial cell number (Fig 3.6F) but inconsistent with the conventional exponential growth hypothesis, which should yield the same growth rate (and same birth and death rates), independent of the initial cell number.

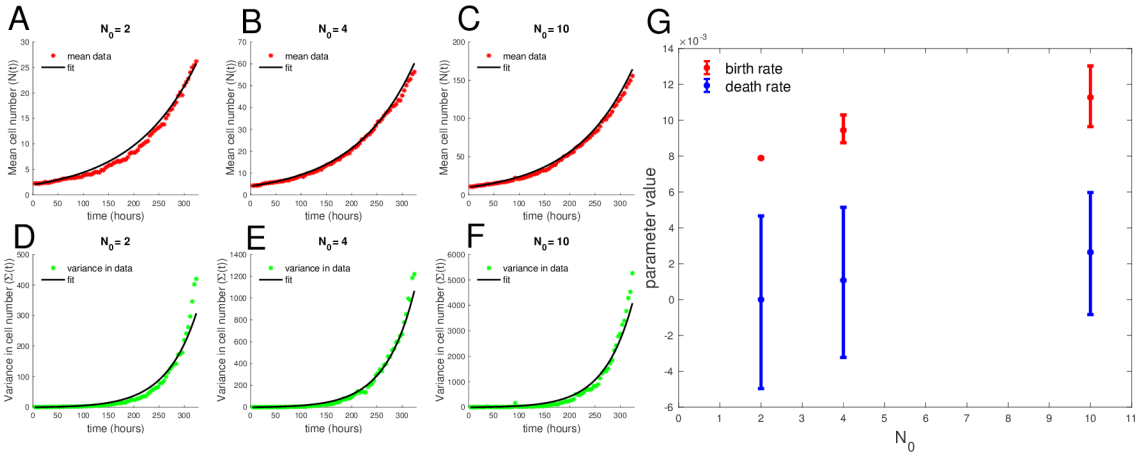

Fig 3.7: Best fit of each initial cell-number means and variances in time to stochastic birth–death model reveals net growth rate increases with initial cell number. (A, B, and C) Data mean over time compared to best fit to model mean. (D, E, and F) Data variance in time compared to best fit to model variance. (G) Best fit birth and death rate parameters for the stochastic birth–death model fit to each initial condition, with confidence intervals determined from profile likelihoods. Parameter estimates for birth rates by initial cell number are $b_2 = 0.00793\ [0.00785 – 0.00794]$, $b_4 = 0.00945\ [0.0093 – 0.0096]$, and $b_{10} = 0.0113\ [0.0112 – 0.0114]$, and for death rates, the parameter estimates are $d_2 = 6.67 \times 10^{-18}\ [-0.0005 \text{ to } 0.0001]$, $d_4 = 0.0011\ [0.0008 – 0.0013]$, and $d_{10} = 0.00286\ [0.0025 – 0.0028]$ for N_0 = 2, 4, and 10, respectively. The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

Fit of low seeding density data to all stochastic models reveals a weak Allee effect

The growth data from the initial conditions of N_0 = 2, 4, and 10 were combined and fit to each of the 7 candidate models using the moment-closure approximation workflow described in Fig 3.3 (Fröhlich *et al.*, 2016). The BIC values for each model fit were computed and compared with the minimum BIC value (Fig 3.8A), and the corresponding BIC weights were calculated (Fig 3.8B) based on the goodness of fit and the complexity of the model (number of parameters; Fig 3.8C). We note that both the strong and weak Allee effect on birth models have significantly lower BIC values than the null model of the simple birth–death model (Fig 3.8A), providing strong, consistent evidence for the

presence of an Allee effect in some form in this data set. Using the BIC weights to evaluate statistical significance between the models revealed that the weak Allee effect on birth is more likely than the strong Allee effect on birth model, with a BIC weight of essentially 1 to 0 for the weak Allee effect on birth versus the strong Allee effect on birth model. The best fit of the weak Allee effect on birth model to the mean and variance of the data is shown in Fig 3.9A and 3.9B, respectively (See Figures 3.23-3.28 for the fit of the data to all 7 candidate models).

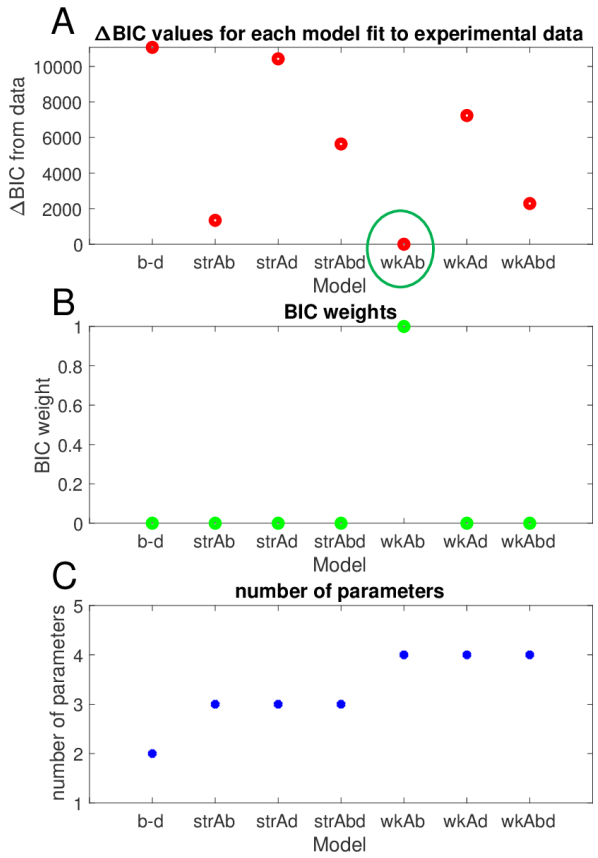

Fig 3.8: Weak Allee model on birth best describes BT-474 in vitro growth data. (A) *Δ*BIC values for the fit of the data to each of the 7 candidate stochastic growth models shows that the weak Allee model on birth exhibits the lowest BIC value. (B) BIC weights for each model indicate that the weak Allee model on birth is significantly better than all other models. (C) Number of parameters in each model as a measure of model complexity. The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git BIC, Bayesian Information Criterion

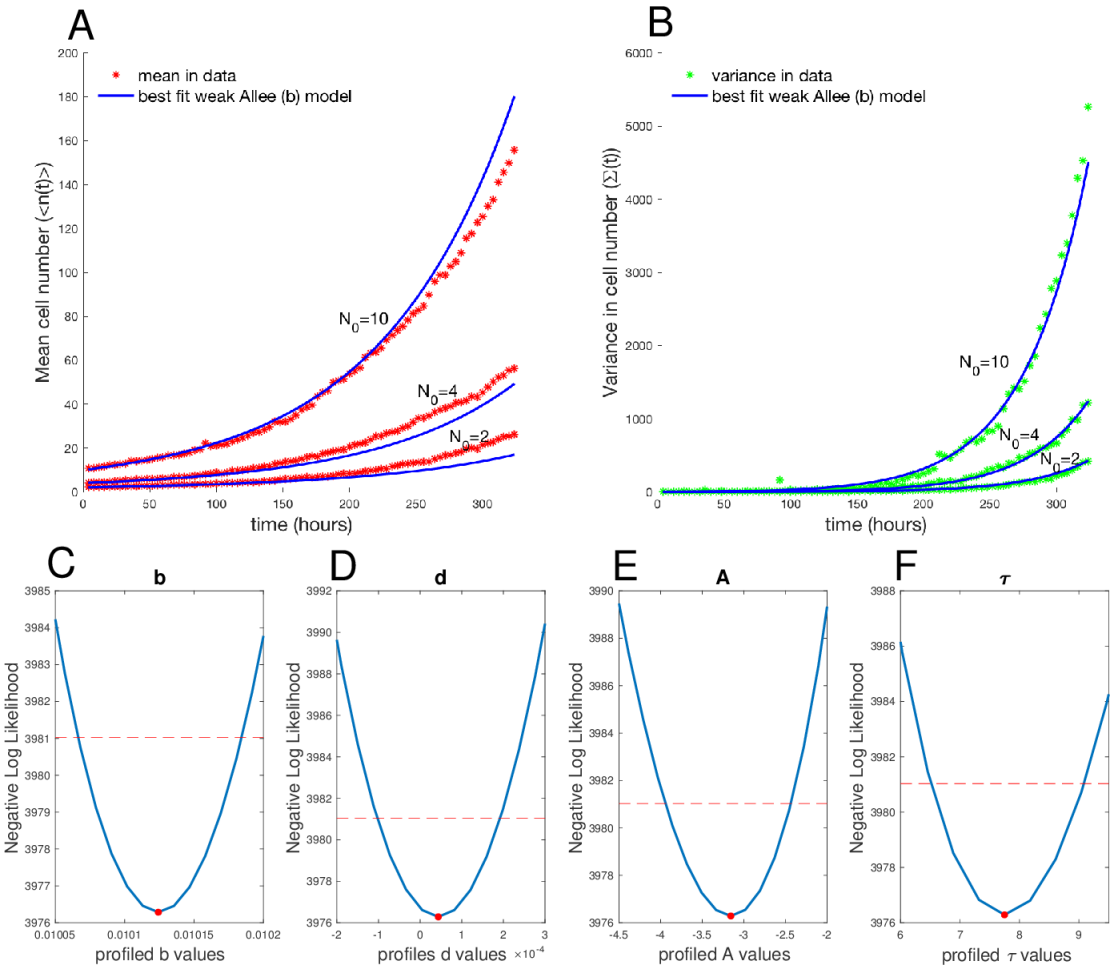

Fig 3.9: Mean and variance of data fit to a weak Allee model on birth. (A) Best fit of data mean to model mean displays the model fits the data well over all 3 initial conditions and over the time course. (B) Best fit of data variance to model variance displays the model fits the data well. (C) Profile likelihood analysis of birth rate around maximum likelihood $b = 0.0101$ $[0.010068 – 0.010181]$. (D) Profile likelihood analysis of death rate around maximum likelihood $d = 4.3613 \times 10^{-5}$ $[-7.270 \times 10^{-5}$ to $1.599 \times 10^{-4}]$. (E) Profile likelihood analysis of Allee threshold $A = -3.1576$ $[-3.8593$ to $-2.4559]$. (F) Profile likelihood analysis of the overall shape parameter $\tau = 7.480$ $[6.8871 – 9.0393]$. The data and code used to generate this figure can be found at https://github.com/brocklab/Johnson-AlleeGrowthModel.git.

The profile likelihoods used to determine the 95% confidence intervals of the best fitting parameters of $b = 0.0101$ $[0.010068–0.010181]$, $d = 4.3613 \times 10^{-5}$ $[-7.270 \times 10^{-5}$ to $1.599 \times 10^{-4}]$, $A = -3.1576$ $[-3.8593$ to $-2.4559]$, and $\tau = 7.480$ $[6.8871–9.0393]$ are

displayed in Fig 3.9C, 3.9D, 3.9E, and 3.9F respectively. The discrepancy between the goodness of fit in the model mean and variance compared to the data is likely because an unbiased approach (as is prescribed in (Fröhlich *et al.*, 2016)) was used to weight the fit of both the mean and variance equally, using the likelihood function described in Eq. 24. In theory, the relative weighting of the value of these 2 outputs could be tuned to reduce the error between the model and measurements in either the mean or variance. The results of model selection for the weak Allee effect model for the BT-474 data indicates that, outside of the effects of demographic stochasticity, any initial cell number is predicted to, on average, develop into a growing cell population, but the growth rate is expected to be significantly slower at low cell numbers.

DISCUSSION

The availability of single-cell resolution live imaging of cancer cell growth in a controlled in vitro setting starting at the population size of a single cell allowed us to examine in detail the influence of absolute cell number in a cell population on growth rate. Using mathematical modeling, we investigated the departure from simple first-order exponential growth kinetics in which the growth rate is proportional to the population size (cell number). Cell–cell interactions, as best known from quorum sensing in bacteria (Jiang *et al.*, 2019), underlie the cell-number dependence of growth rates. Most work on such dependence have been concerned with the slowing of growth with increasing cell number, e.g., due to approaching the carrying capacity of the cell culture. Here, we focus on the initiation of cell growth from a few individual cells and ask whether cooperative behavior

or the Allee effect, as it is known from ecology, can be detected in a departure from exponential growth kinetics as predicted by mathematical models that consider the Allee effect. Because at the early stages of growth (from one cell or a few) growth kinetics is subjected to stochastic fluctuations due to small cell numbers, we formulated stochastic models that consider the Allee effect. We have demonstrated a framework for testing the relevance of a set of stochastic models of cancer cell growth applied to high-throughput, single-cell resolution data.

The 7 distinct candidate stochastic models of growth describe various modifications of the exponential growth model by incorporating growth-rate dependencies on the size of the population. The average behaviors of these models are examined in the deterministic form, and corresponding stochastic models that lead to the average behavior are developed. To test the relevance of the proposed stochastic models, the moment-closure approximation method (Fröhlich *et al.*, 2016) for parameter estimation in stochastic models (Fig 3.3) is applied to the high-throughput cell growth data. We first validated our framework by computational simulation of growth trajectories using a model of intermediate complexity. The parameter estimation framework was applied to the simulated data, confirming the ability of the framework to properly identify the underlying model structure and the true parameters. The framework is applied to a data set with a number of replicates from 3 initial conditions of N_0 = 2, 4, and 10 BT-474 breast cancer cells. The fit of this growth data reveals that the weak Allee model with decreasing birth probability at a low cell number best describes the observed in vitro growth dynamics.

The presence of an Allee effect, even in the nutrient- and space-rich cell culture setting, implies that cancer cells likely exhibit cooperative growth. The ubiquitous cellular heterogeneity in tumors suggests that cooperative interactions between distinct subsets of cells must be present in order to maintain the observed heterogeneity. Evidence for noncell autonomous growth via eco-evolutionary interactions was recently observed by Kaznatcheev and colleagues (Kaznatcheev *et al.*, 2019), from which they observed a fitness benefit to combining fluorescently labeled cancer associated fibroblasts from parental and resistant cell lines and observed a benefit in growth rate of each independent cell type. Other microscopic experimental systems in which frequency dependent fitness effects have been considered include *Escherichia coli*, yeast, and other cancer cell types (Kaznatcheev *et al.*, 2019). Recent work by Marusyk and colleagues (Marusyk *et al.*, 2014) has found evidence for noncell autonomous proliferation using a mathematical modeling framework, showing that the null hypothesis of no clonal interactions can be easily rejected in favor of a model that considers a specific clone that helps support the growth of all other clones. Additionally, studies in which clonal diversity has been manipulated by combining clones in culture have demonstrated that the presence of diverse clones is necessary to obtain the observed growth rate achieved in multiclonal parental cell cultures (Wangsa *et al.*, 2018).

Single-cell and clonal analysis has enabled the detection of secreted growth inducing factors, such as ILII (Marusyk *et al.*, 2014), Wnt1 (Cleary *et al.*, 2014), IGFIII (Archetti, Ferraro and Christofori, 2015), and other paracrine factors (Hoelzinger, Demuth and Berens, 2007; Scheel *et al.*, 2011) in certain clones that result in an increased growth rate in the surrounding nonproducing clones. Bioinformatic analysis of single-cell gene

expression data has allowed for the identification of specific subsets of cells that produce high levels of certain ligands and coexist in a population with cells that contain high expression levels of the cognate receptors (Graeber and Eisenberg, 2001; Zhou *et al.*, 2017; Kumar *et al.*, 2018). Prior to single-cell analysis capabilities, these types of interactions were not readily detectable from bulk gene expression measurements. In such data, the coexpression of a ligand and its cognate receptor in the same sample (a cell population) has by default been interpreted as autocrine signaling (Graeber and Eisenberg, 2001). Both paracrine and autocrine signaling are likely to play a significant and varying role in tumor growth.

In the field of tumor growth modeling, a few studies have considered the role of the Allee effect and the importance of incorporating it to describe and predict the effects of cooperative growth. Bottger and colleagues (Böttger, Hatzikirou and Voss-böhme, 2015) developed a stochastic model in which an Allee effect naturally manifests based on assumptions that cancer cells can either exist in a migratory state or a proliferative state. Additional theoretical work has focused on spatial interactions between cancer cells and incorporated the Allee effect in a model for spatial spreading of cancer (Sewalt *et al.*, 2016). However, most classical tumor growth models rely on the assumption that early stage growth dynamics match the single exponential growth model (Winsor, 1932; Speer *et al.*, 1984; Benzekry *et al.*, 2014; Lima *et al.*, 2016; Pacheco, 2016; West and Newton, 2018). The weak Allee effect revealed in this work provides evidence that descriptions of early stage growth dynamics, which are relevant to progression, relapse, and metastasis, may be improved by taking into account the expected slowing of growth at low cell numbers.

Beyond improving predictions of tumor growth and relapse dynamics, a model that considers the Allee effect may help to explain how cancer cell populations are able to go extinct after therapy despite the prediction of the log-kill hypothesis, which states that the probability of a cell being present after treatment, if a tumor is initially large, is greater than zero (Poleszczuk and Enderling, 2018).

Although much work in tumor biology has led to an appreciation for cancer as an evolutionary process, a focus on cancer cells as ecosystems of interacting species or subpopulations may yield new insights. The possibility of exploiting ecology for the treatment of tumors based on studies in conservation biology about extinction and control of invasive species has been previously proposed (Gatenby, 1991; Mcgregor, Axelrod and Axelrod, 2008; Chen and Pienta, 2011; Basanta *et al.*, 2013; Korolev, Xavier and Gore, 2014; Amend and Pienta, 2015; Amend *et al.*, 2016; Han *et al.*, 2016; West and Newton, 2018; Axelrod and Pienta, 2018; Kimmel *et al.*, 2019). However, this is the first work to our knowledge that has explicitly tested for the presence of the Allee effect in a regime in which low cancer cell populations can be measured and fit to a number of stochastic model structures representing different biological hypotheses about the Allee effect. Our finding is consistent with preclinical (Panigrahy *et al.*, 2012) and clinical observations (Neufeld *et al.*, 2017; Spiteri *et al.*, 2018) of threshold-like behavior of tumor growth or slowed tumor growth following resection. Evidence for the Allee effect is also consistent with evidence of cooperation among cancer cell subclones as has been amply demonstrated (Brown *et al.*, 1990; Axelrod, Axelrod and Pienta, 2006; Cleary *et al.*, 2014; Marusyk *et al.*, 2014; An *et al.*, 2015; Archetti, Ferraro and Christofori, 2015). An understanding of subpopulation

interactions and their molecular mediators that drive the observed Allee effect offer new approaches to manipulate cancer cell growth dynamics in favor of extinction. Allee effect models have been used to compare the impact of alternative management scenarios on threatened or exploited populations that are not readily accessible to experimentation (Courchamp, Berec and Gascoigne, 2008). Although the models themselves are phenomenological, the principles behind them, such as growth promoting cooperation, have been confirmed by ecological observations. The concept of cooperation promoting growth is intuitive to both the ecologist and modeler, in a similar fashion to the way we understand the carrying capacity term in the logistic growth model to represent the biological phenomena of slowing growth due to finite resources and space, and in the same manner, we intend knowledge of the Allee effect to be useful in a variety of contexts.

This study, which seeks to establish feasibility of detection and mathematical description of the Allee effect by observing growth kinetics, has obvious limitations with respect to biological interpretation of the relevance of results. Most notably, we apply the modeling and analysis framework to an in vitro data set for a single breast cancer cell line. The in vitro system may not faithfully represent in vivo growth dynamics, although we expect, and others have shown evidence that(Neufeld *et al.*, 2017; Spiteri *et al.*, 2018), the Allee effect would only be more pronounced in vivo. An in vitro setting provides cells with all of the growth factors, nutrients, and space to robustly grow at low cell densities, whereas these factors may be less abundant for tumor cells in vivo at a low cell density. Although numbers of replicates for each initial condition N_0 were relatively high (20 to 50 replicates) compared with typical growth studies, an increase in the number of replicates would likely

lead to an improvement because the variance in the data should increase in accuracy with increasing sample size. In order to confirm that the Allee effect is a hallmark of tumor growth, a wide range of tumor types will need to be investigated. Additionally, the model presented here is phenomenological; we do not infer the mechanisms by which an Allee effect may be occurring such as in (Böttger, Hatzikirou and Voss-böhme, 2015; Sewalt *et al.*, 2016), nor do we explicitly develop a model of subpopulation interactions as had been done in (Marusyk *et al.*, 2014). Future work will focus on investigating the molecular and cellular mechanisms for an Allee effect and developing a model of heterotypic subpopulation interactions that also considers phenotypic plasticity (Piyush B Gupta *et al.*, 2011; Zhou *et al.*, 2014; Pisco and Huang, 2015; Jolly *et al.*, 2017).

This work provides a framework for in-depth investigation of mathematical models of stochastic growth that incorporate the Allee effect and shows that an Allee effect model may be more suitable to describing early stage tumor growth dynamics than the exponential model. The potential role of the Allee effect opens a variety of new possibilities for understanding and controlling tumor growth. Biological mechanisms of cooperative growth that may be critical for cell populations to enter a highly proliferative regime need to be further investigated, because these mechanisms may be critical to preventing and predicting metastases and tumor relapse.

CHAPTER 3 SUPPLEMENTARY FIGURES, TABLES, AND TEXT

Figures

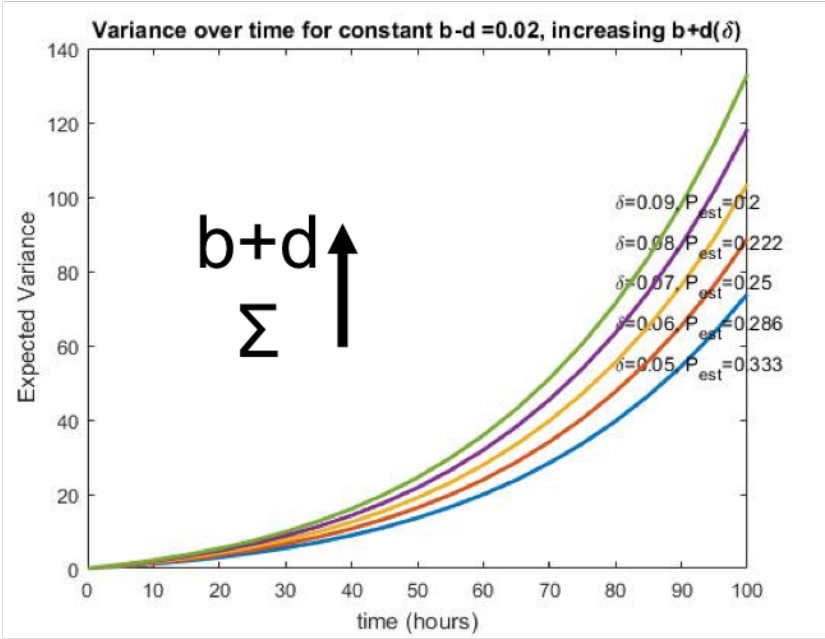

Figure 3.10: Illustrative example demonstrates that increasing magnitude of birth and death rate parameters increases variance in time, enabling the identifiability of b+d. The net growth (b-d) was held constant and the magnitude of b and d were simultaneously increased in order to demonstrate the effect on the time evolution of the variance. This example is used to explain intuitively how the measurement of variance in time enables the proper identification of the b and d parameters uniquely, even while the time evolution of the mean cell number remains constant.

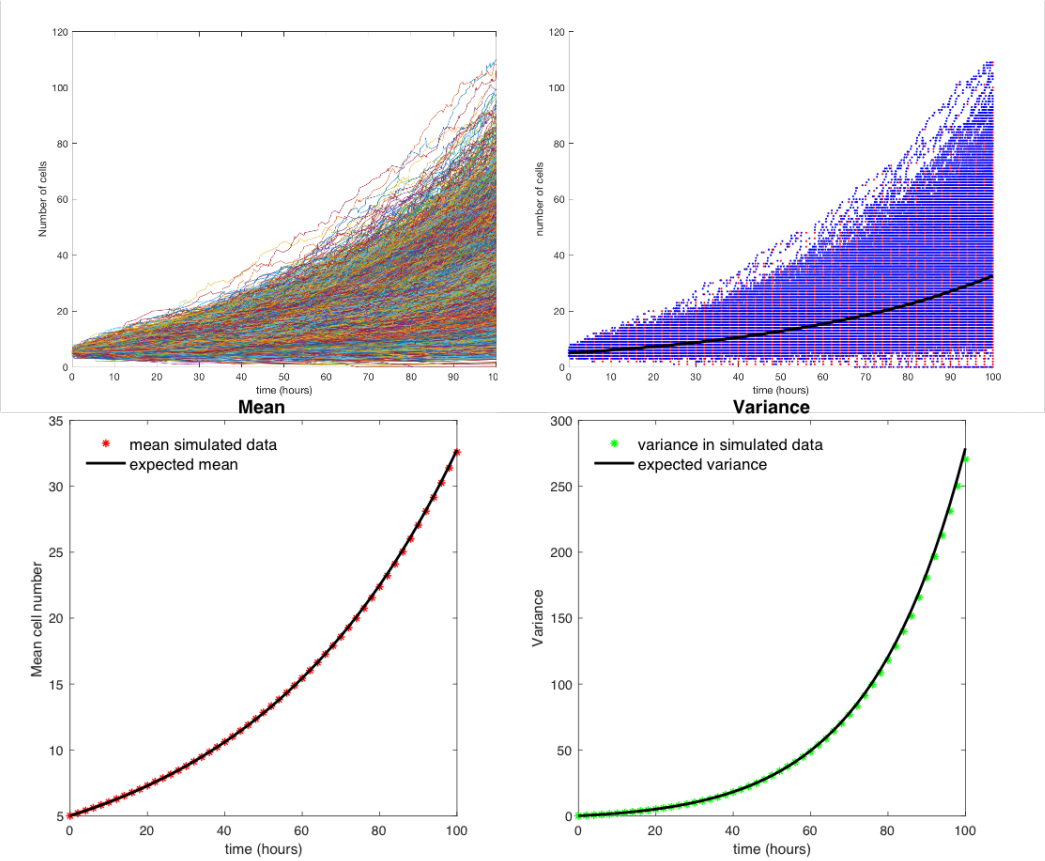

Figure 3.11: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories (a) Example of stochastic growth model output from 5000 simulated cell number trajectories starting at a single cell with birth rate of b = 0.0238 and a death rate of d = 0.005, revealing the expected variability in growth dynamics that is not averaged out at low initial numbers (b) Stochastic growth trajectories uniformly samples every 4 hours (c) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (d) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

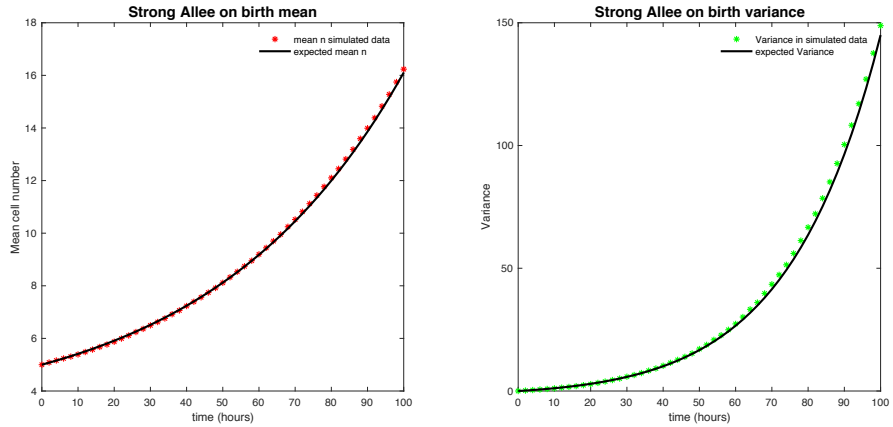

Figure 3.12: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for strong Allee model on birth (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

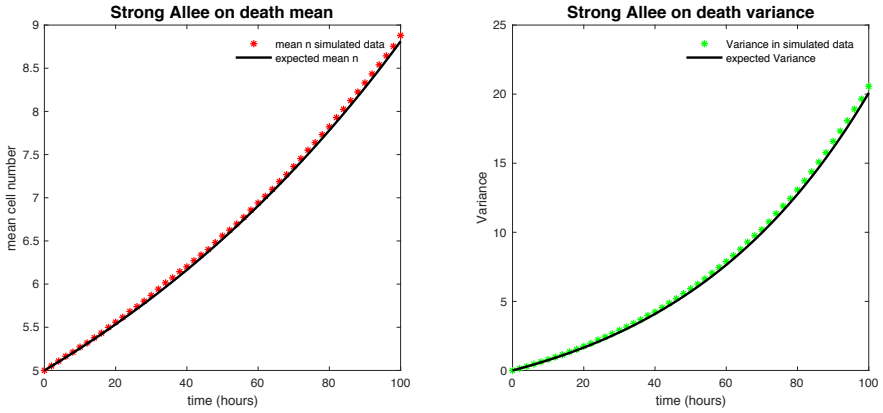

Figure 3.13: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for strong Allee model on death (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

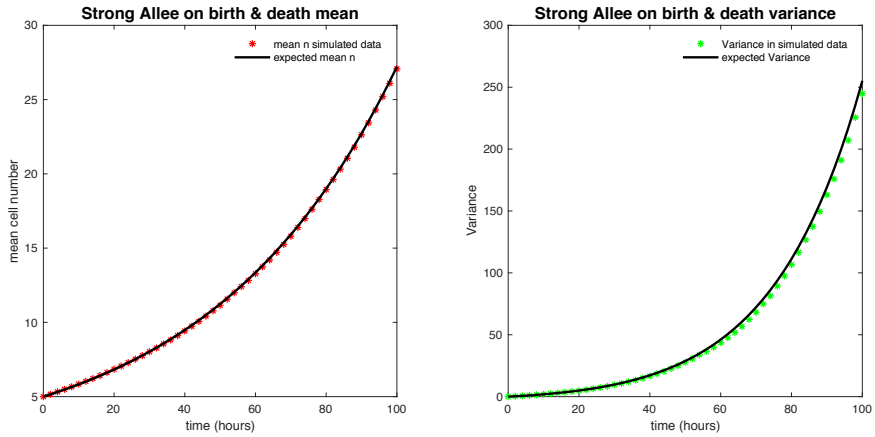

Figure 3.14: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for strong Allee model on birth & death (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

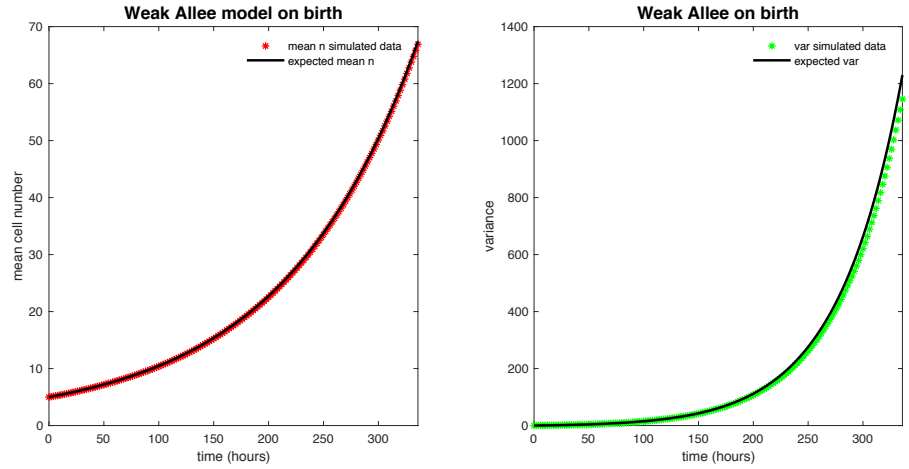

Figure 3.15: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for weak Allee model on birth (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

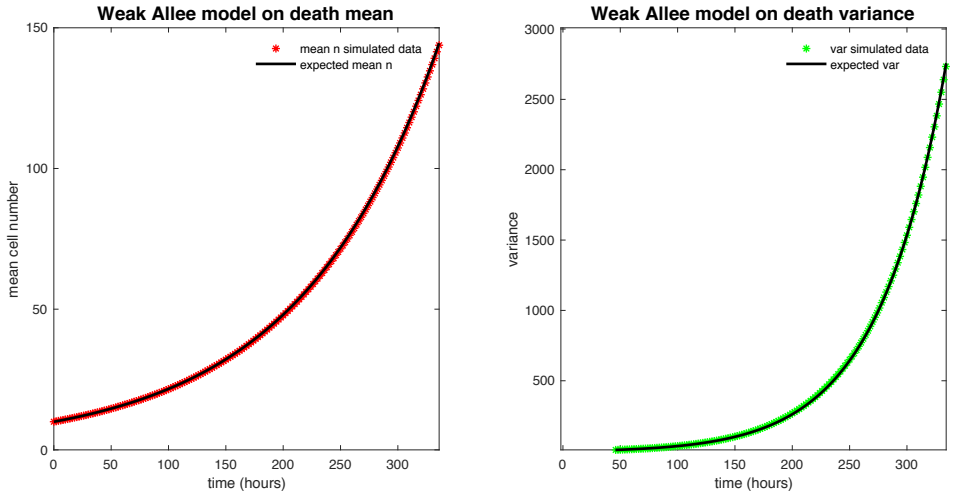

Figure 3.16: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for weak Allee model on death (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

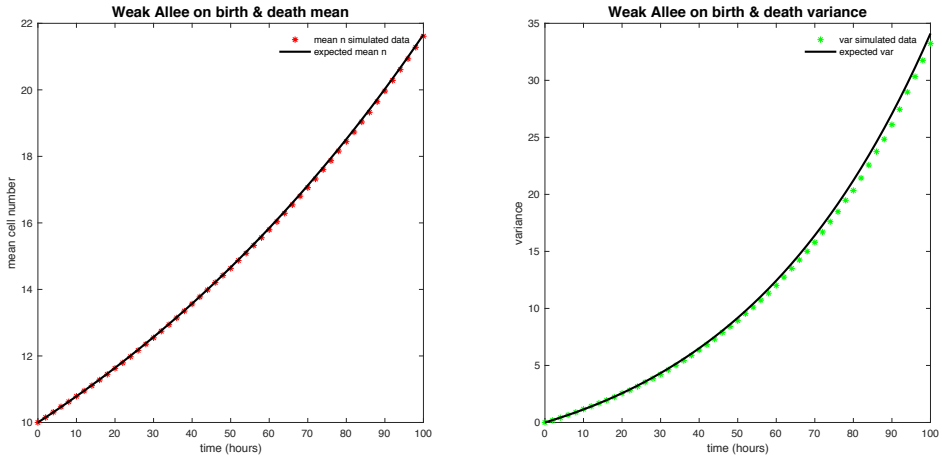

Figure 3.17: Confirmation that moment approach derivations match measured mean and variance from simulated stochastic trajectories for weak Allee model on birth & death (a) Measured mean at each time interval from simulated data with model expected mean as a function of time for the true parameters overlaid. (b) Measured variance at each time interval from simulated data with model expected variance as a function of time for the true parameters overlaid.

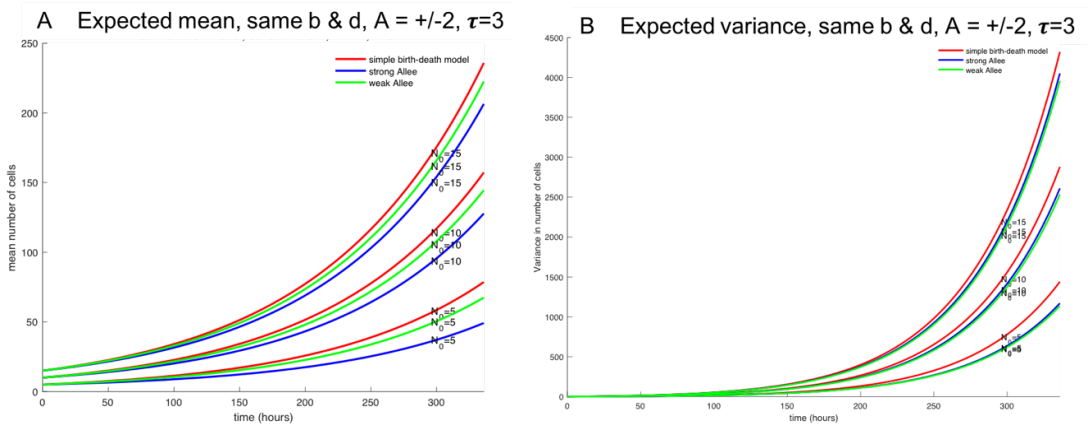

Figure 3.18: Comparison of similar simple birth-death, strong, and weak Allee expectations for the time evolution of the mean and variance. (a) Expected time evolution of the mean cell number for the simple birth-death model (red), the strong Allee model on birth (blue) and the weak Allee model on birth (green) with the same birth and death rates for all but with A=2 for the strong Allee model and A=-2, τ=3 for the weak Allee model indicates significant differences in trajectories for N_0 = 5, 10, & 15 (b) Expected time evolution of the variance in cell number for the same initial conditions and parameters.

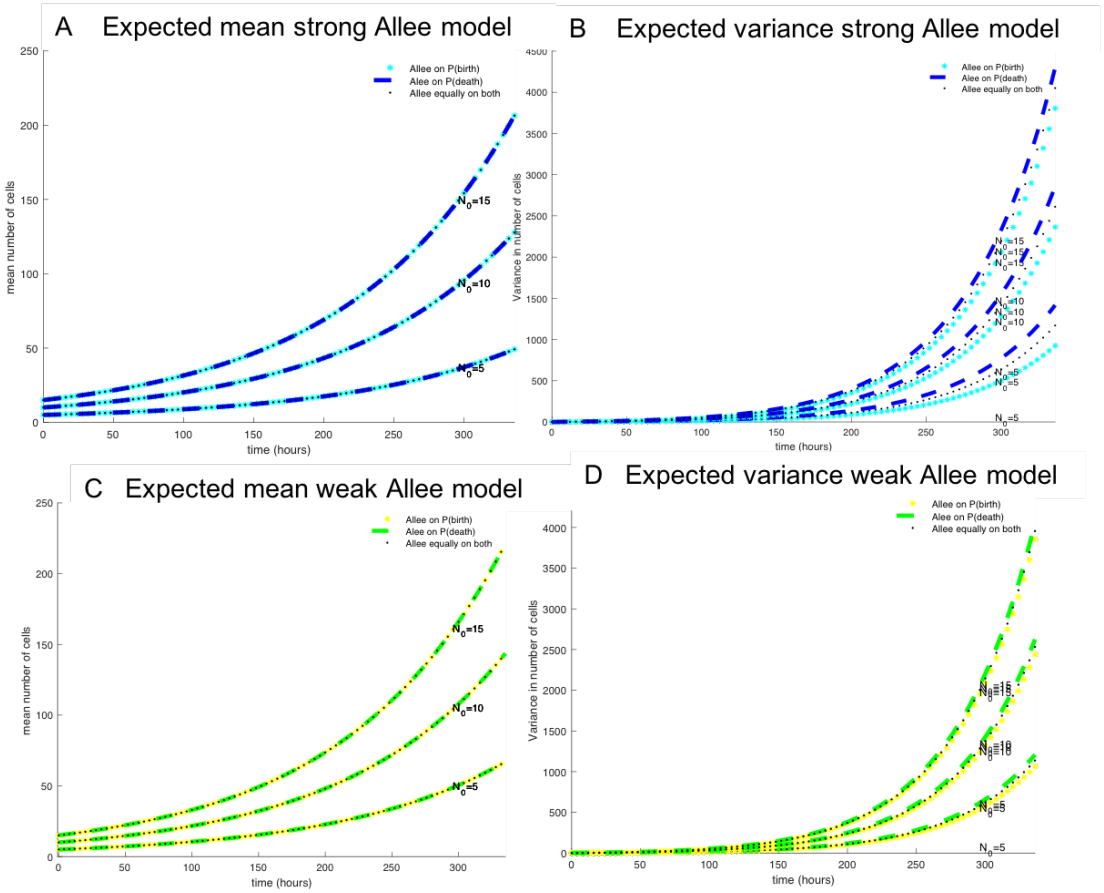

Figure 3.19: Demonstration of effect of Allee mechanism on birth or death probability on the variance. (a) As expected, for constant parameters the mean cell number in time for the strong model is the same for the strong Allee model on birth, death, or both. (b) The expected time evolution of the variance for the strong Allee model on the birth probability (cyan), death probability (dark blue) and them equally (black) (c) As expected, for constant parameters the mean cell number in time for the weak Allee model is the same for the strong model on birth, death, or both. (b) The expected time evolution of the variance for the strong model on the birth probability (yellow), death probability (green) and them equally (black).

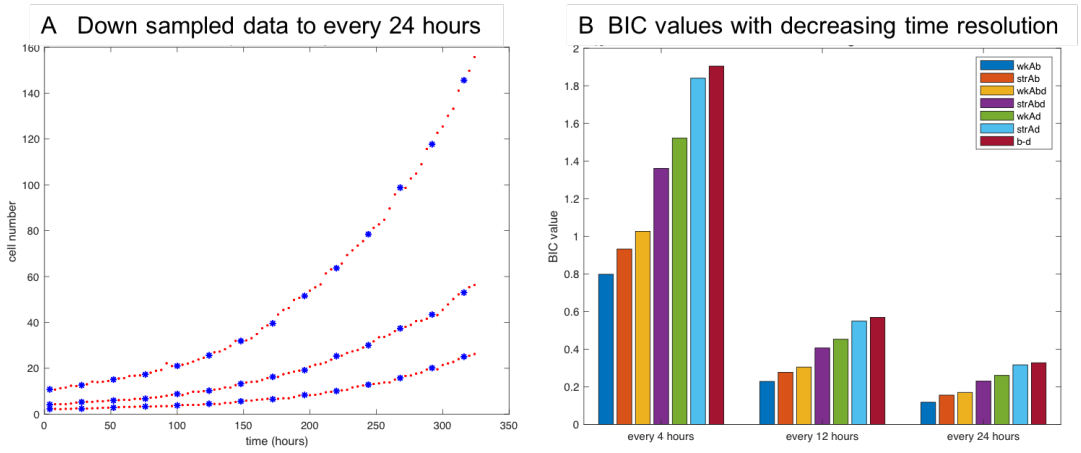

Figure 3.20: Decreased time resolution of data does not change fitting results, weak Allee model on birth is consistently chosen. (a) Example of down sampled time resolution from original data (red) to data every 24 hours (blue) (b) BIC values for each model fit at data sampled every 4, 12, and 24 hours respectively reveals weak Allee model has consistently the lowest BIC value and is chosen every time.

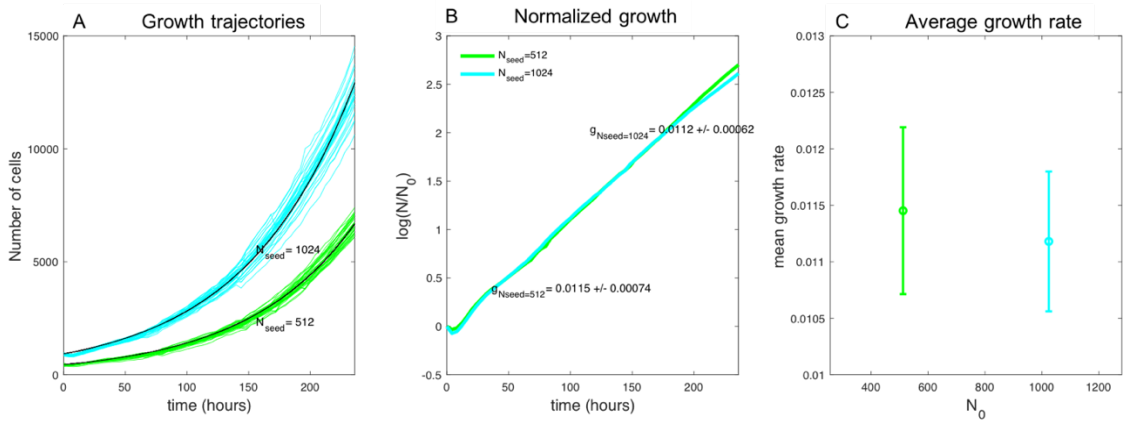

Figure 3.21: Normal cell culture density exhibits expected constant growth rate (a) Thirty growth rate trajectories for seeding of N_0 = 512 (green) & N_0 = 1024 (cyan) (b) Normalized cell number in time by N_0 reveals expected constant growth rate (c) Average growth rate and for N_0=512 (green) of g = 0.0112+/- 0.00062 and N_0 = 1024 g = 0.0115 +/- 0.00074

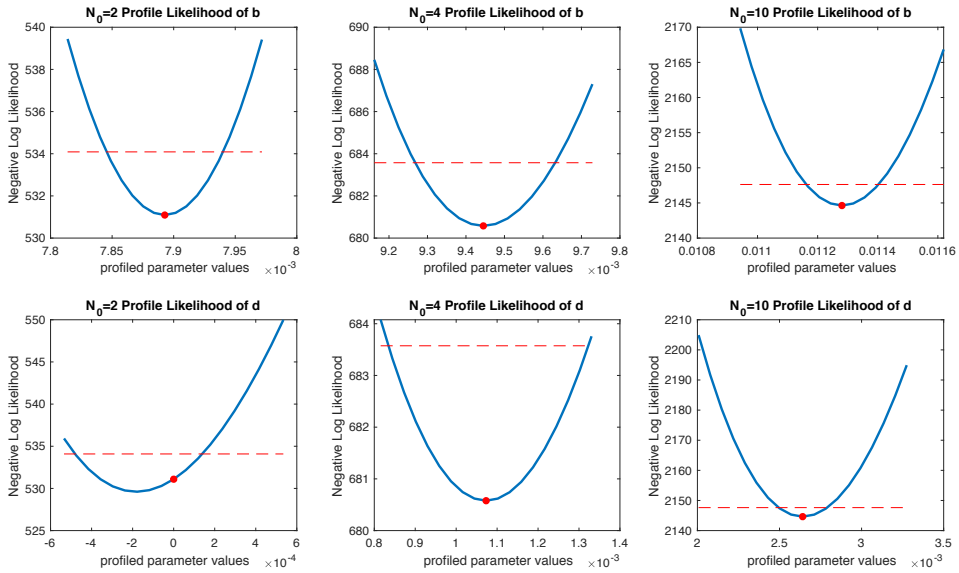

Figure 3.22: Profile likelihood analysis on birth and death rates for individual cell numbers N_0= 2, 4, & 10 reveals practical identifiability of the birth and death rate parameters for datasets of each individual group of N_0 trajectories. (a), (b), (c). Profile likelihood analysis on birth rate parameter for N_0=2, 4, &10 respectively. (d), (e), (f). Profile likelihood analysis on death rate parameter for N_0 = 2, 4, & 10 respectively.

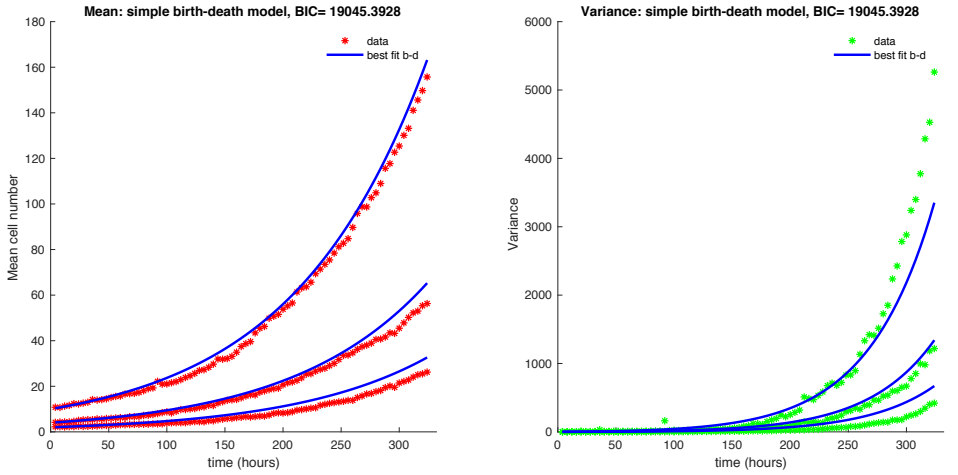

Figure 3.23: Data fit to birth-death model results in a BIC= 1.9e4. (a) Mean of the data (red) to the best fitting birth-death model mean (blue). (b) Variance of the data (green) to the best fitting birth-death model variance (blue).

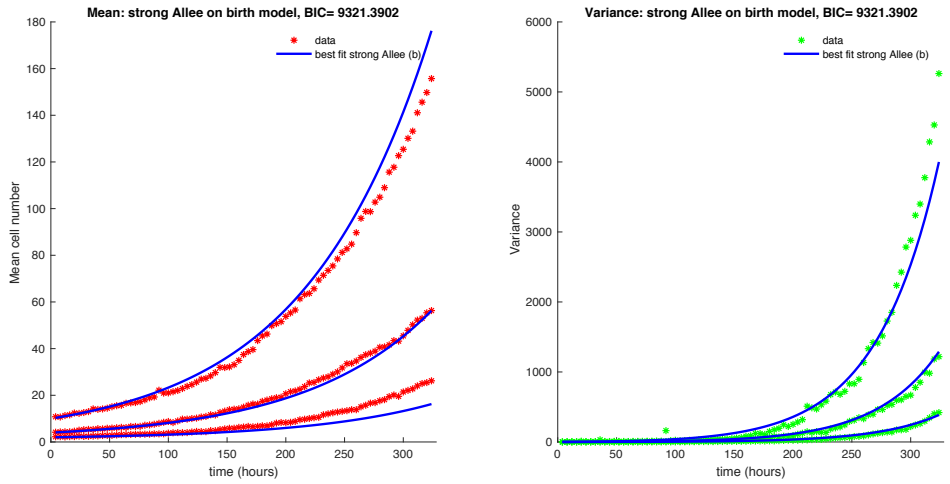

Figure 3.24: Data fit to strong Allee on birth model results in a BIC= 9e3. (a) Mean of the data (red) to the best fitting strong Allee on birth model mean (blue). (b) Variance of the data (green) to the best fitting strong Allee on birth model variance (blue).

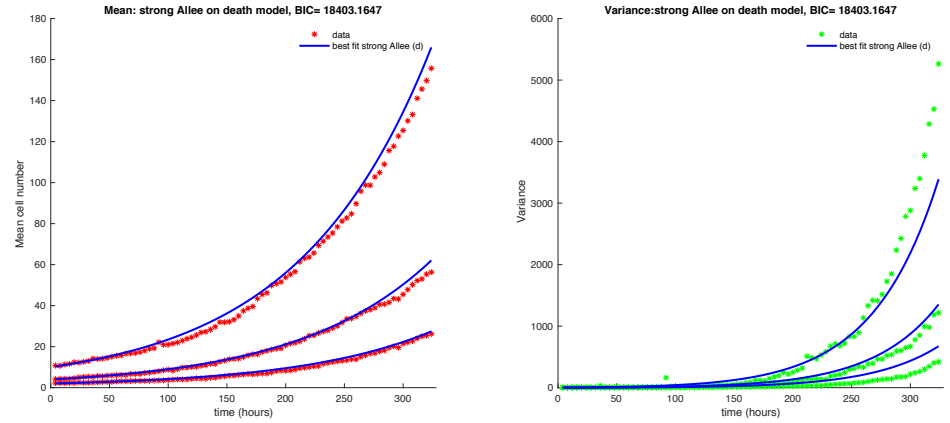

Figure 3.25: Data fit to strong Allee on death model results in a BIC= 1.8e4. (a) Mean of the data (red) to the best fitting strong Allee on death model mean (blue). (b) Variance of the data (green) to the best fitting strong Allee on death model variance (blue).

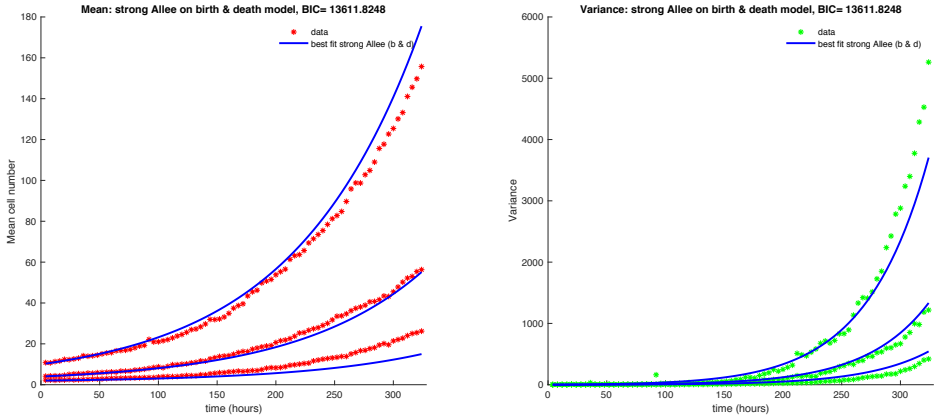

Figure 3.26: Data fit to strong Allee on birth & death model results in a BIC= 1.4e4. (a) Mean of the data (red) to the best fitting strong Allee on birth & death model mean (blue). (b) Variance of the data (green) to the best fitting strong Allee on birth & death model variance (blue).

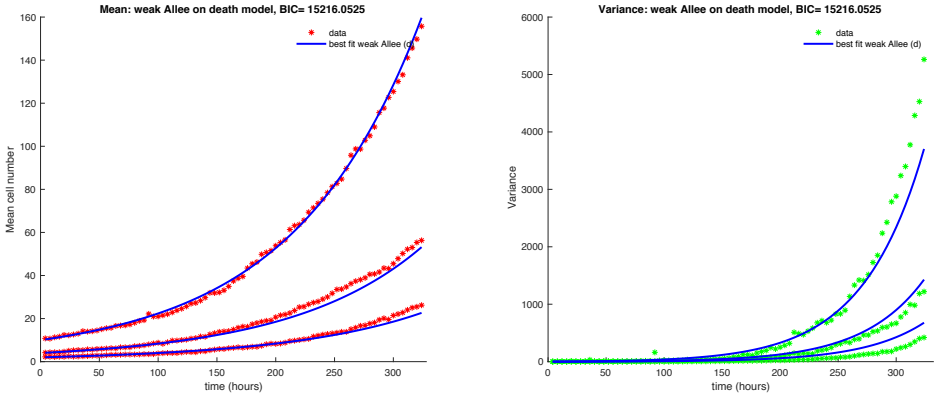

Figure 3.27: Data fit to weak Allee on death model results in a BIC= 1.5e4. (a) Mean of the data (red) to the best fitting weak Allee on death model mean (blue). (b) Variance of the data (green) to the best fitting weak Allee on death model variance (blue).

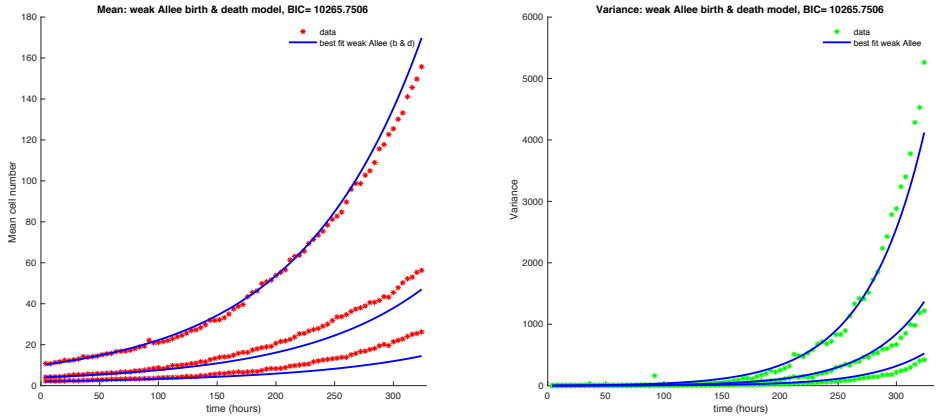

Figure 3.28: Data fit to weak Allee on birth & death model results in a BIC= 1.0e4. (a) Mean of the data (red) to the best fitting weak Allee on birth & death model mean (blue). (b) Variance of the data (green) to the best fitting weak Allee on birth & death model variance (blue).

Tables

Hypothesis	Model	Parameters	Behavior
No Allee effect	$\frac{dN}{dt} = gN$	g = intrinsic growth rate	Cells exhibit cell autonomous proliferation and will grow at a constant rate regardless of cell number
Strong Allee effect	$\frac{dN}{dt} = gN(1-\frac{A}{N})$	A = Allee threshold	If the number of cells is below the threshold A, the population will go extinct. If N is near A, the growth rate will be slowed until N>>A when the growth rate will approach g
Weak/strong Allee effect	$\frac{dN}{dt} = gN(1-\frac{A+\tau}{N+\tau})$	**τ** = overall shape parameter, as **τ** increases, per capita growth flattens	When A is positive, strong Allee effect, when **τ**>|A| and A is negative, the population will grow more slowly when N is near A, but will not go extinct

Table 3.3: Deterministic model structures to describe three distinct tumor growth dynamic model hypotheses.

Stochastic Model	Chemical Master Equation
Birth-death model	$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1}(t) - [d(n)+b(n)]p_n(t) + d(n+1)p_{n+1}(t)$
Strong Allee on birth model	$\frac{dp_n(t)}{dt} = [b(n-1)-(b-d)A]p_{n-1}(t) - [d(n)+b(n)-(b-d)A]p_n(t) + d(n+1)p_{n+1}(t)$
Strong Allee on death model	$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1}(t) - [d(n)+(b-d)A+b(n)]p_n(t) + [d(n+1)+(b-d)A]p_{n+1}(t)$
Strong Allee on birth & death model	$\frac{dp_n(t)}{dt} = [b(n-1)-\frac{(b-d)}{2}A]p_{n-1}(t) - [d(n)+b(n)]p_n(t) + [d(n+1)+\frac{(b-d)}{2}A]p_{n+1}(t)$
Weak Allee on birth model	$\frac{dp_n(t)}{dt} = [b(n-1)-(b-d)(n-1)\frac{(A+\tau)}{(n-1+\tau)}]p_{n-1}(t) - [d(n)+b(n)-(b-d)(n)\frac{(A+\tau)}{(n+\tau)}]p_n(t) + d(n+1)p_{n+1}(t)$
Weak Allee on death model	$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1}(t) - [d(n)+(b-d)\frac{(A+\tau)}{(n+\tau)}+b(n)]p_n(t) + [d(n+1)+(b-d)(n+1)\frac{(A+\tau)}{(n+1+\tau)}]p_{n+1}(t)$
Weak Allee on birth & death model	$\frac{dp_n(t)}{dt} = [b(n-1)-\frac{(b-d)(n-1)}{2}\frac{(A+\tau)}{(n-1+\tau)}]p_{n-1}(t) - [d(n)+b(n)]p_n(t) + [d(n+1)+\frac{(b-d)(n+1)}{2}\frac{(A+\tau)}{(n+1+\tau)}]p_{n+1}(t)$

Table 3.4: Master Equations to describe each of the stochastic model structures.

Text

Text 3.1: Stochastic model simulation using the Gillespie Algorithm

To evolve the stochastic model forward, we implement the Gillespie algorithm (Gillespie, 1977, 2014) by initializing the number of cells to begin with and the rate parameters that describe the probability of either a birth or death event. The Monte Carlo step is performed to generate random numbers that determine the next event (either a birth or a death) to occur as well as the time interval until that event occurs. The probability of a given event to be chosen is proportional to the reaction propensity and the time interval to the next event is exponentially distributed with mean of the reciprocal of the sum of the probability of any of the events occurring. This process is described below:

$$r1 = rand \in [0,1]$$
$$r2 = rand \in [0,1]$$
$$r1 < \frac{P(birth)}{P(birth) + P(death)}$$
$$N = N + 1$$
$$else$$
$$N = N - 1$$
$$\Delta t = \frac{-\log(r2)}{P(birth) + P(death)}$$

Where P(birth) and P(death) are specific to the stochastic model structure (See Table 3.4 for definitions that correspond to each model). These steps were repeated for up to 5000 repetitions. Because the time step was probabilistic, each stochastic trajectory was sampled to obtain a uniform time interval where the number of cells was recorded at each time based on the number at the event equal to or before the interval.

$$\frac{d\langle n \rangle}{dt} = (b-d)\langle n \rangle$$
$$\langle n \rangle(t) = \langle n \rangle_0 e^{(b-d)t}$$

This same derivation (Houchmandzadeh, 2009) was repeated to find up to the 4th moment, and the definition of the variance was used to derive the expected time-derivative of the variance use to identify the magnitude of the birth and death rate.

Text 3.2: Derivation of the moment-closure approximation for the first moment of the birth-death model

Starting from the CME for the birth-death model below, we apply the $\sum n^m$th operator (in the case of the first moment m=1) over the time derivative of the probability of their being n cells at time t to obtain the dime derivative of the expectation of n.

$$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1} - [d(n)+b(n)]p_n + d(n+1)p_{n+1}$$

$$\sum_{n=-\infty}^{n=\infty} np_n(t) = \langle n \rangle$$

$$\frac{d\left(\sum_{n=-\infty}^{n=\infty} np_n(t)\right)}{dt} = \frac{d\langle n \rangle}{dt}$$

We apply the $\sum n^m$th operator to each term on the RHS of the CME. We can then use the fact that the summation from -∞ to ∞ is the same over n-1, n+1, and n to transform the p_{n+1} and p_{n-1} to p_ns by transforming each term on the RHS so that n-1= n, and substituting into each multiplicative term accordingly as shown below:

$$\frac{d\left(\sum_{n=-\infty}^{n=\infty} np_n(t)\right)}{dt} = \sum_{n=-\infty}^{n=\infty} nb(n-1)p_{n-1} - \sum_{n=-\infty}^{n=\infty} n[d(n)+b(n)]p_n + \sum_{n=-\infty}^{n=\infty} nd(n+1)p_{n+1}$$

$$\sum_{n=-\infty}^{n=\infty} n = \sum_{n=-\infty}^{n=\infty} n-1$$

$$\frac{d\left(\sum_{n=-\infty}^{n=\infty} np_n(t)\right)}{dt} = \sum_{n=-\infty}^{n=\infty} (n+1)b(n)p_n - \sum_{n=-\infty}^{n=\infty} n[d(n)+b(n)]p_n + \sum_{n=-\infty}^{n=\infty} (n-1)d(n)p_n$$

Factoring out p_n and applying like terms:

$$\frac{d\left(\sum_{n=-\infty}^{n=\infty} np_n(t)\right)}{dt} = \sum_{n=-\infty}^{n=\infty} b(n)p_n - \sum_{n=-\infty}^{n=\infty} d(n)p_n$$

Applying the $\sum n^m$th operator gives the time derivative of the expected first moment in terms of the first moment of n itself. In this case, we can solve this analytically to obtain the moment-approach approximation for the mean cell number of the stochastic birth-death

process described by exponential growth with a growth rate equal to the birth rate minus the death rate.

$$\frac{d\langle n\rangle}{dt} = (b-d)\langle n\rangle$$
$$\langle n\rangle(t) = \langle n\rangle_0 e^{(b-d)t}$$

This same derivation (Houchmandzadeh, 2009) was repeated to find up to the 4th moment, and the definition of the variance was used to derive the expected time-derivative of the variance use to identify the magnitude of the birth and death rate.

Text 3.3 Confirmation that derivations of mean and variance for each model match the mean and variance from simulated data with known parameters

The moment-approximation derivations from the CME were confirmed to match the measured moments from simulated data from the Gillespie algorithm. In Figure 3.11, five thousand trajectories are simulated from the stochastic birth-death model (Eq. 5 &6) with an initial cell number on N_0=5 and a birth rate of b = 0.0238 cells/ hour and a death rate of d = 0.005 cells/hour (Figure 3.4A). The stochastic simulation trajectories were sampled every 4 hours, and the mean and variance in cell number were calculated at each time point (Figure 3.4B). Figure 4C shows the measured mean from simulated data and the expected mean are a near perfect match, and Figure 3.4D shows the measured and expected variance are as well. See Figures 3.11-3.16 for confirmation that the expected mean and variance from the remaining six stochastic models match the measured mean and variance from simulated data.

We repeated this procedure of simulating 5000 trajectories and computed the measured mean and variance for the remaining the strong Allee stochastic model family and the extended Allee model. We directly compared this to the solution of the moment approach approximation system of ODEs presented in Table 3.1.

Text: 3.4. Theoretical identifiability of the structural models using the differential algebra approach applied to the simple birth-death model as an example

The differential algebra approach (Meshkat, Sullivant and Eisenberg, 2015; Brouwer *et al.*, 2017) requires equations be written to describe the process being modeled and the measurements available. In this case, we are modeling the linear two compartment system of differential equations that describes the time evolution of the mean and the variance in the bulk cell number. Because our experimental system is able to capture a high number of growth trajectories at low initial cell densities, we are able to measure the mean and variance throughout time for each initial condition. For the birth and death model, this leads to the following set of model and measurement equations from Table 3.1.

$$\dot{n} = (b-d)n$$
$$\dot{V} = 2(b-d)V + (b+d)n$$
$$y_1 = n$$
$$y_2 = V$$

Next, we can rewrite these in terms of the measurable outputs y_1 and y_2. We solve for n from the derivative of V differentiate, then set this equal to the derivative of n.

$$y_1 = \frac{\dot{y}_2 - 2(b-d)y_2}{b+d}$$

$$\dot{y}_1 = \frac{\ddot{y}_2}{b+d} - \frac{2(b-d)\dot{y}_2}{b+d} = (b-d)y_1$$

$$0 = \frac{\ddot{y}_2}{b+d} - \frac{2(b-d)\dot{y}_2}{b+d} - (b-d)y_1$$

The coefficients in front of the measurable outputs are the identifiable parameter combinations, which we set equal to a_1, a_2, and a_3. We then use substitution and replacement to solve for the parameters b & d in terms of the identifiable combinations.

$$a_1 = \frac{1}{b+d}$$

$$a_2 = \frac{-2(b-d)}{b+d}$$

$$a_3 = b-d$$

$$b+d = \frac{1}{a_1}$$

$$b = \frac{1}{a_1} - d$$

$$a_2 = \frac{2(\frac{1}{a_1} - 2d)}{\frac{1}{a_1}}$$

$$d = \frac{2-a_2}{4}$$

$$b = \frac{1}{a_1} + \frac{a_2}{4} - \frac{1}{2}$$

If we can isolate each parameter in terms of identifiable combinations (a1, a2, and a3) alone, then the parameters are structurally identifiable, as is shown here by the isolation of b & d. Note that the unique identifiability of b and d, not just (b-d) would not be identifiable without the measurement of the variance (specifically because the time derivative of the variance is proportional to b+d), as is explained in Figure 3.10. We performed this analysis

for all seven stochastic model structures and found all parameters to be uniquely structurally identifiable.

Text 3.5: Model Selection using Bayesian Information Criterion and BIC weights

To investigate competing hypotheses about the underlying structure of tumor growth dynamics, the seven distinct stochastic models were compared using the Bayesian information criterion (BIC) (Raftery, 1999; Loos *et al.*, 2018). The BIC takes into account both goodness of fit of the model and penalizes for complexity of the model in terms of number of parameters, and has been shown to be an inexpensive approximation to the Bayes factors, which gives the favor of a model over another (Loos *et al.*, 2018). In order to ensure the method was not overweighing goodness of fit, the data was down-sampled from the true data collection interval of every 4 hours to every 36 hours to demonstrate that down-sampling changed the magnitudes of the BIC values but did not affect the order of the BIC values of each model relative to one another (Figure 3.20). To evaluate statistical significance between models with BIC values that were very close to one another, the methods presented in Waenmakers & Farrell et al (Wagenmakers and Farrell, 2004) of BIC weighting were used which are given by:

$$w_i(BIC) = \frac{\exp(-\frac{1}{2}\Delta_i(BIC))}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\Delta_k(BIC))}$$

From this equation, each model is assigned a relative weight, whose sum add to 1 based on the probability that it is the most parsimonious model to describe the data.

[4]Chapter 4: Integrating transcriptomics and machine learning with longitudinal data into a mathematical framework to describe and predict resistance

PREFACE

This work represents a truly integrative, collaborative project consisting of a number of separate but corroborating analysis by all members, past and present, of the Brock lab. The main crux of the project is built upon the lineage-traced scRNA-seq system developed by Aziz Al'Khafaji during his time as a graduate student (Al'Khafaji, Deatherage and Brock, 2018). This novel technological advancement enabled linking of phenotypes via single cell transcriptomes with lineage identity of each individual cell. The functional power of this technology is critical to the machine learning component of this project that will be described, essentially enabling us to "see the future" of individual cell fates. The experimental and computational work to generate these lineage-traced transcriptomics data sets is a separate project in itself, and was performed by Aziz Al'Khafaji, Daylin Morgan, Eric Brenner, and Russell Durrett. This process is both labor and time intensive, and I am very grateful for them for letting her "use" the data set in its complete form for the purpose of this project. Likewise, the ability to characterize with such breadth the treatment response dynamics of this cell line to a series of pulse-treatments of doxorubicin was also no small endeavor, this work was performed by Grant Howard,

---

[4] This chapter is based on a paper in submission that is available in pre-print form at:
Johnson, K.E., Brenner, E., Howard, G.R., Al'Khafaji A., Mo, W., Morgan, D., Gardner, A., Jarrett, A., Sontag, E. D., Yankeelov, T.E., Brock, A. (2020). Integrating multimodal data sets into a mathematical framework to describe and predict therapeutic resistance in cancer. bioRxiv 943738;
doi: https://doi.org/10.1101/2020.02.11.943738
Author Contributions:
KJ and AB designed the study; GH, DM, EB, AG, and AA performed experiments; WM curated the data; KJ, GH, DM, EB, AG, RD and WM analyzed the data; KJ performed mathematical modeling; ES, AJ, TY advised on mathematical modeling, KJ and AB wrote the manuscript with input from all authors; all authors read and approved the manuscript.

again as a part of a larger project to characterize treatment response in different treatment conditions. All authors contributed critical and meaningful feedback regarding the use of these data sets for this project, giving suggestions, explanations, and computational tools for handling the diverse data sets. The work outlined below describes a mathematical framework to integrate molecular level data with population-size data to improve our understanding of treatment response dynamics that can ultimately be leveraged to develop optimal therapeutic regimens.

ABSTRACT

A significant challenge in the field of biomedicine is the development of methods to integrate the multitude of dispersed data sets into comprehensive frameworks to be used to generate optimal clinical decisions. Recent technological advances in single cell analysis allow for high-dimensional molecular characterization of cells and populations, but to date, few mathematical models have attempted to integrate measurements from the single cell scale with other data types. Here, we present a framework that actionizes static outputs from a machine learning model and leverages these as measurements of state variables in a dynamic mechanistic model of treatment response. We apply this framework to breast cancer cells to integrate single cell transcriptomic data with longitudinal population-size data. We demonstrate that the explicit inclusion of the transcriptomic information in the parameter estimation is critical for identification of the model parameters and enables accurate prediction of new treatment regimens. Inclusion of the transcriptomic data improves predictive accuracy in new treatment response dynamics with a concordance correlation coefficient (CCC) of 0.89 compared to a prediction accuracy of CCC = 0.79

without integration of the single cell RNA sequencing (scRNA-seq) data directly into the model calibration. To the best our knowledge, this is the first work that explicitly integrates single cell clonally-resolved transcriptome datasets with longitudinal treatment response data into a mechanistic mathematical model of drug resistance dynamics. We anticipate this approach to be a first step that demonstrates the feasibility of incorporating multimodal data sets into identifiable mathematical models to develop optimized treatment regimens from data.

INTRODUCTION

The development of resistance to chemotherapy is a major cause of treatment failure in cancer. Intratumoral heterogeneity and phenotypic plasticity play a significant role in therapeutic resistance (Ferrall-Fairbanks *et al.*, 2019)(Syed *et al.*, 2019) and individual cell measurements such as flow and mass cytometry (Pyne *et al.*, 2009) and scRNA-seq (Islam *et al.*, 2014) have been used to capture and analyze this cell variability (Guo *et al.*, 2018; Kumar *et al.*, 2018; Wang *et al.*, 2019; Zhao *et al.*, 2019). Although attempts have been made to extract dynamic information from scRNA-seq *via* pseudo-time(Cho *et al.*, 2018) or RNA velocity approaches (Manno *et al.*, 2018), these high-throughput "omics" approaches come from cancer cell populations at a single time point. Snapshot information alone has provided immense insight to the field: illuminating novel molecular insight about distinct subpopulations (Al'Khafaji *et al.*, 2019), developing detailed hypothesis about population structure (Smalley *et al.*, 2019), and even demonstrating the ability to predict clinical outcomes (Ferrall-Fairbanks *et al.*, 2019).

However, outside of the field of differentiation (Stumpf *et al.*, 2017), most "omics" data sets have not been directly integrated with longitudinal population data—which are critical to understanding the dynamics of cancer progression.

Longitudinal treatment-response data in cancer have been used to calibrate mechanistic mathematical models of heterogeneous subpopulations (Matthew T McKenna *et al.*, 2018; Brady *et al.*, 2019; Smalley *et al.*, 2019) of cancer cells. These models describe cancer cells dynamically growing and responding to drug with differential growth rates and drug sensitivities. Knowledge of these model parameters have enabled the theoretical optimization of treatment protocols (Greene, Sanchez-Tapia and Sontag, 2018a; Gevertz, Greene and Sontag, 2019; Greene, Gevertz and Sontag, 2019), and have been applied successfully to prolong tumor control in both mice (Smalley *et al.*, 2019) and patients (Gatenby *et al.*, 2009; Brady *et al.*, 2019). Critical to the success of these modeling endeavors is the ability to identify and validate critical model parameters from available data (Prokopiou *et al.*, 2015). Identifiable and practical models are necessarily limited in their capacity to explain biological complexity based on the available longitudinal data, which is often limited to total tumor volume or total cell number in time. While we have evidence of complex relationships between distinct subpopulations of cells (Al'Khafaji *et al.*, 2019) that give rise to observed behavior, the ability to track these subpopulations longitudinally for use in model calibration and parameter estimation remains a challenge (Howard *et al.*, 2018).

One way to resolve this challenge would be to work with both types of data (the snapshot "omics" data sets to provide details of distinct subpopulations, and longitudinal

population-size data) and use them jointly to inform the calibration of a mechanistic model. In this study, we sought to develop a flexible framework for integrating informatics outputs from high-throughput single-cell resolution data with longitudinal population-size data to demonstrate the feasibility of utilizing multimodal data sources in mathematical oncology. The integration of single cell data into a mathematical modeling framework has been successfully employed in the field of differentiation by quantifying the changing proportion of cells in distinct cell states over time (Stumpf *et al.*, 2017). This approach has yet to be applied to cancer, where the effects of exponential growth and death due to drug exposure results in changes in phenotypic composition that are independent of directed transitions between cell states. To better understand these dynamics, we collect longitudinal population-size data in response to treatment with chemotherapy doxorubicin. We combine this with snapshots of lineage-traced scRNA-seq data and build a classifier to estimate phenotypic composition, via the proportion of sensitive and resistant cells, at distinct time points during treatment response. Despite differences in data acquisition, time resolution, and data uncertainty, we demonstrate that these two measurement sources can be used to estimate cell number in time and phenotypic composition in time, which can be compared to their corresponding model outputs. To reflect varying degrees of confidence in the measurement sources, we develop an integrated calibration scheme that relies on Pareto optimality and demonstrate that the phenotypic composition information is essential for the identifiability of model parameters from data. We validate the model results by demonstrating that they can accurately predict the response dynamics to new treatment regimens. We propose this framework as a crucial next step towards combining tumor

composition information with longitudinal treatment data to improve prediction and optimization of treatment outcomes.

MATERIALS AND METHODS

Experimental model and subject details

Cell culture

The human breast cancer cell line MDA-MB-231(ATCC) was used throughout this study. Cells were maintained in Dulbecco's Modified Eagle Medium (Gibco) and supplemented with 1% Penicillin-Streptomycin (Gibco) and 10% fetal bovine serum (Gibco) under standard culture conditions (5% CO_2, 37°C).

A subline of the MDA-MB-231 breast cancer cell line was engineered to constitutively express EGFP (enhanced green fluorescent protein) with a nuclear localization signal (NLS). Genomic integration of the EGFP expression cassette was accomplished through the Sleeping Beauty transposon system (Kowarz, Loescher and Marschalek, 2015). The EGFP-NLS sequence was obtained as a gBlock from IDT and cloned into the optimized sleeping beauty transfer vector containing the EGFP-NLS expression cassette and the pCMV(CAT)T7-SB100 plasmid containing the Sleeping Beauty transposase was co-transfected into a MDA-MB-231 cell population using Lipofectamine 2000. mCMV(CAT)T7-SB100 was a gift from Zsuzsanna Izsvak (Addgene plasmid #34879) (Mátés *et al.*, 2009). GFP+ cells were collected by fluorescence activated cell sorting. MDA-MB-231 cells are maintained in DMEM (Gibco), 10% fetal bovine serum (Gibco) and 200 μg/mL G418 (Caisson Labs). Cells were seeded into the center 60

wells of a 96 well plate (Trueline) at about 2000 cells per well. During the monitoring and treatment, plates were kept in the Incucyte Zoom, a combined incubator and time-lapsed microscope. Cells were fed fresh media every 2-3 days for up to 5 weeks. HEK293T cells were cultured in DMEM with GlutaMAX supplemented with 10% FBS, 4.5 g/L D-glucose, 110 mg/L sodium pyruvate, streptomycin (100ug/mL) and penicillin (100 units/mL).

Longitudinal treatment response data

The EGFP-labeled subline of MDA-MB-231 breast cancer cells were used for longitudinal treatment response. Cells were passaged into the center 60 wells of 96 well plates at a density of about 2000 cells per well. Two days later, cells were treated with a 24 hour pulse-treatment of doxorubicin at concentrations ranging from 0-1000 nM, with 6 replicates of each dose. Dosed media was applied to cells and treatment response was monitored using the Incucyte. After 24 hours, the dosed media was replaced with normal media and monitoring continued. Cells were fed fresh media every 2-3 days for the duration of the monitoring period (up to 5 weeks).

Lentiviral assembly

Lentiviral assembly was performed using the Lenti-Pac HIV Expression Packaging Kit (GeneCopeia). Two days prior to lentiviral transfection 1.5 million HEK293T cells were plated in a 10 cm tissue culture dish. Forty eight hours after plating, cells were 70−80% confluent and transfected with 9 µL of Lipofectamine 2000 (Thermo Fisher #11668027), 1.5 µg per well of PsPax2 (Addgene #12260), 0.4 µg/well of VSV-G (Addgene #8454), and 2.5 µg of Lenti-Pac HIV mix (GeneCopoeia). Media was replaced 24 hours post transfection with 10 mL DMEM supplemented with 5% heat inactivated FBS

and 20 μL TiterBoost (GeneCopoeia) reagent. Media containing viral particles was collected at 48 and 72 hours post transfection, centrifuged at 500g for 5 minutes, and filtered through a 45 μm poly(ether sulfone) (PES) low protein binding filter. Filtered supernatant was stored at −80 °C in aliquots for later use.

Barcode labeling

MDA-MB-231 cells were transduced with the Cropseq-BFP-WPRE-TS-hU6-N20 lentivirus in growth media with 1 μg/mL polybrene. After 48 hours of incubation, 1000 BFP+ cells were isolated by FACS to establish a population with initial diversity of ~1000 unique barcodes. To reduce the likelihood that two viral particles enter a single cell, the lentiviral transduction multiplicity of infection was kept below 0.1.

Drug treatment of barcoded cells for scRNAseq and recovery

Barcode labeled MDA-MB-231 cells (5 replicate wells) were treated with doxorubicin (550 nM) for 48 hours in growth media, washed and replaced with fresh growth media. Surviving cells were maintained in growth media and passaged up serially from 0.1 x 10^6 to 20 x 10^6 cells.

scRNA-seq

Cryopreserved samples from drug-naïve and two samples of doxorubicin-treated cells frozen at 7 and 10 weeks post-treatment were harvested, sorted by FACS to collect the BFP+ population. Cells were loaded into wells of a Chromium A Chip, and libraries were prepared using the 10XGenomics 3' Single Cell Gene Expression (v2) protocol. Paired end (PE) sequencing of the libraries was conducted using a NovaSeq 6000 with an S1 chip (100 cycles) according to the manufacturer's instructions (Illumina).

Plasmid assembly for isolation of lineages

After selecting the lineages of interest for isolation, an array of barcodes was assembled as described in (Al'Khafaji, Deatherage and Brock, 2018). Briefly, oligonucleotide pairs for the barcode of interest were ordered with specific overlapping sequences to both direct assembly of barcode array and integration into the plasmid for isolation. The barcode arrays were ligated, and gel purified to proceed with only a fully assembled array in cloning. The fully assembled barcode array was cloned into the BbsI site with standard restriction digest cloning. This double stranded barcode array was inserted into a plasmid backbone upstream of a minimal core promotor (miniCMV) and sfGFP to generate the Recall plasmid. This was repeated with individual barcodes of interest.

Recall of isolated sensitive and resistant clones by COLBERT

Barcoded MDA-MB-231 cells were seeded in 6 well plates and transfected using Lipofectamine 3000 (ThermoFisher) with 225 ng dCas9-VPR-Slim and 275 ng Recall Plasmid per well. Forty eight hours after transfection, GFP+ cells were single cell sorted by FACS into a 96 well plate and spun for 1 minute at 1000g. Sorted cells were expanded until 80% confluency and passaged into a single well of a 48 well plate. Upon first passage following sort, 1/6 of the cells or ~5000 live cells were resuspended in a PCR reaction mix to confirm lineage identity through PCR amplification and subsequent Sanger sequencing of barcode region.

Alignment to reference genome

The GTF file included with cellranger's GRCh38 3.0.0 reference was modified to create a "pre-mRNA" GTF file so that pre-mRNAs would be included as counts in the later analysis. Cellranger's (v3.0.2) *mkref* command was then used to create a pre-mRNA reference from the GTF file and a genome FASTA file from the GRCh38 3.0.0 reference. FASTQ files of the scRNA-seq libraries were then aligned to the new pre-mRNA reference using the *cellranger count* command, producing gene expression matrices. The matrices for the different samples were concatenated into a single matrix using the *cellranger aggr* command with normalization turned off, so that the raw counts would remain unchanged at this point.

Filtering and normalization

The filtered matrices produced by cellranger were loaded into scanpy (v1.4.4)(Wolf, Angerer and Theis, 2018). Cells were annotated by sample and lineage membership. Only cells meeting the following requirements were retained for further analysis: (a) a minimum of 10000 and maximum of 80000 transcript counts, (b) a maximum of 20% of counts attributed to mitochondrial genes, and (c) a minimum of 3000 genes detected. Genes detected in fewer than 20 cells were removed. Normalization was conducted based on the recommendations from multiple studies that compared several normalization techniques against each other (Büttner *et al.*, 2019; Luecken and Theis, 2019; Vieth *et al.*, 2019). In brief, three steps were performed: (a) preliminary clustering of cells by constructing a nearest network graph and using scanpy's implementation of Leiden community detection (Traag, Waltman and van Eck, 2019), (b) calculating size factors using the R package scran (L. Lun, Bach and Marioni, 2016), and (c) dividing

counts by the respective size factor assigned to each cell. Normalized counts were then transformed by adding a pseudocount of 1 and taking the natural log.

Regressing out cell cycle expression signatures

Using a list of genes known to be associated with different cell cycle phases (Tirosh *et al.*, 2019) , cells were assigned S-phase and G2M-phase scores. The difference between the G2M and S phase scores were regressed out using scanpy's *regress_out* function.

Quantification and statistical analysis

Machine learning of cell phenotypes

The machine learning classifier of sensitive and resistant cell phenotypes was built from the normalized, pre-processed single cell gene expression matrix with lineage identities from the pre-treatment time point only. For the cells in the pre-treatment sample, the lineage abundance at the pre-treatment time point (proportion of cells in each lineage) was calculated and compared to the lineage abundance at the combined post-treatment time points. If the lineage was not observed in the post-treatment time points, the lineage abundance post-treatment was assigned a zero. The change in lineage abundance (% post - % pre) was found for each lineage in the pre-treatment time point (See Fig. 3A). Based on this change in lineage abundance distribution, the pronounced tails of the distribution were used for classification. Cells whose lineage abundance change was positive, i.e. the lineage abundance increased post-treatment, were labeled as resistant in the pre-treatment time point. Cells whose lineage abundance change decreased by more than 5% were labeled as sensitive in the pre-treatment time point. This resulted in 815 cells and their corresponding 20,645 normalized gene expression levels being used to form the training set gene-cell

matrix containing a cell's gene expression vector and corresponding identity (with a 0 being sensitive and a 1 being resistant). This gene-cell matrix was then used to build a classifier capable of predicting the identity of new cells based on an individual gene expression vector.

A principal component classifier was built using the methods for eigendigit classification proposed in (Tunio *et al.*, 2018) for performing face recognition. In short, we perform principal component analysis of the gene-cell matrix from the labeled sensitive and resistant cells only. Each labeled cell's gene expression vector is projected into principal component space, made up of the gene loadings of each of the 20,645 genes and the optimal number of principal components. To classify new cells based on their gene expression and corresponding projection into PC space, the k closest cells to the new cell in the labeled class are found using a Euclidian distance metric. If the new cell's k neighbors give the cell a probability of being resistant of greater than a certain threshold, than that cell is classified as resistant. If it is less than the cut-off threshold, it is classified as sensitive. This process is repeated for all unclassified cells from the remaining pre-treatment time points and all of the post-treatment time points. All calculations of principle component coordinates and *knn*-probabilities were found using python's scanpy package. The number of components, neighbors, and threshold probability were optimized via coordinate optimization described below.

PCA classifier hyperparameter optimization

Using the labeled sensitive and resistant cells from the pre-treatment time point, 5-fold CV was used to split the cells into evenly class-balanced groups of training and testing

data sets. Coordinate optimization was then used to iteratively find the optimal number of both nearest neighbors (k) and number of principal components (n) for correctly identifying the class of each cell. Coordinate optimization works by essentially iteratively optimizing the two variables of interest, here k and n, until they no longer change values. In this case, we first set the number of principal components to a single value and iterated through a range of nearest neighbors to find the number which gave the highest mean AUC (area under the curve) over all 5 folds of cross validation (Figure. 4.8A). Once the optimal number of neighbors was found for that number of principal components, the number of neighbors was set to that value and the optimal number of principal components was varied over a range of values, and again the highest mean AUC over all 5 folds of cross validation was found (Figure 4.8B). Then we set the number of neighbors to this value and repeated the search for the optimal number of principal components. This process was repeated until the optimal number of neighbors and number of principal components no longer changed with each iteration. Using the optimal hyperparameters, we projected all of the labeled cells into the full classifier model and found the ROC curve for different probability thresholds for classifying cells as sensitive or resistant. While many appeared to be reasonable, we chose a threshold value of P(resistant)= 0.2 as our cut-off for calling a cell resistant, as this generated a realistic proportion of cells in each class at the pre-treatment time point.

Model of drug resistance dynamics

The mathematical model of drug-induced resistance, in which treatment exposure directly induced phenotypic transitions into the resistant cell state, was introduced in (Greene, Gevertz and Sontag, 2019). Their original model described sensitive cells (S) and

resistant cells (R) independently growing according to logistic growth and independently dying due to drug treatment ($u(t)$) *via* a log-kill hypothesis. The model includes an explicit role for the transition of sensitive cells into resistant cells via a rate of drug-induced resistance (α) which is modeled as a linear function of treatment $u(t)$. Additionally, their full model included additional terms of spontaneous, treatment-independent resistance (ε) proportional to the number of sensitive cells present, as well as a resensitization term (γ) describing treatment-independent transition from the resistant to the sensitive cell state.

$$\frac{\partial S}{\partial t} = r_S S\left(1 - \frac{S+R}{K}\right) - \big(\varepsilon + \alpha u(t)\big)S - d_s u(t)S + \gamma R$$

$$\frac{\partial R}{\partial t} = r_R R\left(1 - \frac{S+R}{K}\right) + (\varepsilon + \alpha u(t))S - d_R u(t)R - \gamma R$$

In order to have the best possible chance of identifying these model parameters from data, we simplified the original model. We assume that the treatment-independent transition into the resistant state (ε) and the resensitization (γ) are negligible, yielding the following system of equations.

$$\frac{\partial S}{\partial t} = r_S S\left(1 - \frac{S+R}{K}\right) - \alpha u(t)S - d_s u(t)S$$

$$\frac{\partial R}{\partial t} = r_R R\left(1 - \frac{S+R}{K}\right) + \alpha u(t)S - d_R u(t)R$$

Where r_S and r_R are the sensitive and resistant subpopulation growth rates and d_S and d_R are the sensitive and resistant subpopulation death rates, assumed to be linearly proportional to the effective dose ($u(t)$). We assume that the sensitive cells grow faster than the resistant cells so that $r_s > r_r$, as is consistent with the mechanism of action of cytotoxic

therapies targeting rapidly proliferating cells (Anderson *et al.*, 2006; Greene, Gevertz and Sontag, 2019). We assume $d_S > d_R$ as sensitive cells should die more quickly in response to drug than resistant cells, by definition. We modeled the effect of the pulse-treatments as single pulses of $u(t)$ whose maximum is given by the concentration of doxorubicin and whose effectiveness in time decays exponentially.

$$u(t) = k_1 C_{drug} e^{k_2 t}$$

The constants k_1 and k_2 were chosen so that $u(t)$ is scaled between 0 and 5 and so that the effective dose decays over a time scale consistent with experimental observations of doxorubicin fluorescent dynamics *in vitro* (McKenna *et al.*, 2017; Matthew T. McKenna, Weis, Quaranta, *et al.*, 2018). Numerical simulations of the forward model for a given treatment regimen were implemented in MATLAB using the backward Euler method.

To evaluate and compare the effect of treatment regimens on the cell population, we utilized an unbiased time-to-event metric proposed for use in evaluating treatment benefit in clinical trials (Johnson *et al.*, 2019), (there called TTB120 time to reach $1.2*N_0$), and here which we call critical time or t_{crit}, defined as the time to reach $2*N_0$. The longer the t_{crit}, the longer the "tumor burden" is held below this threshold, and therefore the more effective the treatment regimen. This critical time can be simulated for a given $u(t)$ in our model and can also be measured experimentally for most doses administered. We therefore use this output, as well as the phenotypic composition at t_{crit} ($\phi_S(t=t_{crit})$), as outputs for performing sensitivity analysis to assess the effect of parameters on the observed drug response.

Sensitivity analysis of model parameters

As part of the model development process, we performed a sensitivity analysis to assess the effect of individual model parameters on the model output. Although there are a number of choices to use for model outputs, we chose to capture the broad drug response of the population using the critical time (t_{crit}), and the phenotypic composition $\phi(t=t_{crit})$ at that time, as we expect these are two outputs we would feasibly observe in an experimental setting, as the time to population rebound and the phenotype observable via scRNAseq or some other phenotypic characterization. We first performed a global sensitivity analysis on the set of parameter bounds that were well outside the parameter ranges of the calibrated parameters and their associated errors. The results of the sensitivity analysis will reveal the most important parameters of the system, causing the greatest variation in outputs. This exercise should identify any model parameters that the model is insensitive to, and therefore may present opportunities to simplify the model to capture the same dynamics while reducing uncertainty by eliminating the number of free parameters to be fit. A Sobol's global sensitivity method is applied, which is a method that utilizes the analysis of variance (ANOVA) decomposition to define its sensitivity indices (Jarrett *et al.*, 2015). As a global method, random sampling is performed twice over the parameter space of the eight parameters (six free, two carrying capacities), with the number of parameters by N simulations matrices denoted by X and Z. The bounds of the global sensitivity analysis were chosen to be well outside of the 95% confidence intervals around each best fitting parameter from the profile likelihood analysis. The total effects are then calculated using the following:

$$\bar{S}_u = \frac{1}{2N\sigma^2} \sum_{j=1}^{N_{samps}} \left(f(x_j) - f(z_j^u, x_j^{-u}) \right)^2$$

Where σ^2 is the variance of the outputs from the first set of N random samples computed from evaluating over all x_j in X, and the function evaluations of $f(x_j)$ and $f(z_j, x_j^{-u})$ are the outputs (t_{crit} or $\phi(t=t_{crit})$) of the model at parameter values x_j compared to the function evaluated at parameter values z_j for one parameter, and x_j for all the remaining parameters. The total effects were calculated for each parameter value for outputs of both critical time (t_{crit}) and phenotypic composition ($\phi(t=t_{crit})$) for four doses ranging from 0 to 500 nM. Large sensitivity indices between parameters and model outputs characteristics indicate that small changes in the parameter values will result in large variations in the output behavior. For this investigation, to ensure the convergences of the indices, a base simulation size of N=5000 is chosen, resulting in (5000 x 2 x 4 doses x 2 outputs x 8 parameters=640,000) simulations to generate the indices. For this study, only the total effects of the model outputs of t_{crit} and $\phi(t=t_{crit})$ are reported. Specifically, the critical time and phenotypic composition at critical time is recorded for each random simulation and each dose, and per the Sobol method, the total effects indices derived from the variances of these outputs is calculated, which account for variations in individual parameters as well as additional effects resulting from the combined variation of parameters. A sensitivity cut-off of 0.05 is used, indicating parameters that cause less than 5% of the total variation of that model output.

To perform a local sensitivity analysis, we varied each parameter independently from a single parameter set chosen from the set of Pareto optimal sets. To perturb each parameter, we chose a high parameter value of two times its optimal value, and a low parameter value of half its optimal value. We used these high and low parameter values, holding all other parameters constant, and ran the forward model and recorded the response over a range of doxorubicin doses from 0-500 nM, for both the effect in critical time (t_{crit}) and phenotypic composition at critical time ($\phi(t=t_{crit})$). For each independent parameter perturbation, we computed a high and low sensitivity score for the the *i*th parameter, for the two model outputs (t_{crit} or $\phi(t=t_{crit})$) as:

$$S_i^+ = \sum_{j=1}^{n_{doses}} \left(f_j(x_{opt}) - f_j(x_{high})\right)^2$$

$$S_i^- = \sum_{j=1}^{n_{doses}} \left(f_j(x_{opt}) - f_j(x_{high})\right)^2$$

Which is the sum-squared difference between the output values (t_{crit} or $\phi(t=t_{crit})$) for each jth dose in the range of doses, for both the high and low parameter sets, for each *i*th parameter. The sum of the high and low sensitivity scores for each parameter were than ranked for the two outputs of t_{crit} and $(\phi(t=t_{crit}))$. This analysis reveals the most important parameter in driving changes in output behavior of the model locally around the best fitting parameters.

Model fitting with multiple measurement sources

To perform model fitting, we used two sources of measurement data: cell number in time ($N(t)$) in response to the pulsed doxorubicin treatments, and estimates of the phenotypic composition, $\phi(t)$, at three time points total (before and two post-treatment). The data were collected in two separate experimental settings, with two different carrying capacities, which we refer to as K_N and K_ϕ. The longitudinal cell number data was recorded in 96 well plates, resulting in a different carrying capacity than the lineage-traced single cell RNA sequencing experiment in which the population was expanded out to a 15 cm dish due to the need for large cell numbers for running on the 10x Genomics system. The carrying capacity of the longitudinal data, K_N, was found by fitting the untreated control to a logistic growth model and allowing both the effective growth rate of the total population (g_{eff}) and K_N to be fit to the data (See Figure 4S.11).

$$\frac{\partial N}{\partial t} = g_{eff} N \left(1 - \frac{N}{K_N}\right)$$

We set this carrying capacity in the model going forward for fitting the longitudinal data. For the carrying capacity of the single cell RNA sequencing experiment, K_ϕ, we used Thermo-Fisher published "Useful Numbers for Cell Culture" as an estimate(Thermo Fisher Scientific, no date), where the manufacturer cites the number of cells at confluency of 20 million cells. Going forward, we fit the remaining 6 parameters of θ=[ϕ_0, r_S, r_S/r_R *ratio*, α, d_S, d_R/d_S *ratio*] where these represent: the initial fraction of sensitive cells prior to treatment, the sensitive cell growth rate, the ratio of the resistant to sensitive cell growth rate, the rate of drug-induced resistance, the sensitive cell death rate, the ratio of resistant to sensitive cell death rates, respectively. All six parameters were found to be globally

sensitive in one or more of the treatment conditions when looking at either t_{crit} or $\phi(t=t_{crit})$, and so we decided it was reasonable to try to fit them all from the observed data. We note r_S/r_R ratio and d_S/d_R ratio are used for ease of parameter estimation. Since we assumed that $r_r<r_s$ and $d_r<d_s$, we can search the ratio, r_S/r_R ratio = r_R/r_S and d_S/d_R ratio = d_R/d_S, between 0 and 1 when performing parameter estimation.

To estimate the model parameters θ, we used both measurement sources $N(t)$ and $\phi(t)$ and a regularization term, λ, which is allowed to vary between 0 and 1 to reflect varying degrees of confidence in the two measurements sources. The data were fitted using a weighted-sum-of-squares-residual function described below:

$$J(\theta)=\frac{\lambda}{n_{\phi(t)}}\sum_{j=1}^{n_{\phi(t)}}\frac{(\widehat{\phi_j}-\phi_i(\theta,u_j))^2}{\sigma_{\phi_j}{}^2}+\frac{(1-\lambda)}{n_{N(t)}}\sum_{k=1}^{n_{doses}}\sum_{i=1}^{n_{N(t)}}\frac{(\widehat{N_{i,k}}-N_i(\theta,u_{i,k}))^2}{\sigma_{N_{i,k}}{}^2}$$

For the $N(t)$ data, the uncertainty in the data (σ^2_N) at each time point was quantified using the standard deviation of the cell number over the six replicate wells. For the uncertainty in the $\phi(t)$ estimates, we compute the Bernoulli sample variance of

$$\sigma_\phi^2=\frac{\phi(1-\phi)}{n}$$

where n is the number of Bernoulli samples (which here is the number of cells in the data set) at each of the three time points. However, this leads to an underestimate in the uncertainty in the $\phi(t)$ estimates, which depend significantly on where the threshold is chosen. For this reason, we added an uncertainty term of technical noise σ_{tech}=0.01 to this estimate. In reality, the magnitude of the uncertainty in the $\phi(t)$ is not necessarily known,

and the introduction of the regularization term, λ, in practice allows us to vary the degree of certainty we have in each measurement source relative to the other.

The key feature of the introduction of the regularization term λ means that we can tune the joint objective function to favor minimizing error in $N(t)$ and $\phi(t)$. In other work in the biomedical field using multi-objective function optimizations, the number of data points from each measurement source is typically similar, as most data is acquired longitudinally (Jarrett, Bloom, *et al.*, 2018). However, in this case we have significantly higher time and dose resolution in our $N(t)$ data (472 data points) compared to our $\phi(t)$ data (3 data points), and thus chose to include normalization terms in our objective function (Eq. 3) to account for the different resolutions of the data $N(t)$ and $\phi(t)$ data. Because the data come from distinct measurement sources, the robust quantification of comparative uncertainty is not known *a priori*, as we do not intuitively know whether or not the $\phi(t)$ estimates from scRNA-seq are inherently more or less reliable than the longitudinal population size data. We expect this problem to be present for any measurements taken from different measurement sources. Thus, we introduce the regularization term λ, which enables tuning of the certainty in favor of one measurement source over the other. We observe a trade-off in goodness of fit where if we assign a high value to λ, very close to 1, this puts more confidence on our $\phi(t)$, and if we assign a lower value to λ, this puts more confidence in our $N(t)$ data, and favors minimizing the error in that fit.

We first perform parameter estimation using weighting normalized only by the number of data points, with a value of $\lambda = 0.5$ which we call λ^*. We use a multistart search

algorithm where we randomly initialize the guess of the initial parameter vector over a range of reasonable parameter space for 100 initial parameter sets. We use the *fminsearch* function in MATLAB to search for a set of parameters, θ, that minimized $J(\theta)$ for $\lambda=\lambda^*$ for each initial guess. We then select the parameter set that produces the lowest objective function value, $J(\theta)$. The results of this optimization are presented in order to show an example of a single calibrated parameter set compared to the observed data, as well as to test the identifiability of those parameter values, which we call θ^*. This set of parameter values was also used for the local sensitivity analysis.

In order to allow for flexibility and generalizability of the approach for multimodal data sets, we sought to find more than a single optimal parameter set, but a "front" of solutions that could take into consideration the potentially varying degrees of confidence in the two types of measurement sources. We pulled from the field of economics to introduce a concept known as Pareto optimality (Censor, 1977), in which our set of Pareto optimum parameter sets reflects solutions in which an improvement in the fit to $N(t)$ leads to a trade-off resulting in a worse fit in $\phi(t)$. To find the Pareto front set of solutions, we varied the regularization term λ from one which only considers the $N(t)$ data ($\lambda=0$), to one which only considers the weighs the $\phi(t)$ data ($\lambda=1$). We generated a vector of 1000 ordered λ values and iterated through 1000 optimizations at each value of λ. We used the concept of homotopy continuation (Coetzee and Stonick, 1996) to initialize the guess for each optimization as the best fitting parameter set θ from the previous iteration. For each optimization, we recorded all of the parameter values, the sum-of-squares error in $N(t)$, the

sum-of-squares error in $\phi(t)$ the CCC in $N(t)$, and the CCC in $\phi(t)$. The results of the initial optimization are shown in Supp. Fig. S4.4A, colored by their λ value. Next, we filtered the parameter sets, by only keeping those whose parameter values led to a CCC in both $N(t)$ and $\phi(t)$ greater than 0.8 (Figure 4.11B). From that filtered parameter set, we then found the Pareto boundary by removing any parameter sets where there existed another parameter set with a lower error in $N(t)$ and $\phi(t)$ (Figure 4.11C). The resulting parameter sets formed a front, where as we increase the regularization term λ, the error in $\phi(t)$ fit to the data decreased as the error in the $N(t)$ data increased. With this set of Pareto front parameter sets, we looked at the individual parameter values and examined how they varied as we varied λ and thus as we improved the accuracy in fit to one data set over another (Figure 4.12). We could then look at the distribution of parameter values to see which parameter values were stable across objective function weightings, and which were most dependent on the weight of the data sets relative to one another (Fig 4.6L-Q, Figure 4.13). The parameter values chosen all fell well within the 95% CI of θ^* (Figure 4.13).

Structural identifiability of model parameters

We will demonstrate the structural identifiability of the individual model parameters using the differential algebra approach. Structural identifiability of a model and its parameters from a set of measurable outputs tells us that in theory, given perfect data, it is possible to uniquely identify model parameters. Structural identifiability is a pre-requisite for practical identifiability of model parameters from observed data. We start by

presenting the non-dimensionalized model and measurement equations, assuming we can measure both N(t) and ϕ(t).

$$\frac{\partial S}{\partial t} = (1 - (S + R))S - \alpha u(t)S - d_s u(t)S$$

$$\frac{\partial R}{\partial t} = p_R(1 - (S + R))R + \alpha u(t)S - d_R u(t)R$$

$$N(t) = S(t) + R(t)$$

$$\phi(t) = \frac{S(t)}{S(t) + R(t)}$$

We assume all parameters are non-negative and $0 < p_r < 1$ represents the relative growth rate of the resistant population with respect to the sensitive population scaled by the carrying capacity, and $p_r < 1$ assumes that resistant cells grow more slowly than sensitive cells. In work by Greene et al (Greene, Sanchez-Tapia and Sontag, 2018a), they demonstrate that, if they assume d_r=0, i.e. resistant cells are not killed by drug, and that the initial state of the population is completely comprised of sensitive cells (i.e. N_0=S_0), than the remaining parameters are uniquely identifiable from observations of total cell number alone.

We would like to extend this analysis by determining the identifiability of a new experimental system in which not only can $N(t) = S(t) + R(t)$ be observed, but so also can the fraction of cells in each state over time, here denoted as $\phi(t)$. Under these circumstances, we want to test the identifiability of the model which now allows for a net-positive death rate due to drug, d_R, and can have any composition of initial sensitive and resistant cells.

We follow the same arguments outlined in (Greene, Sanchez-Tapia and Sontag, 2018a), along with the complete explanation of the approach with illustrative examples, for the case of multiple outputs from (Sontag, 2017). We start by formulating the dynamical system relevant to our in vitro experimental system. Of note, even though we separately measure $N(t)$ and $\phi(t)$ at discrete time points, since this analysis is for structural identifiability and assumes perfect, noise-free data, we will transform the observable outputs of $N(t)$ and $\phi(t)$ into:

$$S(t) = \phi(t)N(t)$$

$$R(t) = (1 - \phi(t))N(t)$$

Treatment is initiated at time t=0, at which we make no assumptions about the composition of the population such that $S(0) = S_0$, $R(0) = R_0$. Here $0<S_0+R_0<1$. We note this is due to the non-dimensionalization in which we now track the proportion of confluent cells i.e. $S(t) = \frac{S\prime(t)}{K}$ and $R(t) = \frac{R\prime(t)}{K}$ (see (Greene, Sanchez-Tapia and Sontag, 2018a)) for additional details. We can now formulate our system in input/output form as:

$$\dot{x}(t) = f\big(x(t)\big) + u(t)g\big(x(t)\big)$$

$$x(0) = x_0$$

Where f and g are:

$$f(x) = \begin{pmatrix} (1 - (x_1 + x_2))x_1 \\ p_r\big(1 - (x_1 + x_2)\big)x_2 \end{pmatrix}$$

$$g(x) = \begin{pmatrix} -(\alpha + d_s)x_1 \\ \alpha x_1 - d_r x_2 \end{pmatrix}$$

and $x(t) = (S(t), R(t))$. As is standard in control theory, the output is denoted by the variable y which in this work corresponds to $S(t)$ and $R(t)$ obtained from the transformations of the measured variables $N(t)$ and $\phi(t)$

$$y_1(t) = h_1\big(x(t)\big) = x_1(t)$$

$$y_2(t) = h_2\big(x(t)\big) = x_2(t)$$

A system in this form is said to be uniquely structurally identifiable if the map $(p, u(t)) \rightarrow (x(t,p), u(t))$ is injective (Brouwer *et al.*, 2017; Sontag, 2017; Eisenberg, 2019), where p is the vector of parameters to be identified. In this instance $p = (S_0, R_0, d_s, d_r, \alpha, p_r)$, the initial states and the parameters. Local identifiability and non-identifiability correspond to the map being finite-to-one and infinite-to-one, respectively. Our objective is then to demonstrate unique structural identifiability for model system and hence recover all parameter values p from the assumption of perfect, noise-free data.

To analyze identifiability, we utilize results appearing in (Greene, Sanchez-Tapia and Sontag, 2018a) and (Sontag, 2017), where a differential-geometric perspective is used. For the structural identifiability, we hypothesize that we have perfect (hence noise-free) input-output data is available of the form of y_1 and y_2 and its derivatives on any interval of time. We then, for example, make measurements of:

$$y_1(0) = h_1(x_1(0))$$

$$\dot{y_1}(0) = \frac{\partial}{\partial t}\Big|_{t=0} h_1(x_1(t))$$

$$y_2(0) = h_2(x_2(0))$$

$$\dot{y_2}(0) = \frac{\partial}{\partial t}\Big|_{t=0} h_2(x_2(t))$$

We can relate their values to the unknown parameter values p. If there exists inputs $u(t)$ such that the above system of equations may be solved for p, the system is identifiable. The right-hand sides of the above the equation for $x(t)$ may be computed in terms of the Lie derivatives of the vector fields f and g. The Lie differentiation L_xH of a function H by a vector field X is given by:

$$L_xH(x) = \nabla H(x) \cdot X(x)$$

Iterated Lie derivatives are well-defined, and should be interpreted as the function composition, so that for example $L_yL_xH(x) = \mathrm{L_y}(L_xH)$ and $L_x^2H(x) = L_x(L_xH)$.

Defining observable quantities at the zero-time derivatives of the generalized output $y = h(x)$:

$$Y(x_0, U) = \frac{\partial^k}{\partial t^k}\Big|_{t=0} h(x(t))$$

Where $U \in R^k$ is the value of the control $u(t)$ and its derivatives evaluated at $t = 0$: $U = (u(0), u'(0), \ldots u^{k-1}(0))$. The initial conditions x_0 appear due to evaluation at t=0. The observation space is then defined as the span of the $Y(x_0, U)$ elements:

$$F_1 = span_R\{Y(x_0|U) \in R^k, k \geq 0\}$$

We also defined the span of iterated Lie derivatives with respect to the output vector fields $f(x)$ and $g(x)$:

$$F_2 := span_R\{L_{i1}, \ldots L_{ik}h_j(x_0)|(i_1, \ldots i_k) \in \{g, f\}^k, k \geq 0, j \in \{1,2\}\}$$

As is outlined in (Sontag, 2017), (Wang and Sontag, 1989) proved that F_1=F_2, so that the iterated Lie derivatives F_2 may be considered as the set of "elementary observables". Hence, identifiability may be formulated in terms of the reconstruction of parameters p from elements in F_2. Parameters p are then identifiable if the map

$$p \to \{L_{i1}, \dots L_{ik} h_j(x_0) | (i_1 \dots i_k) \in \{g, f\}^k, k \geq 0, j\, j \in \{1,2\}\}$$

Is one-to-one. For the remainder of this analysis, we investigate the mapping defined here, because if one can reconstruct the values of p from the elementary observables (evaluated at the initial state), we can uniquely identify the parameters. This enables us to find the Lie derivatives for the two outputs $h_1(x)$ and $h_2(x)$, which will be found in terms of the parameters p and x_1 and x_2. Then we can recall the evaluation at t=0 given by $x_0 = (S_0, R_0)$, and our ability to observe these at t=0 allows us to set $x_1 = S_0$ and $x_2 = R_0$ and isolate the parameter p recursively from the observables and the Lie derivatives.

Using the input-output system written in terms of f and g we can write the following Lie derivatives:

$$L_f h_1 = (1 - \mathrm{x}_1 - \mathrm{x}_2)\mathrm{x}_1$$

$$L_f h_2 = p_r(1 - \mathrm{x}_1 - \mathrm{x}_2)\mathrm{x}_2$$

$$L_g h_1 = (\alpha + \mathrm{d}_s)\mathrm{x}_1$$

$$L_g h_2 = \alpha x_1 - d_r x_2$$

$$L_f L_g h_2 = \alpha x_1(1 - x_1 - x_2) - d_r p_r x_2(1 - x_1 - x_2)$$

Recursively solving using $x_0 = (S_0, R_0)$ to find the parameters p:

$$S_0 = h_1(x_0)$$

$$R_0 = h_2(x_0)$$

$$p_r = \frac{L_f h_2}{R_0(1 - S_0 - R_0)}$$

$$d_r = \frac{1}{R_0(1 - p_r)}\left(\frac{L_f L_g h_2}{1 - S_0 - R_0} - L_g h_2\right)$$

$$\alpha = \frac{L_g h_2 + d_r R_0}{S_0}$$

$$d_s = \frac{L_g h_1}{S_0} - \alpha$$

Since $F_1 = F_2$, all of the above Lie derivatives are observable via appropriate treatment protocols. Thus by incorporating knowledge of $\phi(t)$, all parameters in system 1 are structurally identifiable. This represents an improvement over the identifiability with $N(t)$ alone as a measurable output and allows us to introduce a non-zero d_R parameter, which we have reason to believe based on experimental evidence, is the more biologically relevant scenario.

RESULTS

Utilizing a Model of Sensitive and Resistant Subpopulations to Describe and Optimize Drug Response Dynamics

To describe and predict the dynamics of cancer cells in response to treatment, we use a mechanistic model that describes sensitive and resistant cell subpopulations growing,

dying, and transitioning from the sensitive, S, to resistant, R, state as a direct result of treatment (Greene, Gevertz and Sontag, 2019).

$$\frac{\partial S}{\partial t} = r_S S\left(1 - \frac{S+R}{K}\right) - \alpha u(t) S - d_S u(t) S$$

$$\frac{\partial R}{\partial t} = r_R R\left(1 - \frac{S+R}{K}\right) + \alpha u(t) S - d_R u(t) R$$

(Eq. 1)

In this model (Fig 4.1A), sensitive and resistant cells grow *via* a logistic growth hypothesis at their own intrinsic growth rates (r_S and r_R) and a joint carrying capacity (K), which will either take the value of K_N for the carrying capacity of the cells in the longitudinal treatment experiment or K_ϕ for the carrying capacity of the cells in the scRNA-seq experiment. Sensitive and resistant cells are killed by the drug at a rate of d_S and d_R respectively, that is proportional to the number of cells in each subpopulation and the effective dose, $u(t)$, following the log-kill hypothesis. By definition, we set $d_S > d_R$ such that sensitive cells will be more susceptible to death due to treatment than resistant cells. Treatment drives cells from the sensitive subpopulation into the resistant subpopulation at a rate α, which is linearly proportional to the number of sensitive cells present and $u(t)$.

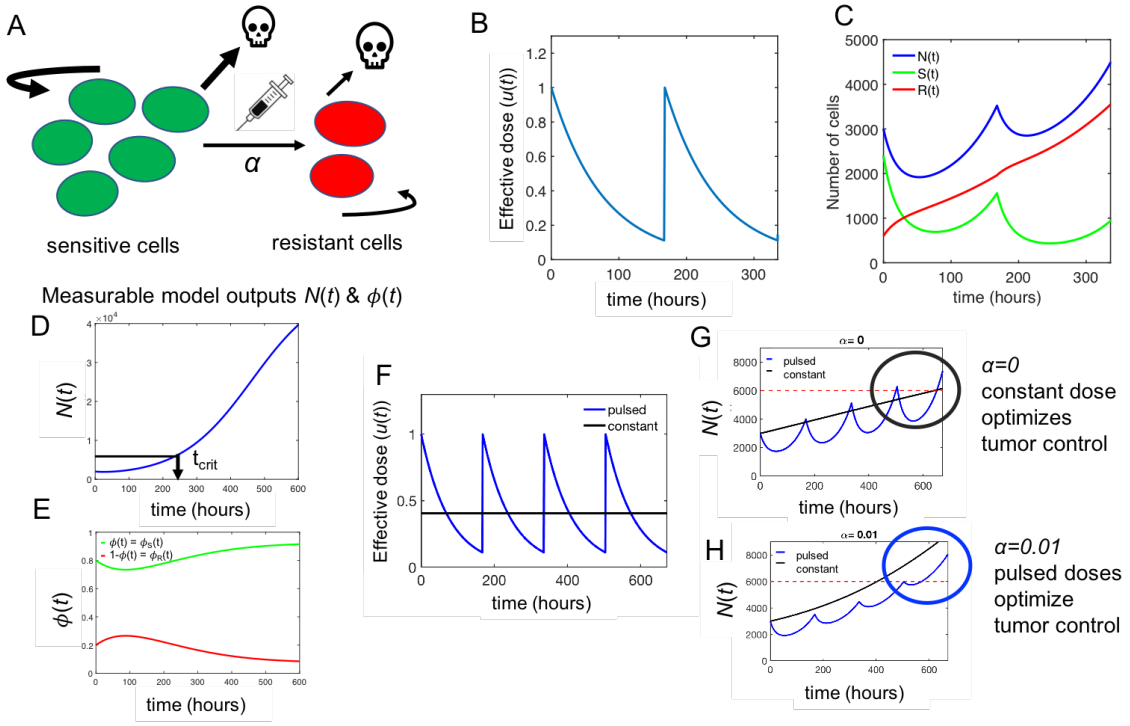

Fig 4.1. Mathematical Model of Treatment-induced Resistance and its Implications. A. Sketch of the model structure (Eq 1). The model describes sensitive and resistant subpopulations growing exponentially at independent growth rates. In response to treatment, sensitive and resistant cells are killed by the drug. The exposure to drug drives sensitive cells into the resistant phenotype. B. Input effective dose dynamics ($u(t)$) for pulse treatment of doxorubicin chemotherapeutic, where exponential decay is assumed (Eq. 2). C. Example of model predicted tumor dynamics under repeated pulse treatments. Sensitive (green) and resistant (red) subpopulation dynamics are predicted by the model. D. Example model predicted total cell number in time in response to a single pulse treatment. The efficacy of a treatment regimen is quantified by the time to reach $2*N_0$, which we call t_{crit} with a longer t_{crit} indicating a more effective treatment. Experimentally, we can only measure total cell number longitudinally. E. Fraction of cells that are sensitive (green) and resistant (red) in the population over time in response to a single pulse treatment. The phenotypic composition is measured using single cell transcriptomics at discrete time points. F. Pulsed (blue) and constant (black) effective dosing regimens ($u(t)$). The constant dose is equal to the average of the pulsed dose over time for ease of comparison (see text for details). G. Example trajectory of total cell number in time for a constant dose (black) and a pulsed dose (blue) for the case where there is no drug-induced resistance ($\alpha = 0$), indicating that optimal tumor control (longer critical time) is reached for the constant dose (black) compared to the pulsed dose (blue). H. Example trajectory of total cell number in time for a constant dose (black) and a pulsed dose (blue) for the case where the drug does induce resistance ($\alpha > 0$), indicating that in this case the optimal tumor control is reached by applying pulse treatments

To investigate the effect of different treatment regimens, we make a simple assumption about the pharmacokinetics of pulsed drug treatments, assuming exponential decay of the effective dose, $u(t)$, of the drug, as has been shown by others in greater detail (McKenna *et al.*, 2017; Matthew T. McKenna, Weis, Quaranta, *et al.*, 2018).

$$u(t) = k_1 C_{drug} e^{k_2 t},$$

(Eq. 2)

where C_{drug} is the concentration of doxorubicin in nM, k_1 is a scaling factor used to non-dimensionalize the effective dose, and k_2 is an estimated rate of decay of the effect of doxorubicin pulse-treatment on breast cancer cells. The effective dose decays over a time scale consistent with experimental measurements of doxorubicin fluorescence dynamics *in vitro* (McKenna *et al.*, 2017; Matthew T. McKenna, Weis, Quaranta, *et al.*, 2018). An example of the model-predicted treatment response dynamics (Fig 4.1C) for a pulse treatment given once every week (Fig 4.1B) demonstrates the response and relapse trajectory in cell number in time ($N(t)$ in blue), along with the underlying phenotypic dynamics of $S(t)$ and $R(t)$. The result of the treatment is that cells in the sensitive population either die or transition to the resistant state, leading to an increase in $R(t)$ over time even as $S(t)$ decays and rebounds. For numerical simulations of Eq. 1, we refer to Methods: Model of drug resistance dynamics.

While we can model the dynamics of heterogeneous subpopulations in terms of number of cells in each phenotypic state, most experimental or clinical workflows only allow for measurement of the total cell number ($N(t)$) over time (Fig 4.1D), as single

markers of resistance cannot usually be tracked throughout treatment. One unbiased metric to evaluate the response of cell populations to different drug treatments is to measure the time to return to some multiple of the initial cell number (Johnson *et al.*, 2019). We define the critical time (t_{crit}) as the time it takes for the total cell number to reach double the initial cell number at the onset of treatment (Fig 4.1D). This metric has been shown to be consistent with "patient benefit" in comparing treatment protocols in pharmacology (Johnson *et al.*, 2019). We employ it here as a single endpoint to evaluate the impact of a treatment on a cell population and to evaluate our model's predictive capabilities as compared to experimentally measured values of critical time. For a given treatment, while we may not feasibly be able to monitor resistant and sensitive cell number longitudinally, we can estimate the phenotypic composition, which we define here by the sensitive cell fraction, $\phi_S(t)$ (or simply $\phi(t)$), which we will use as a shorthand in the remainder of the manuscript), throughout treatment response from our model (Fig 4.1E). Model outputs of $N(t)$ and $\phi(t)$ can be used directly to compare to measurements of cell number in time and phenotypic composition in time following a drug treatment. A full description of the parameters in the model system are described in Table 4.1.

Parameter	Description	Units	Determination
r_S, r_R	Growth rate of sensitive and resistant cell subpopulations	hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
α	Drug-induced rate of transition from sensitive to resistant state	nM^{-1} x hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
d_S, d_R	Death rate of sensitive and resistant cell populations due to drug, $d_R < d_S$	nM^{-1} x hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
ϕ_0	Initial proportion of sensitive cells	number of cells	Fit from $N(t)$ & $\phi(t)$ data
K_N	Carrying capacity for the longitudinal treatment experiment performed in a 96 well plate to measure $N(t)$	number of cells	Fit from $N(t)$ untreated control
K_ϕ	Carrying capacity of the scRNAseq experiment performed in a 10 cm dish to measure $\phi(t)$	number of cells	Fixed
t_{crit}	Time for the number of cells to return to two times the initial cell number (N_0)	hours	Data, fit, and predicted
k_1	Scaling factor to non-dimensionalize concentration in nM of doxorubicin	nM^{-1}	Fixed
k_2	Estimated rate of decay of effect of doxorubicin after pulse-treatment	hour $^{-1}$	Fixed

Table 4.1. Description of model parameters to describe resistance dynamics. Descriptions of the parameters either from measured data (Data), fit of the model to the $N(t)$ (Fit from $N(t)$) or $\phi(t)$ (Fit from $\phi(t)$), the model assumptions (Fixed), or predicted from the parameter estimation from the fitted model (Predicted). We fit for six free parameters in the calibration scheme, as listed by the first four rows of the table.

Previous work has demonstrated the theoretical implications of treatment-induced resistance on identifying optimal treatment regimens (Greene, Gevertz and Sontag, 2019). Here we also found that for a resistance-preserving therapy (i.e., $\alpha = 0$), a constant dosing regimen optimizes tumor control (black line Fig 4.1F), leading to a longer critical time than the pulsed treatment (Fig 4.1G), whereas for a resistance-inducing therapy (i.e., $\alpha > 0$) a pulsed treatment regimen (blue line Fig 4.1F) optimizes tumor control (Fig 4.1H). To

compare the effects of a constant versus pulsed dose, we simulated the effect of a constant dose (black line Fig 4.1F) equal to the mean value over the time interval simulated of the pulsed dose (blue line Fig 4.1F) in an attempt to reflect realistic toxicity constraints that would be present in a clinical setting when developing treatment regimens. This analysis, as well as further work to utilize this modeling framework to develop optimal treatment protocols (Greene, Sanchez-Tapia and Sontag, 2018a) indicates that identifying these model parameters is essential to implementing more sophisticated treatment strategies in a practical clinical setting. While (Greene, Sanchez-Tapia and Sontag, 2018a) show that the critical model parameters are theoretically structurally identifiable from population size data alone, we seek to demonstrate how this model can be practically identified from *in vitro* data using both longitudinal population size data ($N(t)$) and snapshot outputs of the phenotypic composition ($\phi(t)$) at a few time points, enabled by recent advances in lineage tracing (Al'Khafaji, Deatherage and Brock, 2018; Al'Khafaji *et al.*, 2019) and scRNA-seq technologies. We present this project workflow in the experimental setting as proof-of-concept of the ability to properly identify key model parameters from multimodal data sets, with the hopes that the approach of integrating snapshot with longitudinal data sets will eventually be brought to the clinic to develop optimized treatment regimens for existing therapeutic agents.

To demonstrate the feasibility of integrating multimodal data sources into a cohesive modeling framework, we employ an experimental *in vitro* model system of MDA-MB-231 triple negative breast cancer cells exposed to the chemotherapeutic doxorubicin. The combined experimental-computational workflow (Fig 4.2) starts by tagging individual

cells with unique barcodes that are integrated into the genome and expressed as sgRNA's; this COLBERT cell barcoding platform has been described previously (Al'Khafaji, Deatherage and Brock, 2018). The barcode-labeled cell population is expanded to generate the naïve population for these studies (305 unique barcodes represents 305 clonal subpopulations). Cells are then treated with doxorubicin (LD95, 550 nM) for 48 hours and allowed to recover; scRNA-seq is performed prior to treatment and from two parallel replicates after the population had regrown following the pulse treatment.

The transcribed barcode sequence is measured with other transcripts in scRNA-seq. Clones which significantly increase in abundance after treatment are labeled as resistant and those which decrease are labeled as sensitive. We then map the resistant and sensitive functional phenotypes to the transcriptomes of the individual cells they correspond to at the pre-treatment time point. A machine learning classifier is built based on the labeled cell identities and their transcriptomes, and we can apply this classifier to each of the "unknown" cell identities (phenotypes) from the remaining samples. Estimating the binary phenotype of each individual cell from a sample taken throughout treatment response, we quantify the phenotypic composition ($\phi(t)$) at each time point that scRNA-seq was performed. This phenotypic composition measurement can then be combined with longitudinal population size data from drug treatments at different concentrations, compared to corresponding model outputs, and serve to calibrate the mathematical model of drug-induced resistance (Figure 4.8).

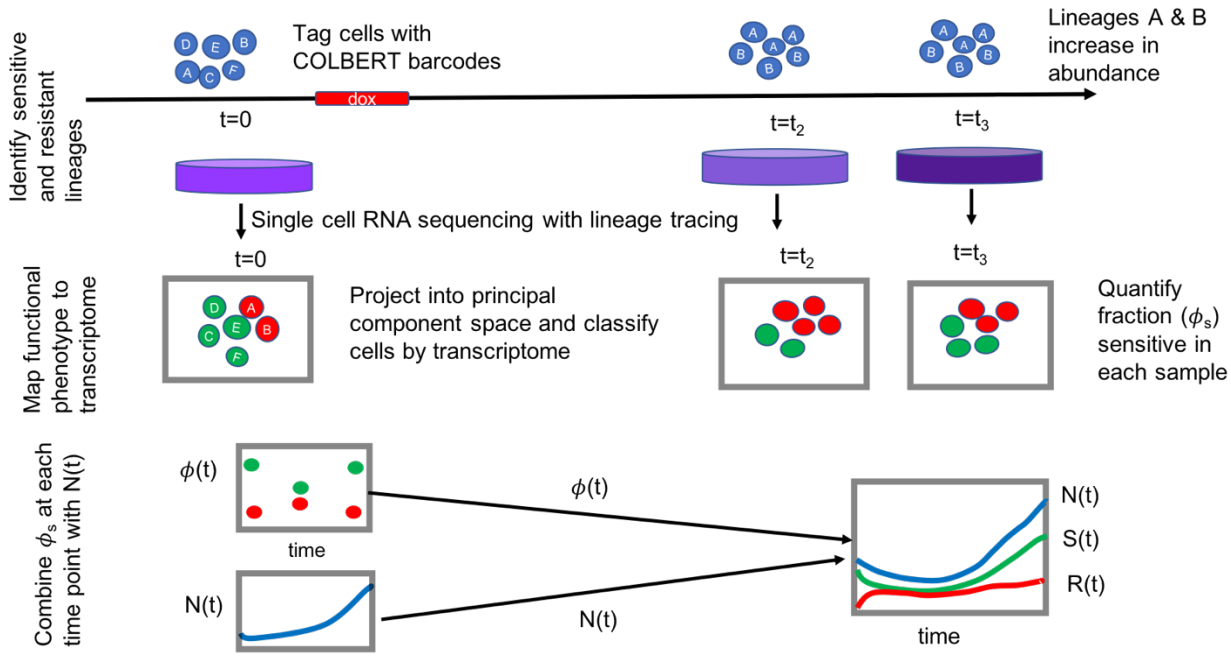

Figure 4.2. Schematic of the workflow for identifying model parameters from data. At $t = 0$ prior to treatment, individual cells are tagged with a unique, heritable, expressed COLBERT barcode. Cells are treated with a pulse treatment of doxorubicin and allowed to recover from treatment, at which time the barcode abundance is quantified. Lineages whose barcode abundance increased from pre- to post-treatment are assumed to have been in a phenotypic state at $t = 0$ that conferred them more resistant to drug than cells whose barcodes significantly decreased in abundance after treatment. Samples of the population were taken before and from parallel replicates sampled at two different time points after treatment for scRNA-seq. The transcriptomes in the pre-treatment samples of the cells tagged with resistant lineages are assigned resistant and the cells tagged with sensitive lineages are assigned sensitive. Using the gene-cell matrix and labeled class identities of sensitive or resistant from the pre-treatment time point only, a classifier is built using Principle Component Analysis (PCA) to distinguish between sensitive and resistant cells. The classifier is applied to the remainder of transcriptomes of the cells, resulting in a prediction for each cell as either sensitive or resistant. These machine learning outputs are made actionable as state variables by using them to quantify the proportion of sensitive cells ($\phi(t)$) at the three time points. This is combined with separate experiments of longitudinal treatment response dynamics ($N(t)$) of the bulk population of the same cell type, and both serve as measured data to be compared to model predicted outputs for parameter estimation.

Lineage-traced scRNA-seq enables identification of sensitive and resistant phenotypes

To investigate the dynamic changes in phenotypic composition in response to treatment, we sought to characterize gene expression over time at the single-cell level.

ScRNA-seq was performed on a barcode-tagged cell population at three time points: immediately before treatment and at parallel replicate samples taken at 7 and 10 weeks after doxorubicin treatment (see Methods: Drug treatment of barcoded cells for scRNA-seq and recovery). By quantifying the proportion of cells with each lineage identity before and aggregated after treatment, we could identify a functional phenotype associated with the pre-treatment transcriptomes. We quantified changes in lineage abundance (percent of the post-treatment population minus percent of pre-treatment population) of each lineage present in the pre-treatment sample to obtain a distribution of changes in abundance after treatment (Fig 4.3A). Cells whose lineage abundance increased by any amount after treatment were labeled resistant, and cells whose lineage abundance decreased by more than 5% were labeled sensitive (Fig 4.3A). All other cells remained unlabeled. This resulted in a training set of 815 labeled cells and their expression levels of 20,645 genes.

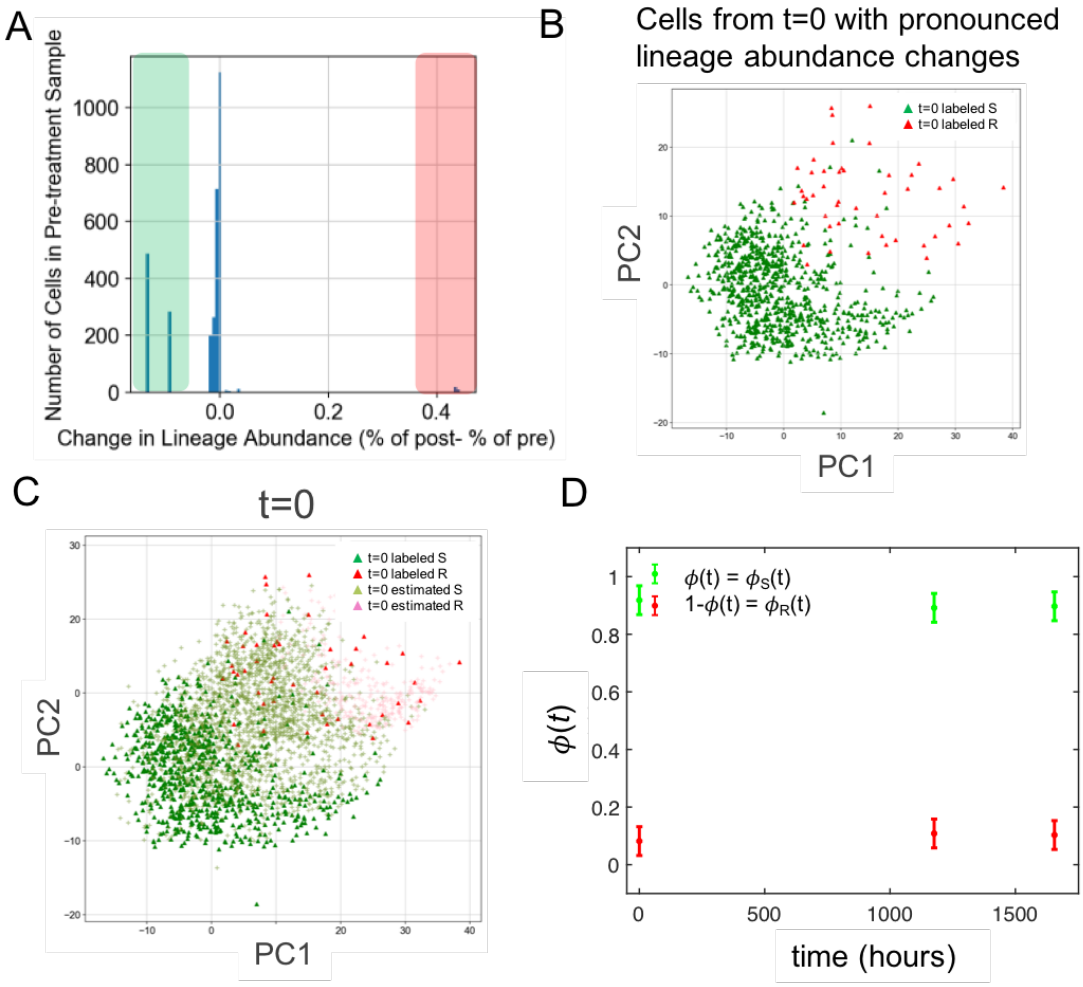

Figure 4.3 Functional Read-out of Changes in Lineage Abundance Allows Mapping of Phenotypes to Transcriptome A. Distribution of changes in lineage abundance from pre- to post-treatment indicates separation of lineages whose cells survive and proliferate and those that are more likely to have been killed by the drug treatment. B. Cells from the extremes of high and low lineage abundance changes (highlighted in green and red in A), projected into principal component space display separation along components (cells are projected into PC1 and PC2 space for visualization, full PC-space is made up of 500 principal components, from the initial 20,645 genes detected). C. Example of remaining cells from $t = 0$ projected onto labeled cells in PC space and estimated as sensitive (olive) or resistant (pink). This was performed for the remaining two time points as well ($t = 7$ weeks and $t = 10$ weeks) D. Proportion of cells classified as sensitive or resistant at each time point is quantified from each samples projection and classification as is displayed in C.

The cells from the identified lineages in the pre-treatment time point were labeled as sensitive or resistant as described above, and the labeled gene-cell matrix was used to build a classifier (based on principal component analysis) capable of predicting whether a cell is more likely to be in a resistant or sensitive state based on its gene expression information alone. See Methods: Machine Learning of Cell Phenotypes for full description of building of the classifier. The optimal hyperparameters of the number of nearest neighbors and the number of principal components for class separation were determined based on 5-fold cross validation and coordinate optimization and were found to be 500 principal components and 73 nearest neighbors (Figure 4.9). A full description of the methods for hyperparameter optimization are outlined in the Methods: Hyperparameter Optimization. Labeled cells are projected into the principle component space, as is displayed visually using projections into only PC1 and PC2 in Fig 4.3B, for sensitive cells (green) and resistant cells (red). To identify the phenotype of new cells that the classifier is not trained on, we project each cell into the principle component space of the labeled cells. A k-nearest neighbor graph is constructed to identify the class of the 73 nearest neighbors in the space, and these are averaged to find the probability of the new cell being in the sensitive or resistant state, where cells above a probability threshold are estimated as sensitive (olive) and below the threshold are estimated as resistant (pink) (Fig 4.3C). This is done for the remaining pre-treatment samples as well as for the cells from the 7 week and 10-week time points (Figure 4.10) based on the single cell gene expression vectors (transcriptomes) of each individual cell. We use the machine learning output to predict each cell as either sensitive or resistant and make these predictions actionable by leveraging

them as measurements of state variables, the proportion of sensitive cells over time $\phi(t)$ (Fig 4.3D). This quantity will be used as one measurement source for model calibration (Eq. 1) (Figure 4.8A & B)

Ability to classify cells as treatment sensitive and resistant enables mechanistic insight into hallmarks of the resistant phenotype

Having identified cells as either resistant or sensitive based on a functional read-out of post-treatment abundance, we can use the class estimates to better understand the transcriptional differences between resistant and sensitive cells. Because the cells largely overlap in principal component space, we use Uniform Manifold Approximation and Projection (UMAPs) as an alternate dimensionality reduction technique for visualization only of the scRNAseq data from the three time points (See Methods: Filtering and normalization). UMAP projections allow for separation of the three time points (Fig 4.4A), and within this projection we can highlight which cells were estimated as sensitive (green) and which cells were estimated as resistant (red) (Fig 4.4E). We can see that the resistant and sensitive cells separate along the first two principal components from Fig 4.4B, where resistant cells accumulate in the upper right quadrant (high in PC1 and PC2) and sensitive cells aggregate in the lower left quadrant (low in PC1 and PC2). Although these principal components only make up a small proportion of the observed variance in gene expression (Figure 4.9D), we see a significant drop off in observed variance after the first few components, indicating that the weights of the genes in these first two components can likely provide us with some mechanistic insight as to which genes are most highly weighted in determining drug-resistance classification. We select a subset of the gene loadings and

plot their direction in the first two principal components (Fig 4.4C), along with a heatmap of the average expression levels for the sensitive and resistant cells for each time point for the top 50 weighted genes in PC1 (Fig 4.4D). The heatmap reveals that the patterns of regulation between genes in the sensitive and resistant cell classes are conserved across the time points. Examples of the differential expression of NEAT1, UBE2S, and TOP2A are shown in Fig 4.4 F, G, and H respectively. Comparing these gene expression maps to the UMAP of resistant and sensitive cell classes (Fig 4.4E) we can see that increased expression in NEAT1 is associated with resistance, while increased expression in UBE2S and TOP2A are associated with the sensitive state. Mechanistically, this corroborates previous findings, as NEAT1 has been shown to be associated with resistance in triple negative breast cancer (Shin *et al.*, 2019). Although we do not perform detailed molecular analysis in this work, the framework presented here to distinguish sensitive and resistant cells over time can be used to perform a more detailed mechanistic investigation of molecular drivers of resistance, and that is an area of future work. For now, we present the results to demonstrate the interpretability of the classifier and its ability to be validated by examining the gene expression levels of known markers of resistance in the components of the classifier.

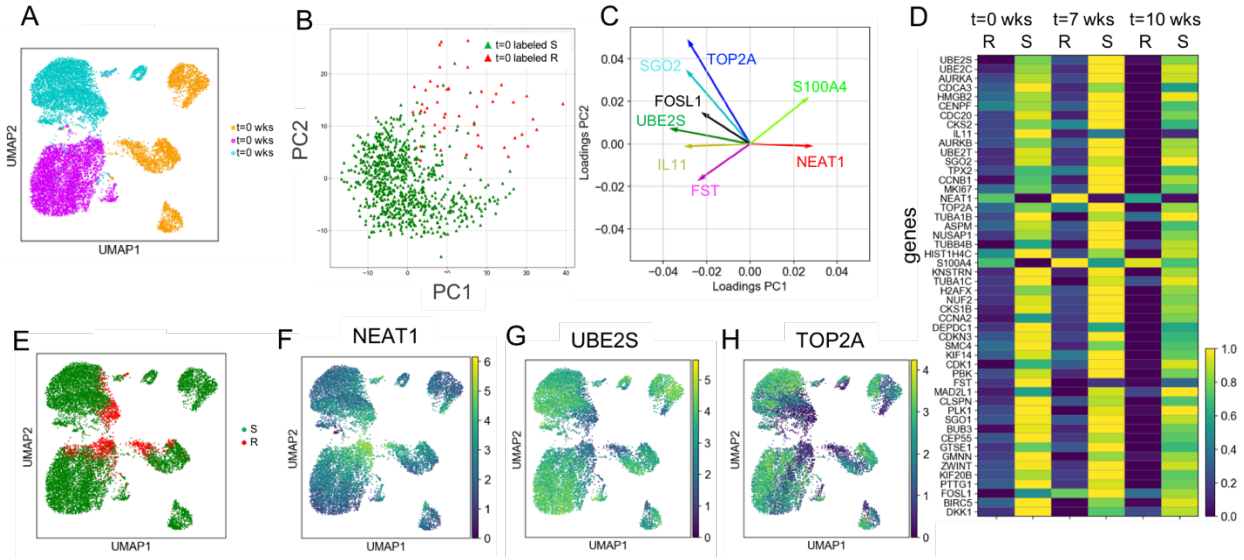

Figure 4.4. Principal Component Analysis and Differential Gene Expression Analysis Provide Molecular Insight into Drug Resistance Interactions A. UMAP projection of single cell transcriptomes colored by time point. B. Labeled sensitive and resistant cells projected into the space visualized by the first two principal components, PC1 and PC2, indicating that resistant cells cluster in the upper right quadrant (high in PC1 and PC2), and sensitive cells tend to cluster in the bottom left quadrant (low in PC1 and PC2). C. Gene loadings for selected genes plotted in the space of the first two principal components illuminates key genes that may be associated with resistance to doxorubicin. D. Heat map of the top 50 genes in PC1 comparing the average expression across the sensitive and resistant cell groups in the three time points, showing a characteristic pattern between sensitive and resistant cells across the three time points. The color bar is scaled within each gene (row). E. Single cells colored by sensitive (*S*) and resistant (*R*) cell classifier labels visualized via UMAP projections indicates drug sensitivity phenotypes cluster together, but not exclusively by the apparent UMAP clustering. F. UMAP projections of cells colored by expression level of NEAT1 indicates high expression of NEAT1 is associated with resistance. G. UMAP projections of cells colored by expression level of UBE2S indicates that high expression of NEAT1 is associated with sensitivity. H. UMAP projections of cells colored by expression level of TOP2A indicates that high expression of TOP2A is associated with sensitivity.

Experimental measurements of population size dynamics in response to varying pulse treatments of doxorubicin

In any attempt to model changes in subpopulation frequencies in cancer, bulk population dynamics reflecting differential growth rates and drug sensitivities need to be taken into account. In pivotal work by (Stumpf *et al.*, 2017) in the field of differentiation, the proportion of cells in three well-defined differentiation states was used to calibrate mathematical models to describe the mechanism of directed transitions in the differentiation process. Over this time scale, it could reasonably be assumed that no significant differential growth rates accounted for changes in composition. However, in cancer and specifically in this study, we monitor populations over much longer time scales and it is necessary to also consider the contribution of differential growth and death rates among subpopulations. This requires measurements of both bulk population dynamics and subpopulation frequencies over time. In the experimental *in vitro* setting, quantifying bulk population size dynamics is quite feasible for a range of treatment conditions, allowing us to observe the differences in response due to various drug concentrations. Cells were treated with a 24-hour pulse of one of 10 doxorubicin concentrations (ranging from 0-1000 nM, n= 6 replicate cell populations for each dose) (Fig 4.5A, 5B) and the cell number monitored throughout regrowth by time-lapse microscopy. The mean and 95% confidence intervals of cell number in time are shown in Fig 4.5C. For each dose, we measured the critical time, t_{crit}, on this measurement of treatment efficacy (Fig 4.5D). As expected, the higher the dose, the longer the population remained below the critical cell number. The cell number in time data, $N(t)$, will be used as another measurement source for model calibration (Eq.1) (Figure 4.8C&D).

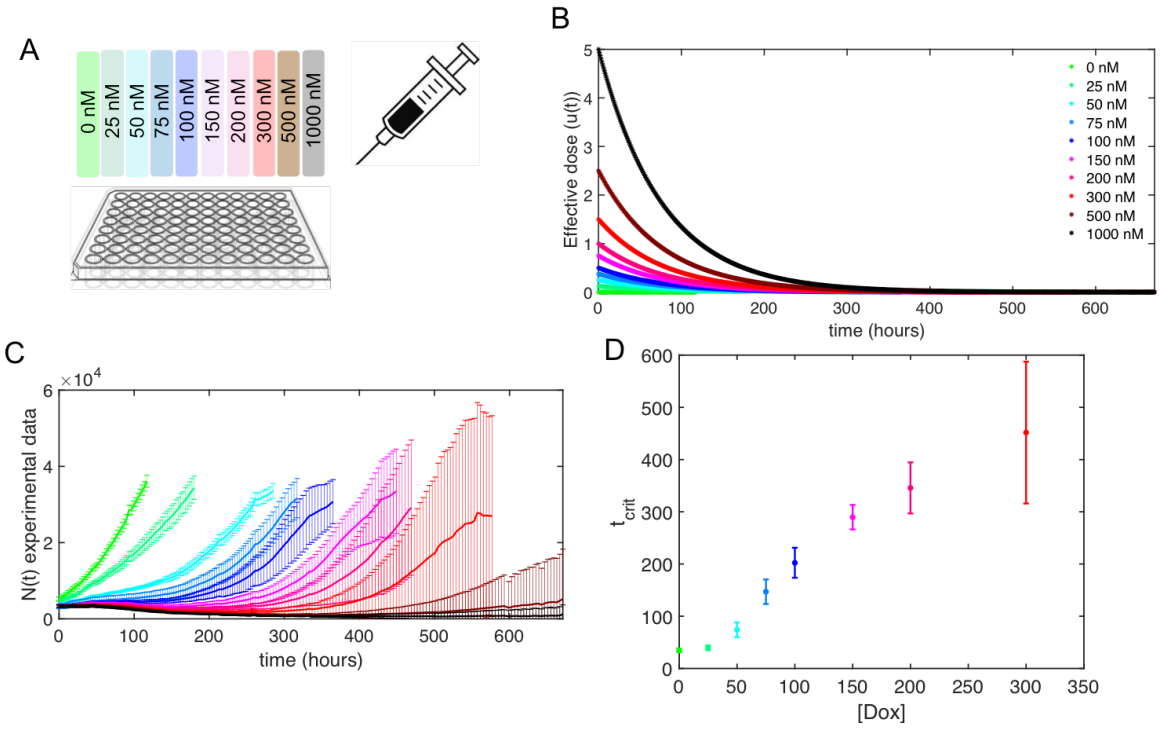

Figure 4.5. Longitudinal Treatment Response Dynamics for Pulse Treatments at Ten Different Drug Concentrations. A. Schematic of experimental set-up using time-resolved fluorescence microscopy to measure the number of MDA-MB-231 GFP labeled breast cancer cells in response to doxorubicin concentrations ranging from 0-1000 nM treated for 24 hours and then allowed to recover in growth media. B. Estimated effective dose dynamics ($u(t)$) of the various pulse-treatments of doxorubicin. C. Number of cells in time, colored by drug concentration as in B, from six replicate wells. Error bars represent 95% confidence intervals around the mean cell number at each time point. Images were converted to cell number estimates every 4 hours. Time of monitoring ranged from 1 week (168 hours) for the untreated control to 4 weeks (672 hours) for the 1000 nM dose. D. Experimental measurements of the t_{crit} for each doxorubicin treatment, legend as in B.

Integrating estimates of phenotypic composition with longitudinal treatment response data is necessary for identifiable model calibration

To utilize all possible pieces of information available about the treatment response of this experimental system, we sought to develop an integrated model calibration scheme

that is capable of integrating information from multimodal data sources. Here, we apply the concept of pareto optimality (Censor, 1977) to reflect the trade-off between goodness-of-fit in each of the two data sources: 1) from longitudinal population data, *N(t)*, sampled at a high temporal resolution and for a number of doses, and 2) machine learning outputs that estimate the phenotypic composition $\phi(t)$ at three distinct time points before and after treatment. For the following dual-objective function, we use a regularization term λ, which can vary from 0 to 1 to reflect our varying confidence in the data from each of the measurement sources. Here we use weighted, non-linear, least squares as the simplest possible calibration method

$$J(\theta) = \frac{\lambda}{n_{\phi(t)}} \sum_{j=1}^{n_{\phi(t)}} \frac{(\widehat{\phi_j} - \phi_i(\theta, u_j))^2}{\sigma_{\phi_j}^{\ 2}} + \frac{(1-\lambda)}{n_{N(t)}} \sum_{k=1}^{n_{doses}} \sum_{i=1}^{n_{N(t)k}} \frac{(\widehat{N_{i,k}} - N_i(\theta, u_{i,k}))^2}{\sigma_{N_{i,k}}^{\ 2}}, \text{ (Eq. 3)}$$

where $n_{\phi(t)}$ is the number of $\phi(t)$ time points, ϕ_j is the experimentally estimated ϕ at time point j, $\phi(\theta,u_j)$ is the model predicted ϕ for a given effective dose u at time j, $\sigma^2_{\phi j}$ is the variance in the measurement of ϕ at time j, $n_{N(t)}$ is the number of total $N(t)$ time points, n_{doses} is the number of different doses applied, $n_{N(t)k}$ is the number of time points in the kth dose, $\underline{N_{i,k}}$ is the measured number of cells at the ith time point for the kth dose, $N(\theta,u)$ is the model predicted number of cells at time i for the kth effective dose, and σ^2_N is the variance in the measurement of N at time i for the kth dose. The resulting objective function $J(\theta)$, minimizes the sum of the squared error in the $\phi(t)$ and $N(t)$ data compared to the model predicted $\phi(t)$ and $N(t)$. The errors are weighted by the experimentally observed uncertainty in those estimates and normalized by the number of $\phi(t)$ and $N(t)$ data points. The results

of this parameter estimation, in terms of weighted error in $N(t)$ and $\phi(t)$ for varying degrees of confidence in each output are shown as the Pareto front set of solutions in Fig 4.6A. See Methods: Model fitting with multiple measurement sources for a description of how this front was found (Figure 4.11). The front is centered around λ^*=0.5, the regularization term value that equally weights the measurement sources based on the number of data points available from each measurement source. The best fitting parameter set resulting from using the objective function with a $\lambda=\lambda^*$ is denoted as θ^* (red dot in Fig 4.6A) and will be used to evaluate goodness of fit and prediction accuracy going forward.

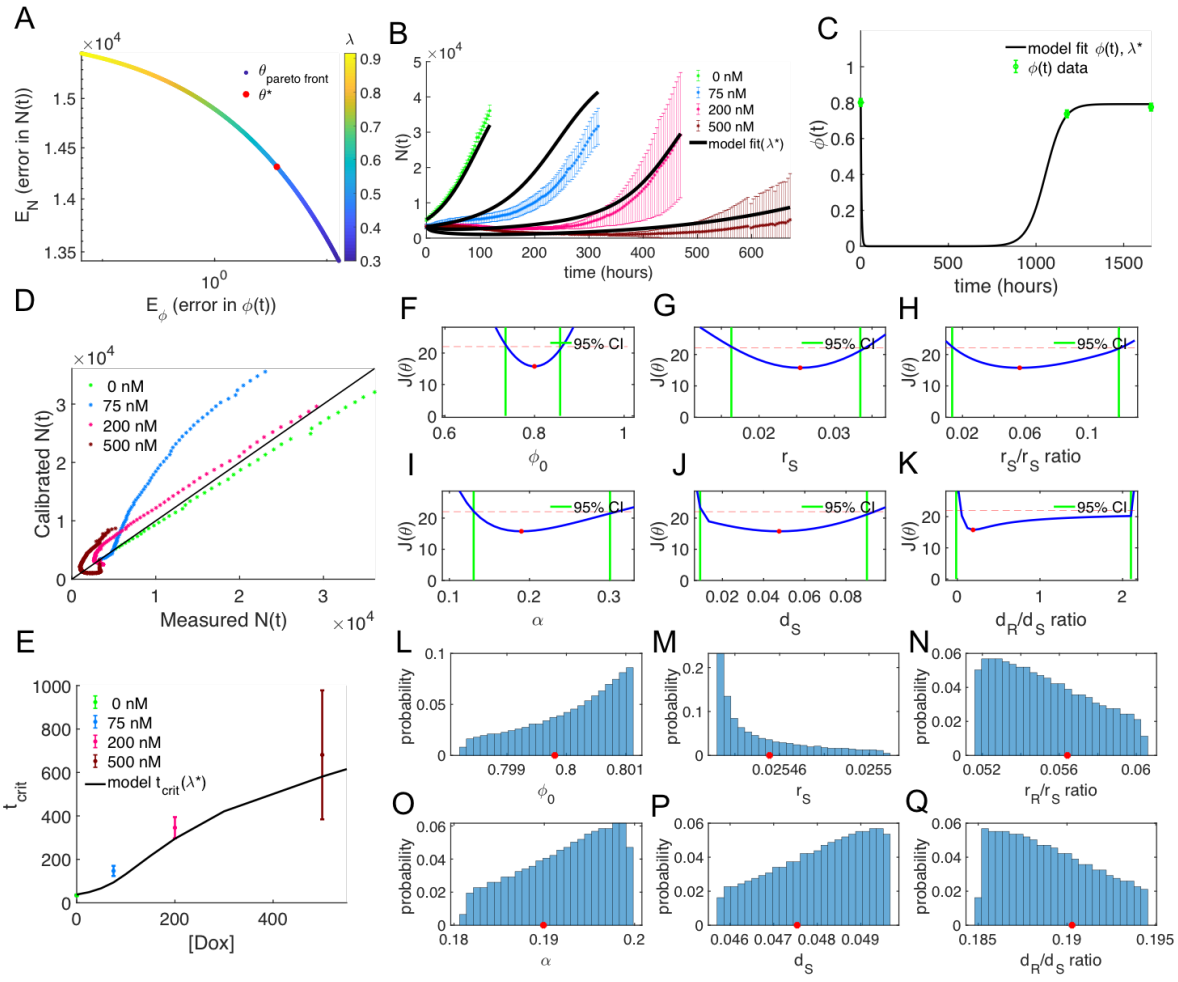

Figure 4.6. continued on next page, Integrated model calibration to incorporate both measurement sources reveals identifiability of model parameters. A. The parameter sets, θ, that fall on the "Pareto front", reflecting a tradeoff between goodness of fit in $N(t)$ and $\phi(t)$. Each dot represents a parameter set, θ, acquired by varying the regularization term, $\boldsymbol{\lambda}$, from 0 to 1 and then filtering solutions (Figure 4S.4) to within a reasonable accuracy in $N(t)$ and $\phi(t)$. The red dot parameter set represents when $\boldsymbol{\lambda}= \boldsymbol{\lambda}^*= 0.5$ in which the weighting between the measurements sources is given equal weight and normalized based on the number of data points in $N(t)$ and $\phi(t)$ measurements. The Pareto front solutions are found by performing multiple optimizations at different values of $\boldsymbol{\lambda}$. Lower values of $\boldsymbol{\lambda}$ (blue) indicate improved fit in $N(t)$ data, whereas higher values of $\boldsymbol{\lambda}$ (yellow) indicate improved fit of the $\phi(t)$ data. B. Calibration results for longitudinal $N(t)$ data from the four doses (0, 75, 200, and 500 nM) used for calibration for the parameter set θ^* (represented by the red dot in A) C. Calibration results for phenotypic composition ($\phi(t)$) data from the same parameter set θ^* , yielding an accuracy in ϕ(t), measured by the concordance correlation coefficient (CCC) of 0.93 D. Measured cell number $N(t)$ verses

model calibrated cell number, yielding a concordance in $N(t)$ of CCC = 0.93. E. Critical time from $N(t)$ data compared to model calibrated critical time for selected parameter set (θ^*) in red in (A) (CCC= 0.88). F. Profile likelihood curvature around the initial sensitive cell fraction (ϕ_0) to determine 95% CI on parameter of ϕ_0=0.80 [0.74, 0.86]. G. Profile likelihood curvature around sensitive cell growth rate (r_S) reveals 95% CI of r_S = 0.026 [0.016, 0.033]. H. Profile likelihood curvature around the ratio of the resistant growth rate to the sensitive cell growth rate reveals a CI of r_R/r_S ratio= 0.056 [0.013, 0.12]. I. Profile likelihood around the drug-induced resistance rate, α of α = 0.19 [0.13, 0.30]. J. Profile likelihood around the death rate due to drug of the sensitive cell death rate with CI d_S = 0.048 [0.0092, 0.90]. K. Profile likelihood curvature of the ratio of the death rate due to drug of the resistant versus sensitive cell fraction with CI d_R/d_R ratio = 0.19 [-0.014, 2.1]. L. Distribution of Pareto front accepted parameter ϕ_0. M. Distribution of Pareto front accepted parameter r_S. N. Distribution of Pareto front accepted parameter resistant to sensitive growth rate. O. Distribution of Pareto front accepted parameter α. P. Distribution of Pareto front accepted parameter d_S. Q. Distribution of Pareto front accepted parameter resistant to sensitive cell death rate.

We use this parameter set, θ^*, (red dot in Fig 4.6A, red dots in Fig 4.6L-Q) to demonstrate an example of the $N(t)$ data fit to the model (Fig 4.6B) and the $\phi(t)$ data fit to the model (Fig 4.6C) with a CCC in $\phi(t) = 0.93$. We demonstrate that the model calibration is fairly accurate at calibrating the $N(t)$ data (Fig 4.6D) with a CCC = 0.93, and is able to capture broader changes over the range of doses by properly matching the critical time (t_{crit}) as a function of dose (Fig 4.6E) for the four doses the model is calibrated on (CCC = 0.9657). In the model development process, we tested to make sure that each of the parameters was sensitive to the relevant model outputs, in this case the critical time (t_{crit}) and the phenotypic composition at critical time $\phi(t=t_{crit})$, for a range of doxorubicin doses. Results from the global sensitivity analysis (See Methods: Sensitivity analysis of model parameters) revealed that all parameters are globally sensitive (i.e. contribute to least 5% of the overall value) in at least one of the model outputs for at least one of the drug doses (Figure 4S.5), except for the carrying capacities (K_N *and* K_ϕ) of the two experimental systems. We used this analysis to inform our decision to set the carrying capacities from separate experiments (Figure 4.13) and literature (Thermo Fisher Scientific, no date) and to try to fit all six remaining unknown parameters.

A key goal of this work is not only to fit the model to multiple data sources, but to demonstrate that the use of the information gained from these dispersed data types is critical in enabling the practical identifiability of the six free model parameters in the mechanistic model. The ability to ensure that model parameters are identifiable from data enables us to have confidence in our interpretation of the values of the model parameters

to be used for making predictions and ultimately decisions, and thus is essential for eventually translating modeling frameworks like the one presented here to real-world settings. A critical first step is to demonstrate the structural identifiability of the system, which was shown (See Methods: Structural identifiability of model parameters) under the assumption of perfect data. Next, in order to test whether the calibrated parameter set θ^* is practically uniquely identifiable from the available data and the objective function (Eq. 3), we utilize the profile likelihood method (Eisenberg, Robertson and Tien, 2013; Brouwer *et al.*, 2017; Eisenberg, 2019). We profiled each parameter independently at a range of values around its best fitting value, θ^*_i, fitting for all the other parameters, and returning the resulting best possible objective function value ($J(\theta)$) for the new optimization problem (Fig 4.6F-K). Parameters that are easily identifiable will result in a tight curvature around the best fitting value, meaning that changing for example the value of the initial sensitive cell fraction (ϕ_0) leads to a large change in the best possible minimized error. Parameters are considered practically identifiable if the curvature of the objective function value crosses above the threshold of the 95% χ-squared distribution (Raue *et al.*, 2009) (red dashed line Fig 4.6F-K). The parameter value at which this threshold is crossed is considered the upper and lower bound of the 95% confidence interval in the parameter value (green vertical lines Fig 4.6F-K), providing an estimate of uncertainty in the best fitting parameter value θ^*_i. The results of this analysis reveal that the six free model parameters are uniquely practically identifiable from the available data. The parameter relationships that result from profiling each individual parameter can be seen in Figure

4.14. In contrast, when the test of practical identifiability was repeated for the case in which the calibration was performed on the $N(t)$ data alone (Figure 4.15), and the results revealed a number of non-identifiable parameters in practice (Figure 4.16). This analysis demonstrated that the incorporation of the snapshot phenotypic composition data was not only a useful additional piece of information, but essential to making the model calibration and parameter estimates identifiable and ultimately useful.

We further investigated the functionality of our Pareto front set of parameter values by examining the resulting distribution of parameter values that are accepted into the front (Fig 4.6L-Q). The distributions of parameter values tell us about which parameters, such as the sensitive cell growth rate, r_S (Fig 4.6L), tend to be very stable regardless of the weighting, whereas other parameters, such as the degree of drug-induced resistance α (Fig 4.6O), are more variable. We observed that the individual parameter values tended to vary systematically with the goodness of fit in $N(t)$ vs. $\phi(t)$ (Figure 4.17), however, all of the parameters in the distributions shown in Fig 4.6L-Q fall within the 95% CI around θ^* (Figure 4.18). We plot a few examples of the Pareto front solution sets fit to the $N(t)$ and $\phi(t)$ data in Figure 4.19, which demonstrates the relatively subtle differences between the fits depending on the weighting of each measurement source.

Model validation using functional isolation of "sensitive" and "resistant" cells predicted from classifier

Because we rely on the machine learning classifier of cell phenotypes from transcriptomic data, we sought to validate our classifier model experimentally to ensure

that cells labeled as “resistant” and “sensitive” were exhibiting these expected phenotypes. Our mathematical model assumes that sensitive cells proliferate more rapidly than resistant cells (i.e. exhibit a higher growth rate) and that resistant cells are capable of improved survival in response to doxorubicin treatment. To test these attributes functionally, we used the COLBERT barcoding system (Al’Khafaji, Deatherage and Brock, 2018) to identify individual lineages from the pre-treatment sample who were labeled as sensitive and resistant based on their changes in lineage abundance, and subsequently isolated them experimentally from the replicate pre-treatment population using the COLBERT recall system (Al’Khafaji, Deatherage and Brock, 2018) (Fig 7A). Once isolated, cells were sorted into single cell clones for functional analysis of growth dynamics and drug sensitivity. Our results confirmed that the cells from the isolated sensitive lineage grow more quickly than the isolated resistant lineage (Fig 4.7B), with overall growth rates of g_S=0.011 and g_R= 0.005 per hour respectively (Figure 4.20). Drug sensitivity was assessed by dosing cells at 400 nM and 2.5 μM for 48 hours and immediately quantifying cell viability via a live-dead assay. The resistant lineage had higher percent viability at both doxorubicin concentrations, with a statistically significant difference in viability at the higher dose (Fig 4.7C).

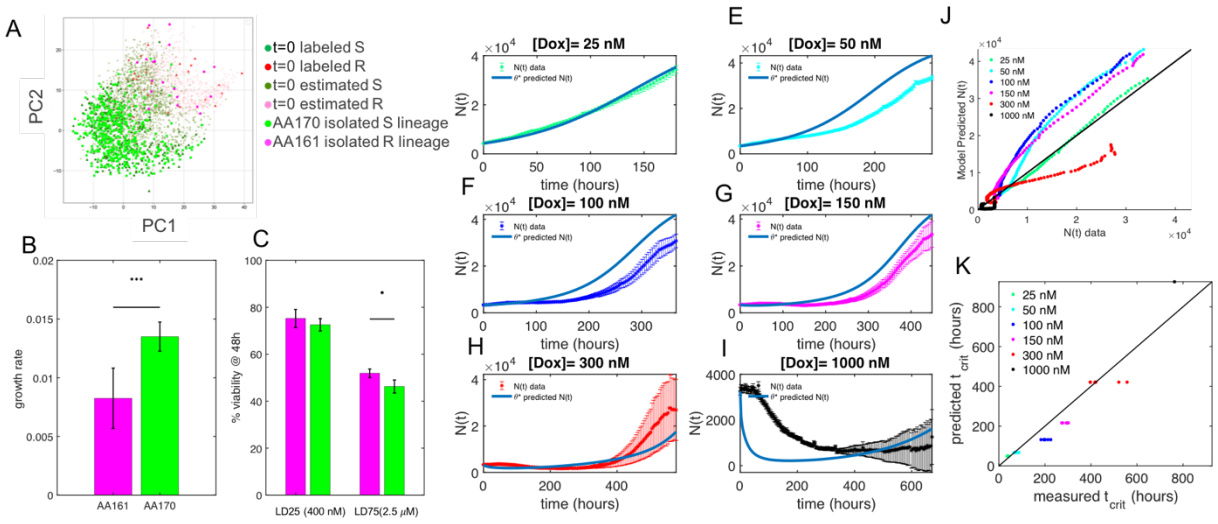

Figure 4.7 Combined Model Validation *via* Lineage Isolation and Prediction of Treatment Response. A. Projection of classified sensitive and resistant cells at the pre-treatment time point into principal component space, with cells from an isolated sensitive lineage (AA170) in bright green, and an isolated resistant lineage (AA161) in hot pink B. Growth rates of the 12 replicate wells of each isolated lineage reveal that the resistant lineage grows significantly more slowly than the sensitive lineage (p = 2.7e-6), as is predicted from the model parameters where $r_S > r_R$. C. Functional testing of the drug sensitivity of each lineage indicates that the cells from resistant lineage (AA161, pink) have a higher resistance, measured by cell viability at 48 hours, at both 400 nM and 2.5 μM doses of doxorubicin, with p-values of p = 0.1942 and p = 0.0023, respectively. D. Prediction of treatment response at 25 nM, E. 50 nM, F. 100 nM, G. 150 nM, H. 300 nM, and I. 1000 nM from θ^* (red dot in Fig 6A). The mean measured cell number in time and 95% confidence interval from six replicate wells are shown for each treatment response. J. Scatterplot of model predicted $N(t)$ versus experimental $N(t)$ data for all 6 new treatment conditions with an overall CCC = 0.89. K. Scatterplot of model predicted critical time from selected parameter set versus experimentally measured critical time, indicating that although we might not be able to precisely predict the exact trajectories of cell number in time for each dose perfectly, we can globally capture the critical time (t_{crit}) for a range of doxorubicin concentrations, despite our model not being trained on these concentrations, with an overall CCC between model predicted critical time and observed critical times of CCC = 0.92.

Multimodal data sources can be leveraged to predict response dynamics to new drug concentrations

A key advantage of leveraging multimodal data sources for parameter estimation is that we can uniquely identify the model parameters and use them to make predictions about the response dynamics to new treatment regimens. We validate the model predictions, obtained from running the model forward with parameter set θ* with input effective doses described in Fig 4.5B for the six remaining pulse treatment of doxorubicin that were not used to train the model. The model predictions compared to the experimental measurements are shown for doses of 25 nM (Fig 4.7D), 50 nM (Fig 4.7E), 100 nM (Fig 4.7F), 150 nM (Fig 4.7G), 300 nM (Fig 4.7H) and 1000 nM (Fig 4.7I). We evaluated the accuracy in all the model predictions over all six unobserved doses and see that we are able to predict the treatment response with reasonable accuracy (Fig 4.7J) with an overall CCC of 0.89 for each model predicted and measured cell number ($N(t)$) in time. When we repeat this calibration, removing that phenotypic composition data (by setting λ=0) we get an overall predictive accuracy of CCC=0.79, indicating the improvement in predictive capabilities with insight of the phenotypic dynamics. While the individual trajectories may not precisely match the data at the 4-hour intervals measured here, they are able to predict the global behavior, *via* predicting the critical time as a function of doxorubicin, very well (Fig 4.7K) with a CCC of 0.92. These results demonstrate the flexibility and predictive capability of this modeling framework, demonstrating its utility in predicting the critical behavior needed to guide optimal-treatment decision making.

DISCUSSION

Recent technological advances have enabled unprecedented, high-throughput single-cell molecular level insight of intratumor heterogeneity (Levitin, Yuan and Sims, 2018; Suvà and Tirosh, 2019). The ability to precisely quantify intratumor heterogeneity (Ferrall-Fairbanks *et al.*, 2019), and illuminate key subpopulations involved in response to treatment (Al'Khafaji *et al.*, 2019), has the potential to improve both prognostic and therapeutics for cancer treatment. These genomic and transcriptomic data sets can direct the choice of specific cancer drugs and illuminate novel resistance pathways, as well as provide a prognostic marker for patients who receive it. Simultaneously, the role of mathematical modeling in oncology has been widely recognized (Rockne *et al.*, 2019) and utilized to improve both our understanding of the dynamic mechanisms of drug response (Jarrett, Lima, *et al.*, 2018; Matthew T. McKenna, Weis, Brock, *et al.*, 2018; Smalley *et al.*, 2019) as well as to develop approaches to guide the design of adaptive patient-specific treatment plans (Gatenby *et al.*, 2009; Prokopiou *et al.*, 2015; Poleszczuk and Enderling, 2018; Brady *et al.*, 2019; Zhang *et al.*, 2019). However, connecting the wealth of "omics" data at the molecular level with temporal dynamics used to calibrate mathematical models for adaptive therapies remains a major challenge in the field.

Recognizing the critical roles of heterogeneity in cancer dynamics, mathematical models of tumor progression often include distinct subpopulations, such as cancer stem cells (Badri *et al.*, 2016; Poleszczuk *et al.*, 2016; Brady *et al.*, 2019), or drug resistant and sensitive subpopulations (Greene *et al.*, 2015; Howard *et al.*, 2018; Gevertz, Greene and Sontag, 2019; Greene, Gevertz and Sontag, 2019). However, despite these models being

calibrated to observed experimental or clinical data, the underlying phenotypic composition that these model calibrations suggest cannot easily be validated, since the degree of resistance or stemness of a cancer cell population in time is not easily measured longitudinally *via* a single biomarker. The majority of these modeling endeavors utilize a single measurement source for longitudinal data acquisition and subsequent model calibration. A few studies utilizing multimodal imaging modalities have harnessed the ability to quantify different aspects of tumor composition—such as vasculature, necrosis, and cellularity, to develop an integrated model calibration of multiple tumor system components (Jarrett, Bloom, *et al.*, 2018; Hormuth *et al.*, 2019). However, this integrated, multimodal approach has not explicitly included inference of the composition of heterogeneous subpopulations taken from separate "omics" datasets for direct model calibration.

Here, we introduce an experimental-computational framework for utilizing multimodal data sets when parametrizing a mechanistic model of drug resistance dynamics in response to treatment in cancer. We demonstrate the applicability of this framework when applied to clonally-resolved scRNA-seq data combined with longitudinal treatment response data from a cancer cell line and assess the ability of the model to predict treatment response dynamics. To this end, we developed a machine learning classifier built upon clonal abundance quantification which estimates the class identity of an individual cell based on its transcriptome. The machine learning outputs classified cell states and were used to assign values to the state variables in the mechanistic model: the number of cells in the sensitive or resistant phenotypic state at each time point. We combined these

estimates of phenotypic composition with population-level treatment response data to calibrate a mechanistic model of drug-resistance dynamics. We validated our machine learning classifier by isolating cells from lineages labeled as sensitive or resistant and testing them functionally. We showed that the presence of multiple measurement sources of data allows for the practical identifiability of the model parameters, which are then used to accurately predict the effect of new drug treatments on the cell population.

The power of mathematical models in oncology, especially those calibrated to real data, is that we can both use them to learn about the underlying mechanisms of the system behavior, and we can harness that knowledge to inform future decision making in an experimental or clinical setting (Yankeelov *et al.*, 2013, 2015). Greene *et al*. (Greene, Gevertz and Sontag, 2019) demonstrate that knowledge of the parameters of the model presented here (Eq.1) can be used to drive optimal treatment protocol decisions; in particular they can help determine (for example) whether pulsed or constant treatment is preferred for a specific patient. The applicability of optimal control theory as it applies to cancer treatments relies on the ability to identify model parameters from data. While it has been shown that the model parameters presented in this paper can be identified from just the bulk population dynamics in theory (Greene, Sanchez-Tapia and Sontag, 2018a), in practice the number of experiments needed to test the conditions is quite difficult if the output is the bulk population dynamics alone. However, the identifiability problem becomes significantly easier if the knowledge of the underlying phenotypic composition is also plausible (See Methods: Identifiability of model parameters). In this work, we leverage high-throughput "omics" data sets, taken at just a few snapshots of time, to estimate the

phenotypic composition and demonstrate the improvement in identifiability of model parameters from including this data alongside longitudinal data.

High-throughput single cell transcriptomics or other types of high throughput snapshot data can give an abundance of information about the heterogeneity and potential mechanisms of resistance of cell populations (Al'Khafaji *et al.*, 2019; Ma *et al.*, 2019). However, the ability to use this information beyond hypothesis generation (Smalley *et al.*, 2019), but to actually inform model calibrations, is still lacking. In this work, we attempted to overcome the problem of practical identifiability of model parameters from observed data by demonstrating how to explicitly integrate snapshot data about the relevant cell subpopulations into a model calibration. We argue that the ability to integrate information from snapshot data with temporal data is essential for the potential for the proposed mathematical oncology models to be practically useful, as these models should not "throw away" information but should instead be able to take into account explicitly as much available data as possible.

The functional characterization of single cells via changes in lineage abundance post-treatment enabled us to identify novel mechanistic insights into which pathways and interactions are critical for surviving treatment response. While clustering of cells by their transcriptomes can enable identification of novel cell states, these cell states are not necessarily relevant to drug-tolerance. Once can see this quite simply in scRNA-seq pipelines as failure to remove cell cycle genes from the analysis reveals that cells will often cluster by cell cycle state (Luecken and Theis, 2019). While states of the cell cycle may be important for certain applications, they are often regressed out. However, we cannot regress

out other unknown phenotypic subpopulations, and thus these are what can emerge from unsupervised clustering algorithms. While these can provide novel insight about population structure, they may not be what is relevant to driving changes in treatment response behavior. Thus, the ability to read-out lineage identities represents a novel functional component that enables us to zoom in at the right phenotypic state-space relevant to our question- what cells are more drug resistant and which are more drug sensitive, and what is driving these changes? Because we used principal component analysis to build a classifier to separate the sensitive and resistant cells, we can look at the differences in gene expression patterns between the groups of cells we identified and propose potential novel interactions and new biomarkers. For example, our analysis reveals TOP2A, NEAT1, and UBE2S as delineators between sensitive and resistant cells. This knowledge can provide the basis for future work investigating the role of these genes and their related pathways in drug-response.

While scRNA-seq has limitations in the clinical setting due to its high cost, in experimental settings barcode labeling fits rather flexibly into existing scRNA-seq workflows and can add a critical functional component to the phenotypic read-out, as we display in this work. In the clinical setting, other types of approaches to learn more about cancer cell composition are being employed in the era of precision medicine. From radiomics to genomics, it is becoming increasingly common for patients to have access to high-throughput measurements, or at least some insight into their mutational burden at certain time points. This information may be integrated into the clinical or tumor board's decision-making process (He and Ahuja, 2015).

We suggest that the approach here could be modified for the available types of snapshot data in different experimental or clinical settings, where it is expected that only sparse data in terms of tumor composition and longitudinal dynamics will be available. This could potentially be overcome in two ways: 1) simplifying the model structure to reduce the number of model parameters, or 2) setting some parameter values to those obtained from the literature, and only allowing a few (key) parameters to be patient-specific, as is performed in (Brady *et al.*, 2019). We anticipate that this integrated approach can be applied flexibly to incorporate and integrate snapshot data about population composition with longitudinal bulk population dynamics. While transcriptomic and longitudinal data have been used together in a number of studies, this is the first work to our knowledge that allows for explicit parameter estimation using multimodal measurement sources of varying time resolutions and enables flexible implementation depending on the degree of confidence in each data source. The synergy of machine learning with mechanistic modeling integrates multimodal datasets and opens up new approaches to describe, predict, and ultimately optimize treatment response in cancer.

CHAPTER 4 SUPPLEMENTARY FIGURES, TABLES, AND TEXT

Figures

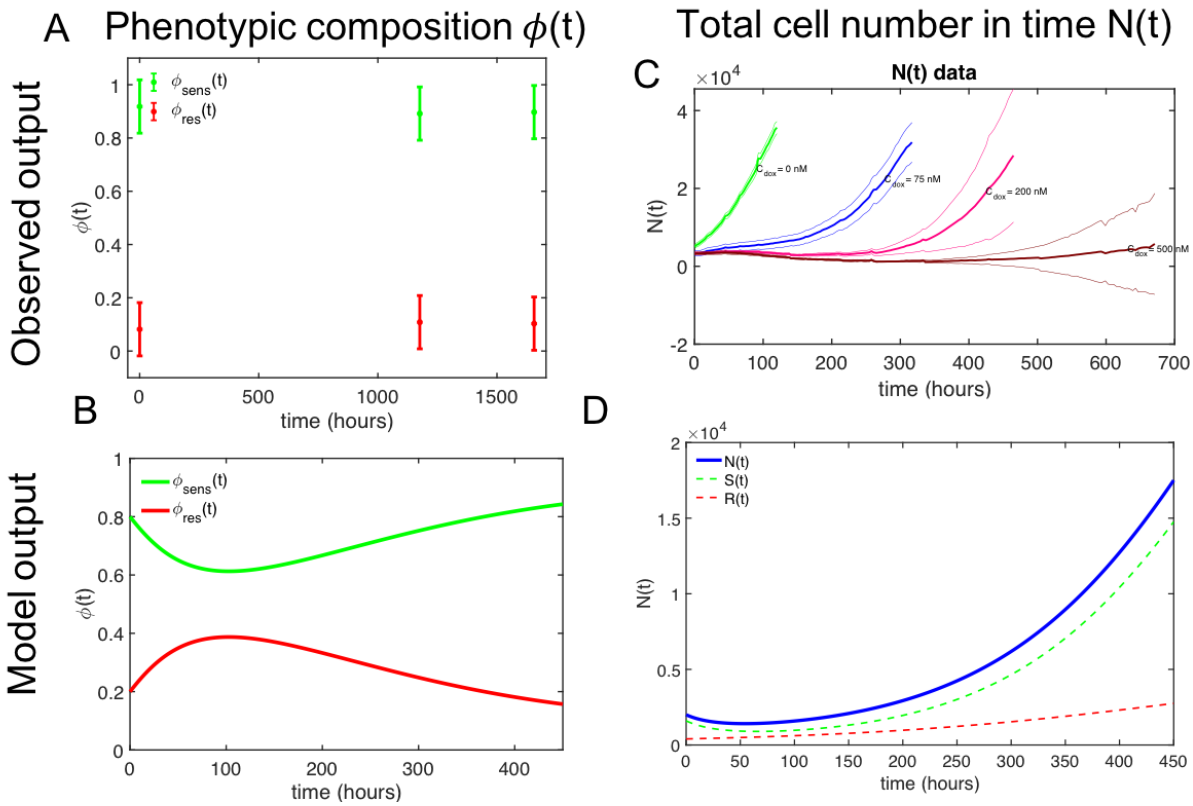

Figure 4.8: Measured and model predicted outputs to be used for parameter estimation from observed data A. Observed estimated fraction of sensitive cells (green) and resistant cells (red) from scRNAseq classifier at three time points ϕ(t). B. Model predicted output of sensitive cell fraction dynamics (green) and resistant cell fraction dynamics (red) for an example parameter set. C. Observed number of tumor cells in time for pulse treatments of doxorubicin at 0, 75, 200, and 500 nM. D. Model predicted output of total cell number in time for a pulse treatment of 75 nM for an example parameter set.

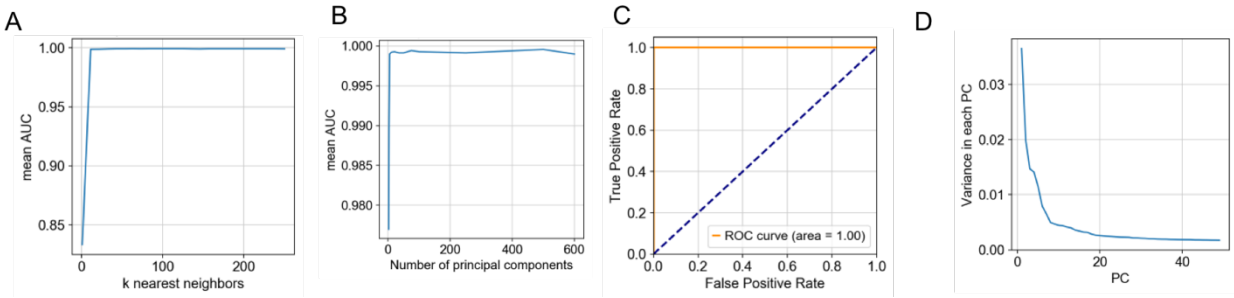

Figure 4.9: Optimization of Principal Component Classifier Hyperparameters use coordinate optimization and 5-fold Cross Validation. A. Number of nearest neighbors used in the classifier versus mean AUC from 5-fold CV to determine optimal number of neighbors of k=73. B. Number of principal components used in the classifier versus mean AUC from 5-fold CV to determine optimal number of components, n=500. C. ROC curve from classifier with optimized number of nearest neighbors and components for separating labeled cells. D. Proportion of variance explained by the principal components drops off sharply for higher PCs.

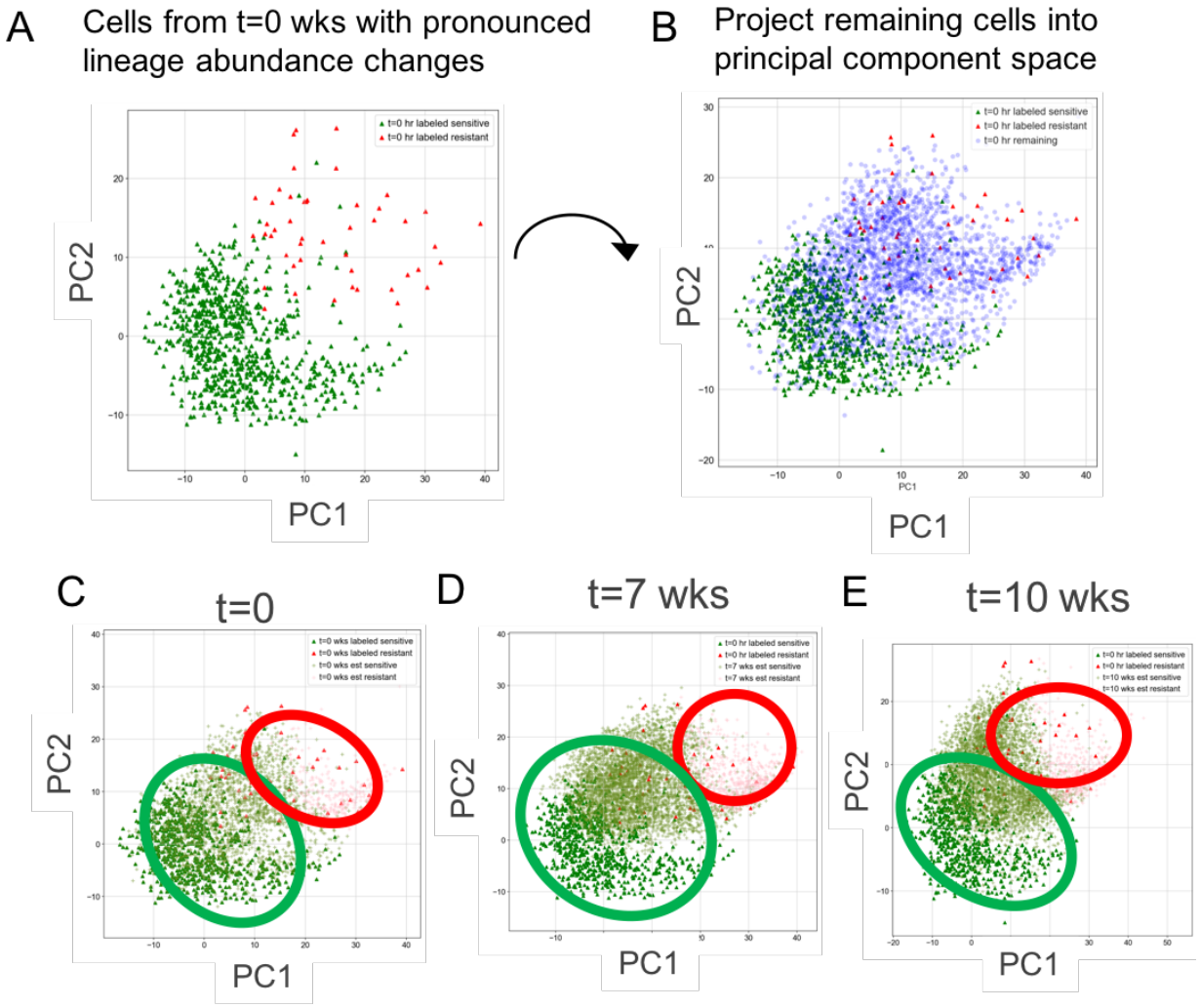

Figure 4.10: Single cell transcriptomes from each time point projected into principal component space and classified using nearest neighbors A. Lineage-abundance guided "labeled" cells projected into principal component space separate along components (PC1 and PC2 shown here for visual effect). B. Unknown cells are projected into the principal component space of the labeled cells. C. Remaining cells from t=0 projected onto labeled cells in PC space and estimated as sensitive (olive) or resistant (green). D. Cells from t=7 weeks projected alongside labeled cells. E. Cells from t=10 weeks projected alongside labeled cells.

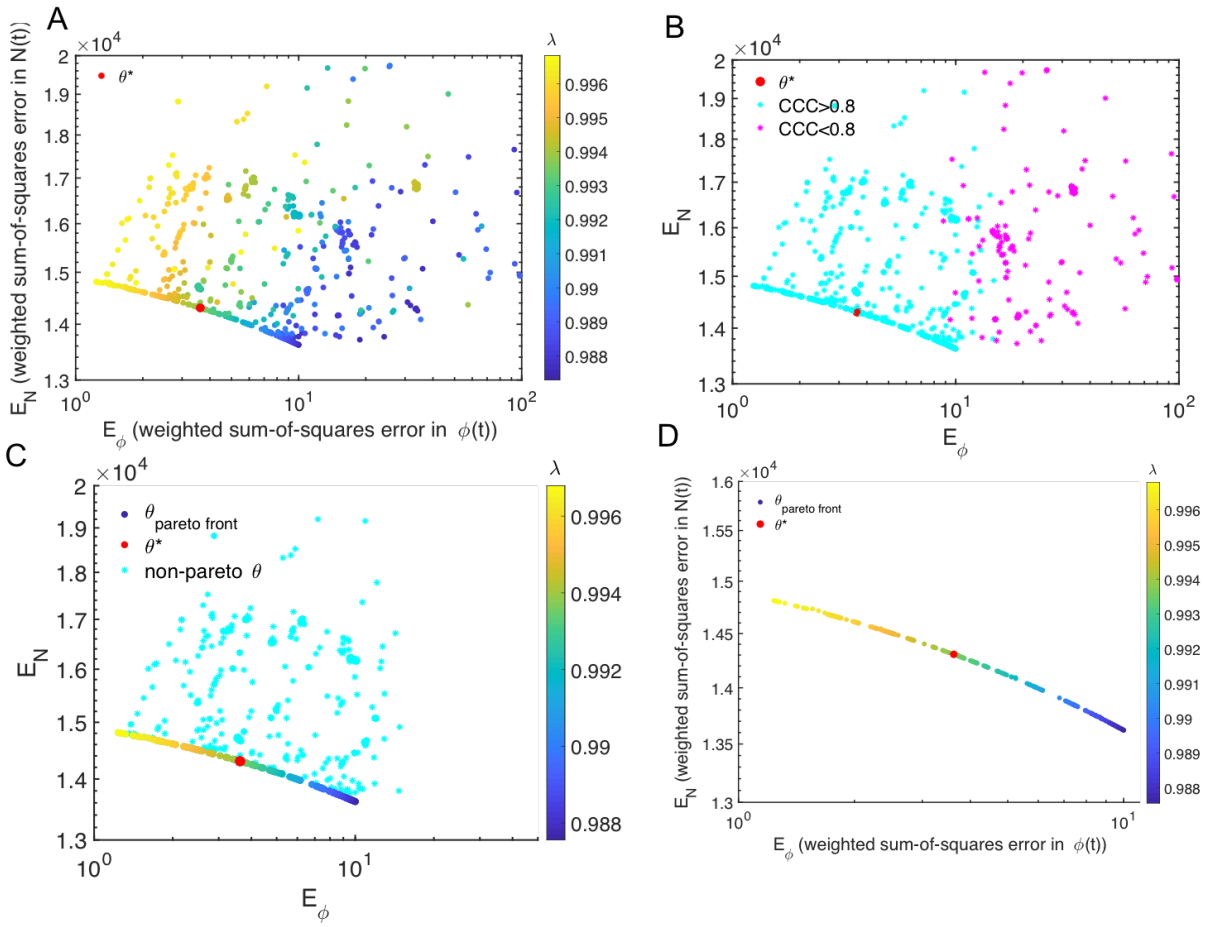

Figure 4.11: Step-by-step refinement of accepted parameter sets to identify solutions along pareto front. A. The resulting weighted sum-squared error in N(t) (E_N) and ϕ(t) ($E\phi$) of 1000 optimizations with the regularization term $\boldsymbol{\lambda}$, varying from $\boldsymbol{\lambda}$=0 (only fitting N(t) data), to $\boldsymbol{\lambda}$=1 (only fitting ϕ(t) data). B. Filtering of parameter sets to require that parameter sets have a CCC>0.8 in both N(t) and ϕ(t). C. Further filtering of parameter sets to remove "non-pareto" solutions- i.e. any parameter sets theta for which there exists another theta with a lower error in both ϕ(t) and N(t). D. The final set of "pareto-front" solutions, which contain parameter sets for which an improvement in error in N(t) comes at a trade-off of a worsening in error in ϕ(t).

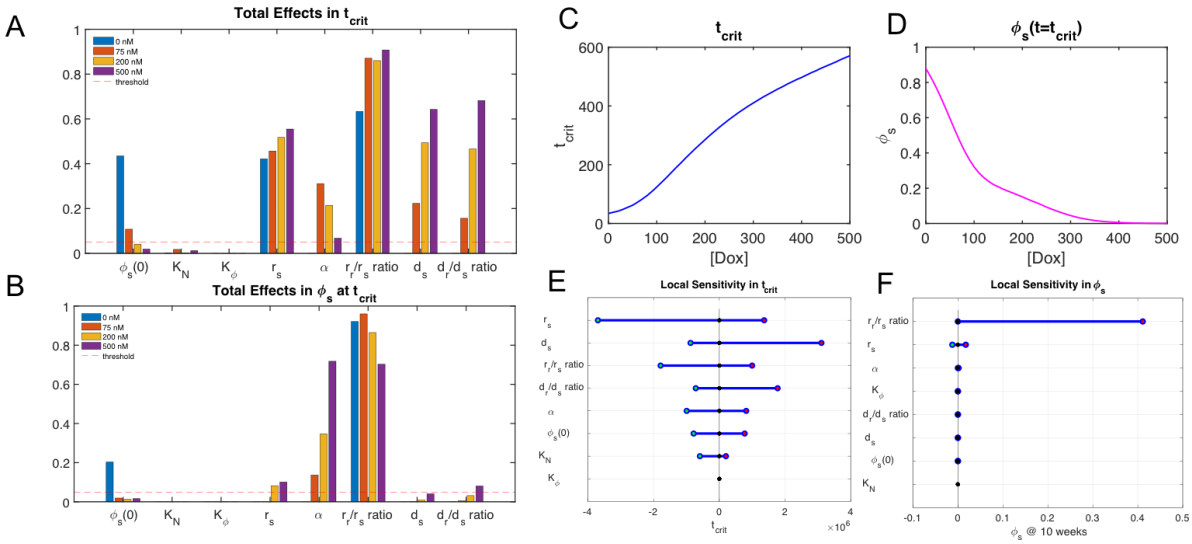

Figure 4.12: continued on next page, Sensitivity Analysis of Model Parameters Reveals All Parameters are Locally and Globally Sensitive Under Treatment. A. Sobol's total effects of each parameter globally on critical time for 0,75, 200, and 500 nM pulse treatments reveals that all fit parameters are above the threshold of sensitivity for at least one of those doses (the parameter contributes at least 5% to the critical time for at least one of the doxorubicin concentrations). B. Sobol's total effects of each parameter globally on sensitive cell fraction for 0, 75, 200 and 500 nM pulse treatments reveals that most fit parameters are above the threshold of sensitivity for at least one of the doses. The carrying capacity of the single cell RNA sequencing experiment (K_2) is the only parameter that is not above the threshold for any sensitivity analysis output or dose, and for this reason supports our decision to set that carrying capacity from a literature value (the expected number of 231 cells at confluence in a 10 cm dish, which the cells were expanded up to). C. An example of the model predicted critical time as a function of doxorubicin concentration, taken from the selected parameter set in red in Fig 5A. Critical time is chosen as an output for model sensitivity because it evaluates treatment response and drug sensitivity in of a cell population:drug concentration combination without biasing for response dynamics that might vary from system to system, and because it is most relevant to what we experimentally are able to observe (i.e. the cells rebounded to 2 times their initial cell number on this day). D. An example of the model predicted sensitive cell fraction at the critical time as a function of doxorubicin concentration, again for the selected parameter set in red in Fig 5A. This was chosen again because of its relevance to experimental workflows, as the time at which the population rebounds to 2 the seeding population might be a good time at which we could perform an experimental analysis of the tumor cell composition (i.e. scRNAseq). E. Local sensitivity in critical time produced by varying the selected parameter set by 50% above and below its value and recording the resulting change in critical time trajectory over a doxorubicin range of 0 to 500 nM. F. Local sensitivity in sensitive cell fraction at critical time produced by again varying the selected parameter set by 50% above and below its

value and recording the resulting change in sensitive cell fraction over a doxorubicin range of 0 to 500 nM.

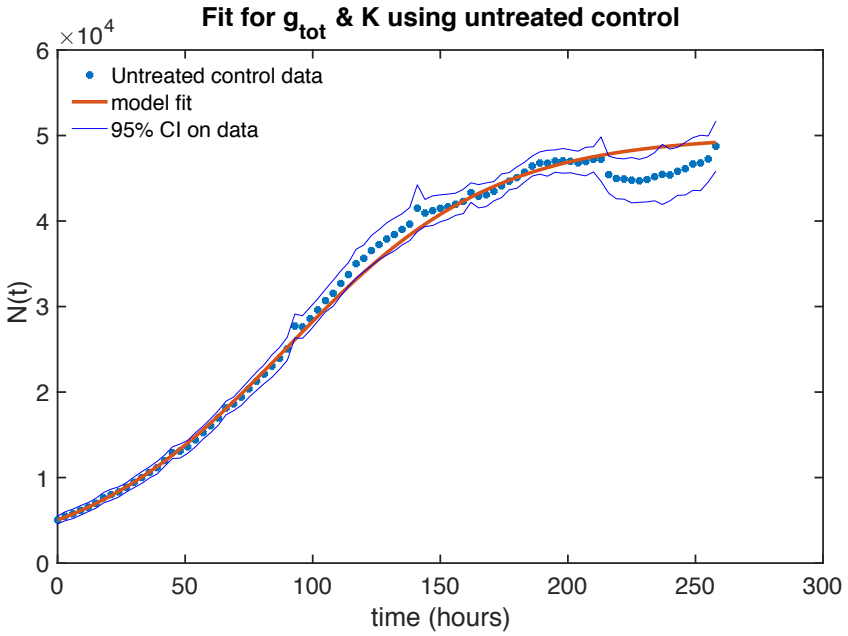

Figure 4.13: Fit to untreated control to find effective dose and carrying capacity of MDA-MB- 231 cells in a 96 well plate.

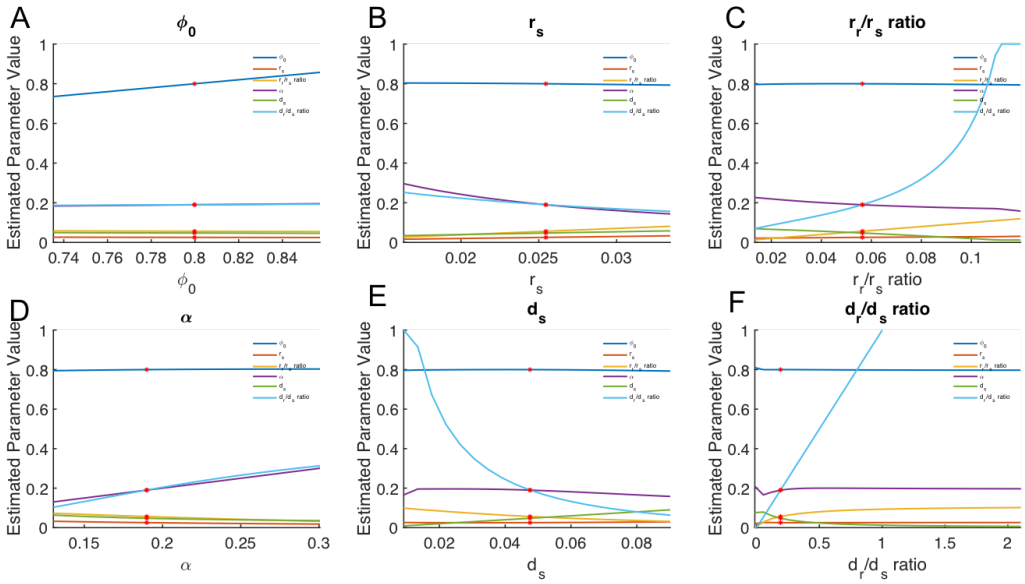

Figure 4.14: Parameter relationships from profile likelihood analysis. Plot of the how each of the remaining 5 parameters varied while "profiling" A. ϕ_0, B. r_s C. r_r/r_s ratio D. α, E. d_s, and F. d_r/d_s ratio. The resulting curves describe the parameter relationships that enable the model to fit the observed data set.

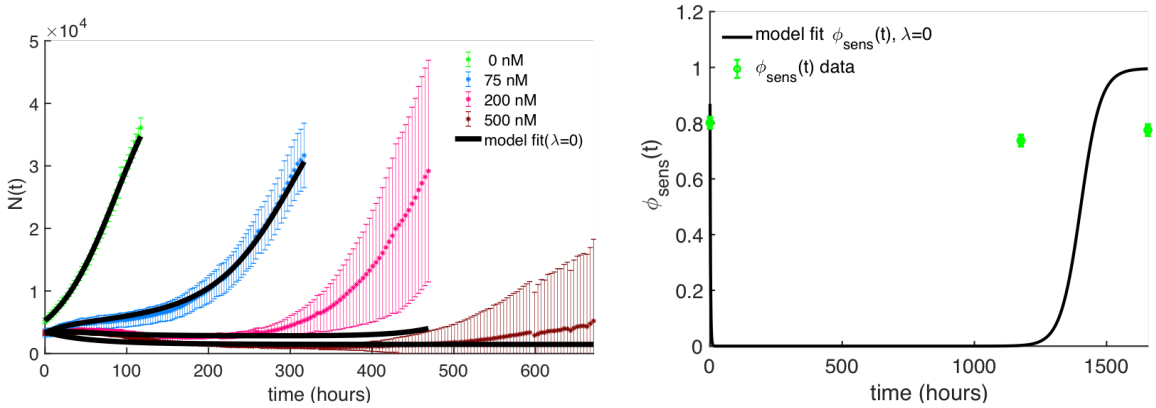

Figure 4.15: Fitting results without incorporating ϕ(t). A. Model fit compared to N(t) when only using N(t) data for calibration. Because the higher doses (200 nM and 500 nM) have a higher data uncertainty, we do not fit these doses well. The CCC of the mean N(t) compared to the model calibrated N(t) is 0.8471. B. Resulting model prediction of ϕ(t) dynamics, based on calibration from N(t) data only, with a CCC of 0.0913.

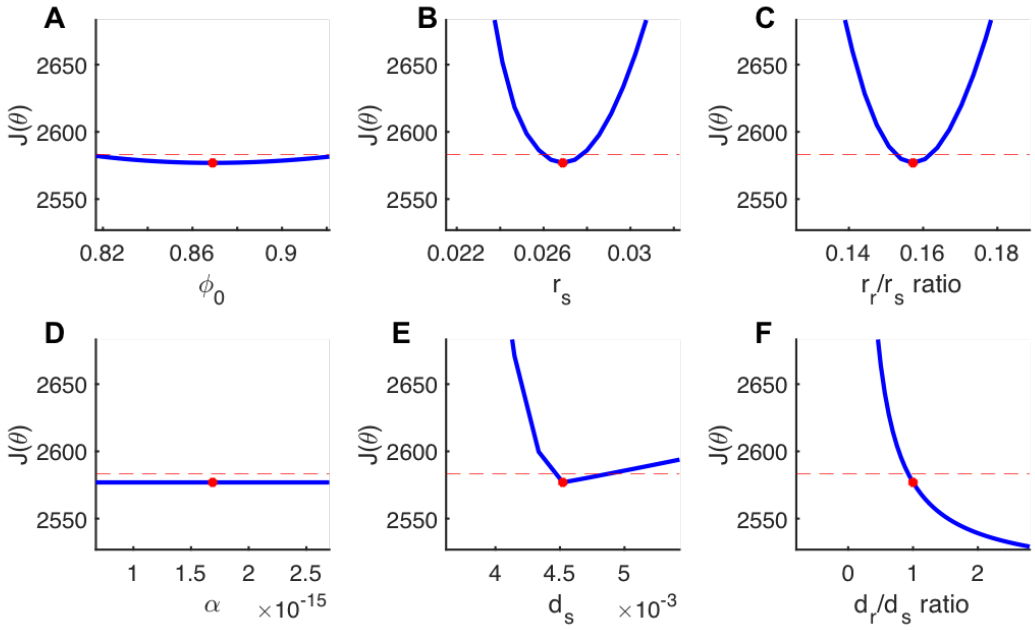

Figure 4.16: Profile likelihood results reveal that not all parameter are identifiable without incorporating ϕ(t). A. Profile likelihood around ϕ_0, the initial sensitive cell fraction, reveals that in this case, parameter is identifiable, as it does eventually cross the 95% χ^2- threshold. B. Profile likelihood around the sensitive cell growth rate, r_s, revealing the parameter is identifiable. C. Profile likelihood around the resistant-to-sensitive cell growth rate ratio reveals the parameter is identifiable. D. Profile likelihood around the drug-induced resistance rate α, revealing the parameter is unidentifiable because none of the profiled values enable the objective function value, J(θ) to cross above the threshold, indicating the value of this parameter within this region does not affect the goodness of fit of the model to the data. E. Profile likelihood around the sensitive cell death rate, d_s, revealing the parameter is identifiable. F. Profile likelihood around the resistant-to-sensitive cell death rate ratio, revealing the parameter is unidentifiable at the upper bound because the profiled values do not cross the objective function threshold, and therefore their value cannot be uniquely identified.

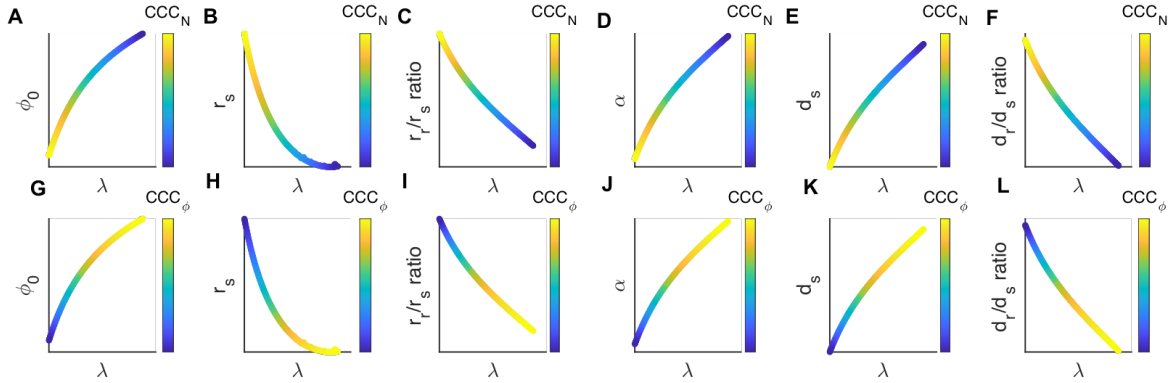

Figure 4.17: Variation of parameter values as a function of the regularization term λ, indicate that parameter values have directional bias in their goodness of fit in N(t) vs. the ϕ(t). A-F. Parameter values over the range of λ in the pareto front, colored by the value's corresponding accuracy in calibration to the N(t) data (CCC_N). G-L. Parameter values over the range of lambda in the pareto front, colored by the value's corresponding accuracy in calibration to the ϕ(t) data (CCC_ϕ).

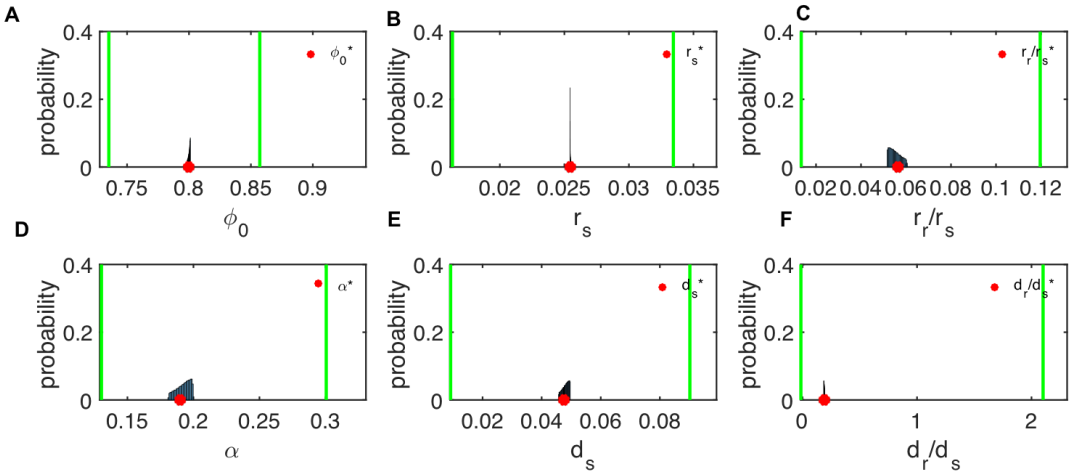

Figure 4.18: Pareto front parameter distributions fall well within the 95 % CI on θ*, found via the prolife likelihood method (Fig 6F-K) and displayed here by the green lines. A. Distribution of pareto front accepted parameter ϕ_0. B. Distribution of pareto front accepted parameter r_s. C. Distribution of pareto front accepted parameter resistant to sensitive growth rate. D. Distribution of pareto front accepted parameter α. E. Distribution of pareto front accepted parameter d_s. F. Distribution of pareto front accepted parameter resistant to sensitive cell death rate.

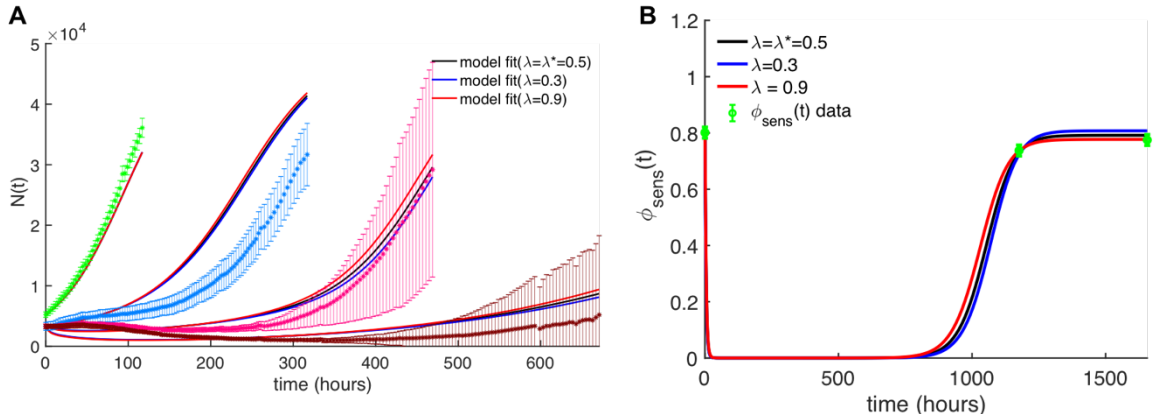

Figure 4.19: Fitting results for the range of pareto-front parameter sets A. Model fit to N(t) for the weighting by number of data points only (black, **λ** = **λ***), the lowest **λ** (**λ** = 0.3) favoring N(t) the most, and the highest **λ** (**λ**= 0.9) favoring ϕ(t) the most. B Model fit to ϕ(t) for the range of accepted pareto front parameter sets.

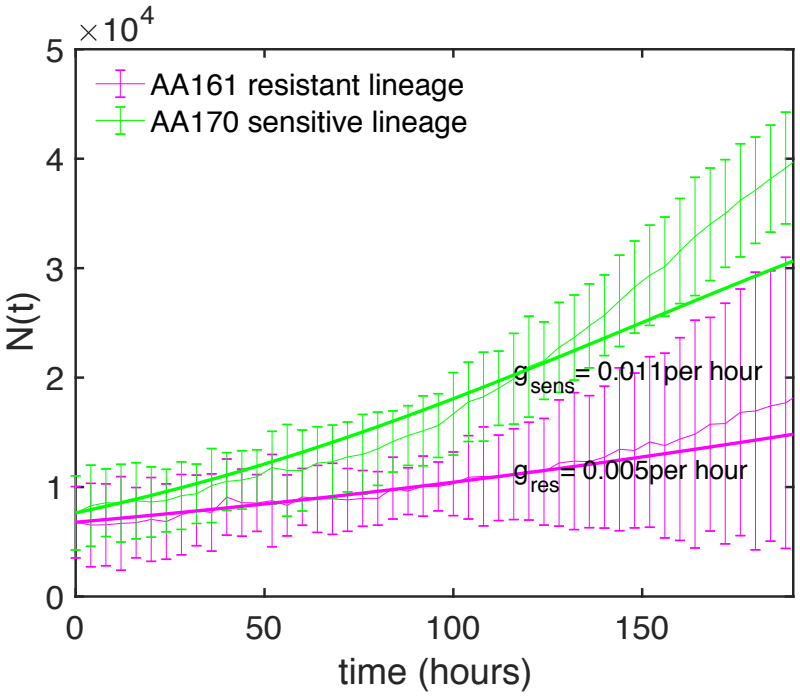

Figure 4.20: Growth dynamics of isolated sensitive and resistant cell lineages indicates that sensitive cells growth on more quickly than the resistant cells, validating our modeling assumptions.

Chapter 5: Conclusions and future work

SUMMARY OF WORK AND FUTURE IMPROVEMENTS

The aim of this research was to utilize experimental data from multimodal sources to improve how we describe, understand, predict, and optimize cancer progression in a research setting. The intention of this work is to demonstrate and validate, first in experimental cell line models, how mathematical frameworks can improve how data is used in the clinical setting. We believe that demonstrating the feasibility of these mathematical methods *in vitro* will help enable their translation and adaptation in real clinical practice to improve how doctors use data to inform treatment decisions.

In Chapter 2, we describe an investigation into the dynamic response of a population of cancer cells following chemotherapeutic treatment. This work was the essence of data-driven, in that it was fully motivated by experimental findings and an attempt to gain an understanding from those experimental findings retrospectively. Breast cancer cells (MCF7) were treated with chemotherapeutic doxorubicin at a high dose for 24 hours, and then maintained to monitor treatment response dynamics. In this work, we were interested in characterizing the drug-sensitivity of the cell population over time as the population responded to the chemotherapy, and so the drug-sensitivity of the recovering cell population to a range of doses of doxorubicin was measured weekly. Dose-response curves were analyzed to quantify the resistance of the population post-treatment in several different ways. The results of this analysis indicated that the population of cells transiently increased in overall resistance around three weeks after initial pulse-treatment.

The methods presented in Chapter 2 reflect a creative solution to make use of the available dose-response curves that were acquired longitudinally. We developed a biological hypothesis about the underlying population structure, positing that either the

population was distributed unimodally with a potential for dynamic shifting of the central resistance, or that the population consisted of a bimodal distribution of resistance in which their exist two subpopulations with differing levels of resistance whose relative proportions of the population potentially shifted over time in response to the treatment. Using model selection criteria for a single fit to the same data set, we came to the conclusion that the more likely underlying population structure was bimodal, indicating evidence of underlying heterogeneity in the breast cancer cell line population. The analysis from this investigation led us to believe that there was a transient increase in the proportion of cells in the more resistant subpopulation around three weeks after treatment response, followed by a return to nearly the same proportions of resistant and sensitive cells.

Although cell counting was performed each week as an attempt to quantify the population-size dynamics was made, this aspect of the experimental design could have been improved to quantify the unperturbed cell number over time. Ideally, to truly understand treatment response dynamics we needed not only instantaneous measures of cell viability as a function of dose, but also more precise and unperturbed measures of population size over time. This was certainly a drawback of the investigation, and future work in Chapter 4 improves upon this by using advances in technology to precisely monitor population size over time after a pulse treatment. Nonetheless, this work allowed us to "probe" the drug-sensitivity of the overall population over time during treatment response, which previously had not been performed systematically following a pulse-treatment of chemotherapy. We acknowledge that a two-state model of sensitive and resistant cells is likely a vast oversimplification of the distribution of cell-states, as they relate to drug sensitivity, that are likely to exist in the population. It is quite possible the distribution is not truly bimodal but instead multimodal, however, we were not able to add additional complexity to our model to test that hypothesis with confidence. Future work in Chapter 4

seeks to expand on this work by performing a more high throughput measurement of the population response to multiple chemotherapy concentration. It also improves on an attempt at assessing the underlying phenotypic composition of the population by instead using single cell gene expression and lineage tracing for high throughput, broad and deep quantification of individual cells within the population before and during treatment response. The work in Chapter 2 provided much of the motivation for the workflow and experimental design outlined in Chapter 4, while also attempting to address some of the shortcomings mentioned.

In Chapter 3, we were interested in investigating the relevance of an ecological phenomena that describes the slowing of per capita growth rate of populations of species at low population densities, known as the Allee effect. This investigation was motivated by observations in both tumors inoculated into mice (Panigrahy *et al.*, 2012) and regrowth of resected human glioblastomas (Neufeld *et al.*, 2017), which both indicated that perhaps cancer cells do not simply grow autonomously at a constant per capita growth rate from the single cell level to the point of a detectable tumor. Anecdotally, it seemed quite plausible that cancer cells might need to reach a critical density before entering the "exponential phase" of growth; however, most mathematical models of tumor growth focus only on deviations from exponential growth at higher population sizes. Thus, we sought to set up a combined experimental-computational design to test whether or not cancer cells, cultured in a controlled, nutrient and space optimal environment, exhibited observable slowing of growth rate at low population densities. The novelty of this investigation was that, unlike in a pre-clinical or clinical setting, we had a technology to detect the growth dynamics of very small cancer cell population sizes at single cell resolution, with a high number of replicates. In order to disentangle the stochastic effects of small population sizes from true cooperative growth via an Allee effect, we developed a stochastic parameter

estimation framework, using the moment-closure approximations described in (Fröhlich *et al.*, 2016) to compare the best structural model to describe the observed data.

In this investigation described in detail in Chapter 3, we took advantage of the novel technological advances that now enabled for high-resolution and high-throughout data that could be readily compared to stochastic simulations of different Allee effect scenarios. The novelty in this framework is in how we utilized the available data- to take advantage of the fact that we could measure the observed variability in cell growth trajectories, and use that information as an input for our model calibration. It is often the case that when we acquire cell population data we measure information about a distribution of values (i.e. protein expression levels by FACS) but oftentimes for data analysis a summary statistic (i.e. mean expression level, or percent of cells above a certain threshold of expression level) is more commonly used. While in most cases, these summary statistics are just as meaningful, in this specific case, understanding the full distribution of cell growth trajectories was necessary due to the expected variability in growth behavior inherent at small population sizes. While in most cases, these summary statistics are just as meaningful, in this specific case, understanding the full distribution of cell growth trajectories was necessary due to the expected variability in growth behavior inherent at small population sizes. In this work, we took the full distribution of replicates of growth trajectories and summarized them, by initial condition, into mean cell number and variance in cell number, in time. However, this potentially masks the full available distribution of the data set, which is shown for the final time point for each initial condition of N_0= 2 (red), 4 (green), and 10 (blue) cells in Fig. 5.1. For example, if the behavior of the cells ended up being bimodal, with cells falling into two subpopulations of slower or faster growers, the analysis in Chapter 3 would fail to capture those dynamics. Future work in this area could focus on how to better make use of the entire distribution of data for parameter estimation,

perhaps by testing each individual trajectory against a few candidate underlying distributions. Additional investigations could employ techniques such as Kalman-filtering or Gaussian processes to compare individual data points and infer underlying distributions over time.

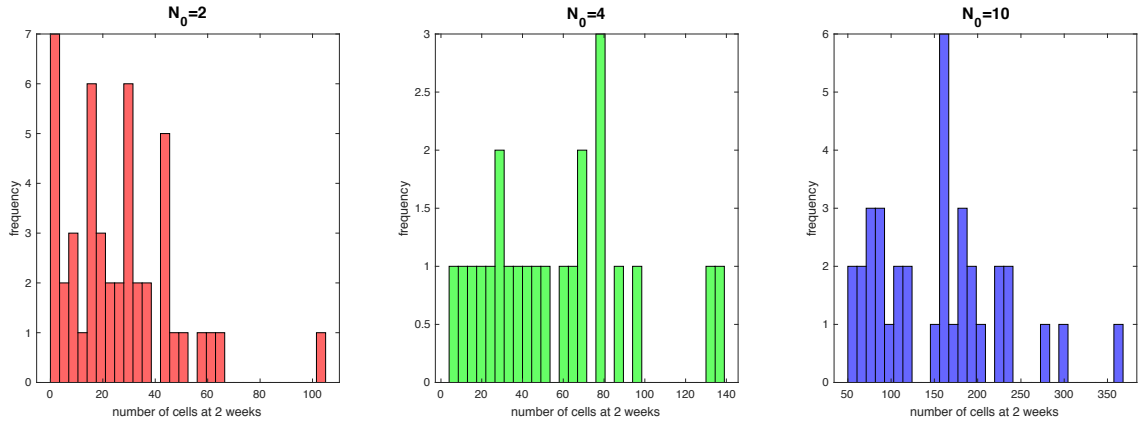

Figure 5.1. Final cell number distributions for initial conditions of 2, 4, and 10 cells. While the current framework utilized the mean and variance of each of these distributions at each time point the data was acquired, it does not explicitly use the full distribution at all time points, which could be problematic if for example we observe a multi-modal distribution that could be indicative of two possible phenotypes related to growth. Future work to improve upon this project might consider how the full extent of the observed distribution at all time points could be more fully utilized.

Our overall conclusion from the analysis in Chapter 3 was that, even in the very controlled setting with optimal cell culture conditions, BT-474 breast cancer cell growth dynamics displayed a weak Allee effect, which results in a decrease in the birth rate as a function of the total population size. We acknowledge that although this model was "selected" using model selection tools (BIC), this simply means it is the most likely of the candidate structural models. All candidate models were composed of the assumption that the cell population was made up of a single homogenous cell type, whose presence facilitated homotypic interactions leading to the positive scaling of growth rate with population size. This resulted in a phenomenological model to describe population dynamics. However, an area of future work is the investigation of a more mechanistic

model capable of characterizing the cooperative interactions that are likely due to distinct subpopulations playing complimentary roles that facilitate growth above a critical density. In order to further investigate these subpopulation interactions, distinct subpopulations of cells within a cell line can be isolated based on expressed of surface receptors, and the theory of cooperation between these cell types can be tested experimentally and analyzed via mechanistic mathematical models of cooperative interactions.

The work in Chapter 3 provided an example of ways in which technological advancements in data collection can lead to the use of novel mathematical frameworks to best make use of these new high throughput, high resolution data sets. This theme continues into the work described in Chapter 4, in which high throughput longitudinal population size data and single cell transcriptomics data are utilized in a single mathematical framework in an attempt to develop a more informed understanding of treatment response dynamics. This work brought together a number of different expertise from the Brock lab. Technological advancements in expressed barcoding of clonal populations using the COLBERT system (Al'Khafaji, Deatherage and Brock, 2018; Al'Khafaji *et al.*, 2019) developed by Aziz Al'Khafaji in the Brock lab enabled functional read-out of the lineage identity of cancer cells before, during and after treatment with chemotherapy doxorubicin. Pre-processing, read alignment, and normalization of the scRNAseq data set was performed by Russ Durrett, Eric Brenner, and Daylin Morgan, all who have developed expertise in the bioinformatics analysis of these transcriptomic data sets. Because these are cancer cell lines with stable phenotypes, Grant Howard separately performed high temporal resolution treatment response measurements of this breast cancer cell line to ten different pulse-treatment of doxorubicin and capture the population size dynamics for 6 replicates for each condition. Without all of their expertise, none of the work described in Chapter 4 would have been possible.

These datasets combined provided an abundance of information about the MDA-MB-231 breast cancer cell line response to treatment, and the major contribution of Chapter 4 was to attempt to make sense of it. These diverse sets of data are integrated in a single mathematical framework in which the lineage-traced transcriptomics data is used to estimate the composition of the population over time, and the longitudinal treatment response data is used to monitor population size over time in response to different drug concentrations. The estimates of the cell identities from lineage-traced scRNAseq represent a relatively forward application of machine learning to biological data. Given a 20,000 gene expression vector, and known class identities from changes in lineage abundance of those cells, we used principal component analysis to build a classifier capable of "predicting" the identity, in terms of drug sensitivity, of cells with unknown drug sensitivity identities. This output was chosen intentionally, because it was of interest to quantify the proportion of cells that were sensitive and resistant at the times that the transcriptomics data was acquired. These estimates went directly into a mechanistic model of treatment response dynamics, providing insight into the phenotypic composition at three time points. While we acknowledge that reducing this breadth of molecular data to a binary classification of sensitive or resistant was quite reductionist, it enabled actionable comparisons for calibrating a model that describes how these subpopulations evolved. Future work should focus on other novel ways in which big data sets can leverage machine learning to potentially interact with mechanistic modeling, providing useful information to improve our ability to inform either of these models. It is a broad theme of this thesis that collaborative interactions across disciplines breeds more relevant scientific contributions, and it is my opinion that the realms of big data, machine learning, and mechanistic mathematical modeling are no different in that the contributions at the intersection of these

fields will prove the most fruitful to developing new advancements in biology and medicine.

While the work in Chapter 4 intended to make the most use out of available data, in hindsight, many things would be done differently to improve the experimental workflow. For one, the doxorubicin concentration given to the cells that were to be sequenced with scRNA-seq would be the same as the treatment condition given to the cells acquired longitudinally, to minimize the potential free variables leading to differences in treatment response. Secondly, acquiring scRNA-seq during the intermediate time points of treatment response would have been crucial to understanding the changes in population composition at that time. However, because cells that have mostly died due to drug are not easily sampled for scRNA-seq, if we were to repeat the experiment we would look to increase the number of cells initially treated, and lower the dose of doxorubicin, enabling sequencing at the intermediate time point which represents when the cell population is likely responding to drug. Additionally, other, more flexible uses of the transcriptomics data set could be used as inputs into a mechanistic model. For example, instead of mandating that a cell either be classified as sensitive or resistant, a new cell could be mapped onto a spectrum of sensitivity, and the shift in these distributions could be used as inputs into a model that allows for a continuous distribution of drug sensitivity. There are a number of other methods one can think of to continue to develop more informed methods to make use of these data sets, and the work described in Chapter 4 is really only the beginning of a number of possible ways to integrate molecular level data into mechanistic modeling frameworks. The approach in Chapter 4 is intended to be seen as an example of one way to achieve this. It is intended to be thought provoking regarding other ways in both the research and clinical setting that multimodal data sets can be integrated to develop more

informed understanding of underlying biological systems to inform experimental or clinical decisions.

Overall, several examples of novel methods of analyzing experimental cancer cell line data of a variety of sources are presented. Although these results are promising, there is room for optimization in each of the studies and potential future applications in both experimental and clinical settings.

THE FUTURE OF PRECISION ONCOLOGY

In this section, I will outline my broader vision for the way in which I believe mathematical oncologists should seek to position themselves in the field of cancer research and clinical oncology. These viewpoints are solely my opinions, and do not necessarily reflect the work presented in this dissertation exclusively, as here I will broadly focus on the potential clinical impact of the mathematical oncologist. While the work presented in this dissertation is based purely on experimental systems using cancer cell line model systems, a career goal of mine is to build upon these experiences focused on improving our understanding and knowledge of these biological systems to those that will directly impact clinical decision-making in oncology.

In this dissertation, much of the data collected was not designed intentionally for the purpose of building mathematical models in oncology. Instead, for example in Chapter 2 and Chapter 4, the data was collected using standard techniques in experimental biology to quantify and deepen the understanding of the biology of the cancer cell line during chemotherapy treatment. Whether that was assessed via a dose-response assay or a scRNA-seq experiment, a diverse set of modalities were used that in theory, contribute to our understanding of the underlying process. However, most mathematical modeling efforts in oncology typically only take into account data amenable to modeling dynamical systems:

longitudinal data. Typically, mathematical oncologists will design experiments in order to capture this data appropriately in a way that is amenable to model calibration and downstream analysis. While I believe it is critical for the mathematical oncologist to play a role in experimental design, I also believe that the field of mathematical oncology needs to be more open and flexible to utilizing the available information, in whatever form it might be in.

In real clinical decision making, physicians are tasked with using all available information about a patient to make their treatment decision, including baseline characteristics such as age, sex, and ethnicity, as well as the available blood levels, immune levels, as well as the "big data" of genomic information, histology, and anatomy from imaging. While it might be easy to criticize the physician for not applying longitudinal data analysis to improve understanding of the dynamics of the tumor growth and treatment response within the patient, this is not their job- it is ours. The physician must focus on caring for patients, ordering the right tests, gathering as much available information as possible. Likewise, the role of the experimental biologist is not to do this type of analysis, for they must be in the lab, performing relevant experiments in search of new and more effective treatment options. It is necessarily the job of the mathematical oncologist in this scenario to be tasked with making sense of all of the available information: baseline characteristics, "omics" data, blood levels, and tumor size dynamics, to help the physician make the most informed treatment decision. Instead of cherry picking the types of data that fit into the world of mathematics, we should instead be trying to bring math, along with simple retrospective analysis of historical data about treatment response outcomes, into the clinic. Instead of doing math for the sake of math, we should be doing math for the sake of helping improve treatment-decision making.

What does this look like in practice? The opportunities are immense, but not straightforward. Perhaps baseline characteristics should be used to determine which mathematical models of treatment response are more relevant for a given patient, or perhaps they should be used to inform the patients most likely parameter values for a given model, or to direct a treatment modality most likely to be effective. But how do these models take into account all the different forms of data that may or may not be available? This question largely remains unanswered, but I think it is the job of the mathematical oncologist to stretch themselves to think about ways to answer this question, how can we help doctors make these decisions, and patients benefit from them? While this is certainly an immense challenge, it is one that we cannot place on the burden of doctors or biologists, and it is one that we ourselves need to try to address, by learning about the problem intimately, via interacting with physicians and understanding their day-to-day decision making process, and by knowing well and making best use of the available data, both retrospectively and as it is observed in real-time. While the task at hand might require a combination of machine learning, mechanistic modeling, phenomenological modeling, and analysis of historical population data, it will certainly not be achieved by only one of these modalities. Working towards this goal will require an integrated, complex approach to come up with new creative solutions that span a variety of disciplines- from basic science, to mathematics and computation, to clinical care. All parties need to have a seat at the table and learn to communicate and collaborate with one another.

In my final year of graduate school, I had the opportunity to attend an Integrated Mathematical Oncology Workshop at the Moffitt Cancer Center. The theme of the workshop was “Tumor Board Evolution”. Before I arrived, I didn’t know what a tumor board was, and so I assumed the workshop would just be an evolutionary approach to modeling cancer, of which I was relatively comfortable in as most mathematical oncology

models include some type of hypothesis regarding the evolution of tumor cells over time. Instead, the workshop was actually challenging the tumor board itself to evolve- by incorporating a mathematical oncologist. This is currently an ongoing experiment at Moffitt Cancer Center, and for this one week my teammates and I got to experience what a tumor board might be like. We met with surgical oncologists who performed surgeries to remove liver metastases from colorectal cancer patients, and they explained how at "tumor board" they presented their challenging cases to the entire team of oncologists: the radiologists, medical oncologists, and various surgeons- with the goal of coming up with a treatment plan for the patient, particularly when the case was not so straightforward. In the course of that week, we learned about the types of information the physicians had available to them about these patients during the course of their treatment, and all the many factors they had to weigh as they were adapting in real-time to the patient's treatment response with the types of data they could use to assess that.

While their options for treatment were sometimes limited, they still always had options, and struggled to select the right ones given the varieties of different pieces of information they had to weigh all at once about a patient. This is where I think the mathematical oncologist needs to step up- perhaps presenting the physician with a historical database of past treatment outcomes for similar patients to the one at hand, as well as providing a framework for interpreting the variety of types of data they have about an individual patient, and updating it in real-time in order to help them determine the course of action with the highest probability of success, given the available information. When I returned from this workshop, I began to tell my friend about this experience, and as I began to explain the "experiment" that we had undergone over the week she interjected- "Wait, so why don't they already have the mathematical oncologists on the tumor board helping

doctors make sense of all these things?" I was immediately struck by the simplicity of this question- why weren't we already doing this? I think we need to be.

CONCLUDING REMARKS

Several new approaches to utilize experimental data in mathematical oncology frameworks have been developed in this dissertation. Although in this dissertation, they are only applied to experimental systems in breast cancer cell lines, the broader goal of this work is to demonstrate how some of these questions could potentially be extended to a clinical setting. It is critical to be able to develop these workflows first in model systems experimentally, where experimental design limitations, as well as ethical and bureaucratic obstacles, are much more limited. We posit that this work represents a critical step forward to the field of mathematical oncology, as it seeks to demonstrate ways in which all different types of data sets can be utilized in an integrated manner to better inform the mathematical models that can drive improvements in treatment decision making. The ability to think critically and come up with creative solutions pulling from the expertise of a variety of disciplines is absolutely necessary to create impactful change in the field of cancer research and clinical care, and we hope that this dissertation work makes a small dent at contributing to this cause.

Bibliography

Abuhammad, S. and Zihlif, M. (2013) 'Gene expression alterations in doxorubicin resistant MCF7 breast cancer cell line', *Genomics*. Elsevier Inc., 101(4), pp. 213–220. doi: 10.1016/j.ygeno.2012.11.009.

Ai, L. *et al.* (2014) 'TRIM29 Suppresses TWIST1 and Invasive Breast Cancer Behavior', *Cancer Research*, pp. 4875–4888. doi: 10.1158/0008-5472.CAN-13-3579.

Al'Khafaji, A. *et al.* (2019) 'Expressed barcodes enable clonal characterization of chemotherapeutic responses in chronic lymphocytic leukemia', *bioRxiv*, pp. 1–24. Available at: https://dx.doi.org/10.1101/761981.

Al'Khafaji, A. M., Deatherage, D. and Brock, A. (2018) 'Control of Lineage-Specific Gene Expression by Functionalized gRNA Barcodes', *ACS Synthetic Biology*. doi: 10.1021/acssynbio.8b00105.

Amend, S. R. *et al.* (2016) 'Ecological paradigms to understand the dynamics of metastasis', *Cancer Letters*. Elsevier Ireland Ltd, 380(1), pp. 237–242. doi: 10.1016/j.canlet.2015.10.005.

Amend, S. R. and Pienta, K. J. (2015) 'Ecology meets cancer biology : The cancer swamp promotes the lethal cancer phenotype', *Oncotarget*, 6(12).

An, M. W. *et al.* (2015) 'Evaluating continuous tumor measurement-based metrics as phase II endpoints for predicting overall survival', *Journal of the National Cancer Institute*, 107(11), pp. 1–7. doi: 10.1093/jnci/djv239.

Anderson, A. R. A. *et al.* (2000) 'Mathematical Modelling of Tumour Invasion and Metastasis', *Journal of Theoretical Medicine*. Hindawi Publishing Corporation, 2, p. 490902. doi: 10.1080/10273660008833042.

Anderson, A. R. A. *et al.* (2006) 'Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment', *Cell*, 127(5), pp. 905–915. doi: 10.1016/j.cell.2006.09.042.

Anderson, A. R. A. and Maini, P. K. (2018) 'Mathematical Oncology', *Bulletin of Mathematical Biology*. Springer US, 80(5), pp. 945–953. doi: 10.1007/s11538-018-0423-5.

Araujo, A. *et al.* (2018) 'Size Matters: Metastatic Cluster Size and Stromal Recruitment in the Establishment of Successful Prostate Cancer to Bone Metastases', *Bulletin of Mathematical Biology*. Springer US, 80(5), pp. 1046–1058. doi: 10.1007/s11538-018-0416-4.

Archetti, M., Ferraro, D. A. and Christofori, G. (2015) 'Heterogeneity for IGF-II production maintained by public goods dynamics in neuroendocrine pancreatic cancer', *Proceedings of the National Academy of Sciences*, 112(6), pp. 1833–1838. doi: 10.1073/pnas.1414653112.

Axelrod, R., Axelrod, D. E. and Pienta, K. J. (2006) 'Evolution of cooperation among tumor cells', *Proceedings of the National Academy of Sciences*, 103(36), pp. 13474–13479.

Axelrod, R. and Pienta, K. J. (2018) 'Cancer as a Social Dysfunction — Why Cancer Research

Needs New Thinking', *Molecular Cancer Research*, 16(9), pp. 1346–1347. doi: 10.1158/1541-7786.MCR-18-0013.

Badri, H. *et al.* (2016) 'Optimization of radiation dosing schedules for proneural glioblastoma', *Journal of Mathematical Biology*. Springer Berlin Heidelberg, 72(5), pp. 1301–1336. doi: 10.1007/s00285-015-0908-x.

Basanta, D. *et al.* (2013) 'Exploiting ecological principles to better understand cancer progression and treatment', *Interface Focus*.

Bauer, G. *et al.* (2018) 'Letter The Science of Living Matter for Tomorrow', *Cell Systems*. Elsevier Inc., 6(4), pp. 400–402. doi: 10.1016/j.cels.2018.04.003.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) 'Approximate Bayesian Computation in Population Genetics', *Genetics Society of America*, 162(December), pp. 2025–2035.

Behjati, S. and Tarpey, P. S. (2013) 'What is next generation sequencing?', *Archives of Disease in Childhood: Education and Practice Edition*, 98(6), pp. 236–238. doi: 10.1136/archdischild-2013-304340.

Benzekry, S. *et al.* (2014) 'Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth', *PLoS Computational Biology*, 10(8). doi: 10.1371/journal.pcbi.1003800.

Bose, I. *et al.* (2017) 'Allee dynamics: Growth, extinction and range expansion', *arXiv*, pp. 1–9.

Böttger, K., Hatzikirou, H. and Voss-böhme, A. (2015) 'An Emerging Allee Effect Is Critical for Tumor Initiation and Persistence', *PLoS Computational Biology*, pp. 1–14. doi: 10.1371/journal.pcbi.1004366.

Brabletz, T. *et al.* (2018) 'EMT in cancer', *Nature*, 18, pp. 128–134. doi: 10.1038/nrc.2017.118.

Brady, R. *et al.* (2019) 'Prostate-Specific Antigen Dynamics Predict Individual Responses to Intermittent Androgen Deprivation', *bioRxiv*. Available at: https://dx.doi.org/10.1101/624866.

Brenner, D. J. (2008) 'The Linear-Quadratic Model Is an Appropriate Methodology for Determining Isoeffective Doses at Large Doses Per Fraction', *Seminars in Radiation Oncology*, 18(4), pp. 234–239. doi: 10.1016/j.semradonc.2008.04.004.

Brock, A., Chang, H. and Huang, S. (2009a) 'Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours', *Nature Re*, 10, pp. 336–342.

Brock, A., Chang, H. and Huang, S. (2009b) 'Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours', *Nature Reviews Genetics*, 10(5), pp. 336–342. doi: 10.1038/nrg2556.

Brock, A. and Huang, S. (2017) 'Precision Oncology : Between Vaguely Right and Precisely Wrong', pp. 1–8. doi: 10.1158/0008-5472.CAN-17-0448.

Brock, A., Krause, S. and Ingber, D. E. (2015) 'Control of cancer formation by intrinsic genetic noise and microenvironmental cues', *Nature Reviews Cancer*, 15(8), pp. 499–509. doi: 10.1038/nrc3959.

Brouwer, A. F. *et al.* (2017) 'A systematic approach to determining the identifiability of multistage carcinogenesis models', *Risk Analysis*, 37(7), pp. 1375–1387. doi:

10.1111/risa.12684.A.

Brown, J. L. *et al.* (1990) 'Clonal analysis of a bladder cancer cell line: tumour heterogeneity experimental model of', *British Journal of Cancer*, 61, pp. 369–376.

Büttner, M. *et al.* (2019) 'A test metric for assessing single-cell RNA-seq batch correction', *Nature Methods*, 16(1), pp. 43–49. doi: 10.1038/s41592-018-0254-1.

Byrne, H. M. (2010) 'Dissecting cancer through mathematics: from the cell to the animal model', *Nature Reviews Cancer*. Nature Publishing Group, 10, p. 221. Available at: https://doi.org/10.1038/nrc2808.

Cao, Y. and Petzold, L. (2006) 'Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems', *Journal of Computational Physics*, 212(1), pp. 6–24. doi: 10.1016/j.jcp.2005.06.012.

Censor, Y. (1977) 'Pareto Optimality in Multiobjective Problems', *Applied Mathematics and Optimization*, 4, pp. 41–59.

Chaffer, C. L. *et al.* (2016) 'EMT , cell plasticity and metastasis', *Cancer and Metastasis Reviews*. Cancer and Metastasis Reviews, (November), pp. 645–654. doi: 10.1007/s10555-016-9648-7.

Chen, K. and Pienta, K. J. (2011) 'Modeling invasion of metastasizing cancer cells to bone marrow utilizing ecological principles', *Theoretical Biology and Medical Modelling*, 8(36), pp. 1–11.

Chen, P. *et al.* (2018) 'Adaptive and Reversible Resistance to Kras Inhibition in Pancreatic Cancer Cells', *Cancer Research*, pp. 985–1003. doi: 10.1158/0008-5472.CAN-17-2129.

Chen, S. *et al.* (2018) 'Dissecting heterogeneous cell-populations across signaling and disease conditions with PopAlign'.

Chisholm, R. H., Lorenzi, T. and Clairambault, J. (2016) 'Cell population heterogeneity and evolution towards drug resistance in cancer: Biological and mathematical assessment, theoretical treatment optimisation', *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(11, Part B), pp. 2627–2645. doi: https://doi.org/10.1016/j.bbagen.2016.06.009.

Cho, H. *et al.* (2018) 'Modelling acute myeloid leukaemia in a continuum of differentiation states', *Letters Biomath*, 5, pp. 1–41. doi: 10.1080/23737867.2018.1472532.Modelling.

Citron, M. L. *et al.* (2003) 'Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: First report of Intergroup Trial C9741/Cancer and Leukemia', *Journal of Clinical Oncology*, 21(8), pp. 1431–1439. doi: 10.1200/JCO.2003.09.081.

CiViC: Clinical Interpretation of Variants in Cancer (no date). Available at: https://civicdb.org/home (Accessed: 14 February 2020).

Cleary, A. S. *et al.* (2014) 'Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers', *Nature*. Nature Publishing Group, 508(1), pp. 113–117. doi: 10.1038/nature13187.

Cloonan, N. *et al.* (2008) 'Stem cell transcriptome profiling via massive-scale mRNA

sequencing', *Nature Methods*, 5(7), pp. 613–619. doi: 10.1038/nmeth.1223.

Coetzee, F. M. and Stonick, V. L. (1996) 'On a natural homotopy between linear and nonlinear single-layer networks', *IEEE Transactions on Neural Networks*, 7(2), pp. 307–317. doi: 10.1109/72.485634.

Cold Spring Harbor Laboratory: Meetings & Courses Program (2019). Available at: https://meetings.cshl.edu/ (Accessed: 14 February 2020).

Courchamp, F., Berec, L. and Gascoigne, J. (2008) *Allee Effects in Ecology and Conservation*. New York: Oxford University Press.

Crick, F. and Watson, J. (1953) '© 1953 Nature Publishing Group'.

Davies, J. E. H. *et al.* (2015) 'Vemurafenib resistance reprograms melanoma cells towards glutamine dependence', *Journal of Translational Medicine*. BioMed Central, pp. 1–11. doi: 10.1186/s12967-015-0581-2.

Deberardinis, R. J. *et al.* (2008) 'Review The Biology of Cancer : Metabolic Reprogramming Fuels Cell Growth and Proliferation', *Cell*, (January), pp. 11–20. doi: 10.1016/j.cmet.2007.10.002.

Deberardinis, R. J. and Chandel, N. S. (2016) 'Fundamentals of cancer metabolism', (May).

Diniz, W. J. S. and Canduri, F. (2017) 'Bioinformatics: An overview and its applications', *Genetics and Molecular Research*, 16(1). doi: 10.4238/gmr16019645.

Duncan, R. P. *et al.* (2014) 'Quantifying invasion risk: The relationship between establishment probability and founding population size', *Methods in Ecology and Evolution*, 5(11), pp. 1255–1263. doi: 10.1111/2041-210X.12288.

Efron, B. (1987) 'Better Bootstrap Confidence Intervals', 82(397), pp. 171–185. doi: https://doi.org/10.1080/01621459.1987.10478410.

Eisenberg, M. C. (2019) 'Input-output equivalence and identifiability: some simple generalizations of the differential algebra approach', *arXiv*, pp. 1–25.

Eisenberg, M. C., Robertson, S. L. and Tien, J. H. (2013) 'Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease', *Journal of Theoretical Biology*. Elsevier, 324, pp. 84–102. doi: 10.1016/j.jtbi.2012.12.021.

Elosegui-artola, A. *et al.* (2017) 'Force Triggers YAP Nuclear Entry by Regulating Transport across Nuclear Pores Article Force Triggers YAP Nuclear Entry by Regulating Transport across Nuclear Pores', *Cell*. Elsevier, 171(6), pp. 1397-1410.e14. doi: 10.1016/j.cell.2017.10.008.

Enderling, H. and Chaplain, M. A. (2014) 'Mathematical Modeling of Tumor Growth and Treatment', *Current Pharmaceutical Design*, 20. doi: 10.1109/ICGEC.2010.193.

Enriquez-navas, P. M. *et al.* (2016) 'Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer', 8(327).

Fallahi-sichani, M. *et al.* (2017) 'Adaptive resistance of melanoma cells to RAF inhibition via reversible induction of a slowly dividing de-differentiated state', pp. 1–24. doi: 10.15252/msb.20166796.

Ferrall-Fairbanks, M. C. *et al.* (2019) 'Leveraging Single-Cell RNA Sequencing Experiments to Model Intratumor Heterogeneity', *Clinical Cancer Informatics*, pp. 1–10. Available at:

https://doi.org/10. 1200/CCI.18.00074.

Fillmore, C. M. and Kuperwasser, C. (2008a) 'Human breast cancer cell lines contain stem-like cells that self-renew , give rise to phenotypically diverse progeny and survive chemotherapy', *Breast Cancer Research*, 10(2), pp. 1–13. doi: 10.1186/bcr1982.

Fillmore, C. M. and Kuperwasser, C. (2008b) 'Research article Human breast cancer cell lines contain stem-like cells that self-renew , give rise to phenotypically diverse progeny and survive chemotherapy', 10(2), pp. 1–13. doi: 10.1186/bcr1982.

Foo, J. and Michor, F. (2009) 'Evolution of Resistance to Targeted Anti-Cancer Therapies during Continuous and Pulsed Administration Strategies', 5(11). doi: 10.1371/journal.pcbi.1000557.

Frohlich, F. *et al.* (2018) 'Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model', *Cell Systems*, 7, pp. 567–579. doi: 10.1016/j.cels.2018.10.013.

Fröhlich, F. *et al.* (2016) 'Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion', *PLoS Computational Biology*, 12(7), pp. 1–28. doi: 10.1371/journal.pcbi.1005030.

Gallasch, R. *et al.* (2013) 'Mathematical models for translational and clinical oncology', *Journal of Clinical Bioinformatics*. Journal of Clinical Bioinformatics, 3(1), p. 1. doi: 10.1186/2043-9113-3-23.

Gardner, S. N. (2000) 'A mechanistic, predictive model of dose-response curves for cell cycle phase-specific and -nonspecific drugs', *Cancer Research*, 60(5), pp. 1417–1425.

Gatenby, R. A. (1991) 'Population Ecology Issues in Tumor Growth', *Cancer Research*, 2, pp. 2542–2548.

Gatenby, R. A. *et al.* (2009) 'Adaptive therapy', *Cancer Research*, 69(11), pp. 4894–4903. doi: 10.1158/0008-5472.CAN-08-3658.

Gatenby, R. A. and Maini, P. K. (2003) 'Cancer summed up', *Nature*, 421(6921), p. 321. doi: 10.1038/421321a.

Gerlee, P. and Altrock, P. M. (2017) 'Extinction rates in tumour public goods games', *Journal of Royal Society Interface*, 14.

Gevertz, J. L., Greene, J. M. and Sontag, E. D. (2019) 'Validation of a Mathematical Model of Cancer Incorporating Spontaneous and Induced Evolution to Drug Resistance', *bioRxiv*, pp. 1–15. Available at: http://dx.doi.org/10.1101/2019.12.27.889444.

Gillespie, D. T. (1977) 'Exact Stochastic Simulation of Coupled Chemical Reactions', *The Journal of Physical Chemistry*, 81(25), pp. 2340–2361.

Gillespie, D. T. (2014) 'The chemical Langevin equation', *The Journal of Chemical Physics*, 297(2000). doi: 10.1063/1.481811.

Goel, S. *et al.* (2011) 'Normalization of the vasculature for treatment of cancer and other diseases', *Physiology Review*, 91, pp. 1071–1121. doi: 10.1152/physrev.00038.2010.

Gordon and Betty Moore Foundation (2020).

Gottesman, M. M. (2002) 'Mechanisms of Cancer Drug Resistance', *Annual Review of Medicine53*, 53, pp. 615–27.

Graeber, T. G. and Eisenberg, D. (2001) 'Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles', *Nature Genetics*, 29(3), pp. 295–300. doi: 10.1038/ng755.

Greene, J. M. *et al.* (2015) 'Modeling intrinsic heterogeneity and growth of cancer cells', *Journal of Theoretical Biology*. Elsevier, 367, pp. 262–277. doi: 10.1016/j.jtbi.2014.11.017.

Greene, J. M. *et al.* (2016) 'Mathematical Modeling Reveals That Changes to Local Cell Density Dynamically Modulate Baseline Variations in Cell Growth and Drug Response', *Cancer Research*, 76(10), pp. 2882–2891. doi: 10.1158/0008-5472.CAN-15-3232.

Greene, J. M. and Gevertz, J. L. (2017) 'A mathematical approach to differentiate spontaneous and induced evolution to drug resistance during cancer treatment', *bioRxiv*, pp. 1–30.

Greene, J. M., Gevertz, J. L. and Sontag, E. D. (2019) 'Mathematical Approach to Differentiate Spontaneous and Induced Evolution to Drug Resistance During Cancer Treatment abstract', *JCO Clinical Cancer Informatics*, pp. 42–49. doi: https://doi.org/10.1200/CCI.18.00087.

Greene, J. M., Sanchez-Tapia, C. and Sontag, E. D. (2018a) 'Mathematical Details on a Cancer Resistance Model', *bioRxiv*, pp. 1–42. Available at: https://dx.doi.org/10.1101/475533.

Greene, J. M., Sanchez-Tapia, C. and Sontag, E. D. (2018b) 'Mathematical Details on a Cancer Resistance Model Mathematical Modeling of Induced Drug Resistance'.

Guo, J. *et al.* (2018) 'Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development', *Cell Stem Cell*. Elsevier Inc., 21(4), pp. 533–546. doi: 10.1016/j.stem.2017.09.003.

Gupta, Piyush B. *et al.* (2011) 'Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells', *Cell*. Elsevier Inc., 146(4), pp. 633–644. doi: 10.1016/j.cell.2011.07.026.

Gupta, Piyush B *et al.* (2011) 'Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells', *Cell*. Elsevier Inc., 146(4), pp. 633–644. doi: 10.1016/j.cell.2011.07.026.

Hafner, M. *et al.* (2016) 'Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs', *Nature Methods*, 13(6), pp. 521–527. doi: 10.1038/nmeth.3853.

Han, J. *et al.* (2016) 'Rapid emergence and mechanisms of resistance by U87 glioblastoma cells to doxorubicin in an in vitro tumor microfluidic ecology', *Proceedings of the National Academy of Sciences*, 113(50), pp. 14283–14288. doi: 10.1073/pnas.1614898113.

Hangauer, M. J. *et al.* (2017) 'to GPX4 inhibition', *Nature Publishing Group*. Nature Publishing Group, 551(7679), pp. 247–250. doi: 10.1038/nature24297.

Hansen, E., Woods, R. J. and Read, A. F. (2017) 'How to Use a Chemotherapeutic Agent When Resistance to It Threatens the Patient', *PLoS Biology*, 15(2), pp. 1–21. doi: 10.1371/journal.pbio.2001110.

Hardeman, Keisha N *et al.* (2017) 'Dependence On Glycolysis Sensitizes BRAF-mutated Melanomas For Increased Response To Targeted BRAF Inhibition', *Nature Publishing*

Group. Nature Publishing Group, (October 2016), pp. 1–9. doi: 10.1038/srep42604.

Hardeman, Keisha N. *et al.* (2017) 'Dependence On Glycolysis Sensitizes BRAF-mutated Melanomas For Increased Response To Targeted BRAF Inhibition', *Scientific Reports*. Nature Publishing Group, 7(October 2016), p. 42604. doi: 10.1038/srep42604.

Harris, L. A. *et al.* (2016) 'An unbiased metric of antiproliferative drug effect in vitro', *Nature Methods*, 13(6), pp. 497–500. doi: 10.1038/nmeth.3852.

He, J. and Ahuja, N. (2015) 'Personalized Approaches to Gastrointestinal Cancers: Importance of Integrating Genomic Information to Guide Therapy', *Surgical Clinics North America*, 95(5), pp. 1081–1094. doi: 10.1016/j.suc.2015.05.002.

Heiden, M. G. Vander *et al.* (2009) 'Understanding the Warburg Effect : Cell Proliferation', *Science*, 324(May), pp. 1029–1034.

Hoelzinger, D. B., Demuth, T. and Berens, M. E. (2007) 'Autocrine factors that sustain glioma invasion and paracrine biology in the brain microenvironment', *Journal of the National Cancer Institute*, 99(21), pp. 1583–1593. doi: 10.1093/jnci/djm187.

Hormuth, D. A. *et al.* (2019) 'Calibrating a Predictive Model of Tumor Growth and Angiogenesis with Quantitative MRI', *Annals of Biomedical Engineering*, 47(7), pp. 1539–1551. doi: 10.1007/s10439-019-02262-9.

Houchmandzadeh, B. (2009) 'Extracting moments from Master Equations', *ArXiv*, 1(2), pp. 1–14.

Howard, G. R. *et al.* (2018) 'A multi-state model of chemoresistance to characterize phenotypic dynamics in breast cancer', *Scientific Reports*. Springer US, (July), pp. 1–11. doi: 10.1038/s41598-018-30467-w.

Huang, S. (2011) 'The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?', *Bioessays*, pp. 149–157. doi: 10.1002/bies.201100031.

Huang, S. (2013) 'Genetic and non-genetic instability in tumor progression: Link between the fitness landscape and the epigenetic landscape of cancer cells', *Cancer and Metastasis Reviews*, 32(3–4), pp. 423–448. doi: 10.1007/s10555-013-9435-7.

Ibrahim-Hashim, A. *et al.* (2017) 'Defining cancer subpopulations by adaptive strategies rather than molecular properties provides novel insights into intratumoral evolution', *Cancer Research*, 77(9), pp. 2242–2254. doi: 10.1158/0008-5472.CAN-16-2844.

Islam, S. *et al.* (2014) 'Quantitative single-cell RNA-seq with unique molecular identifiers', *Nature Methods*, 11(2). doi: 10.1038/nmeth.2772.

Jarrett, A. M. *et al.* (2015) 'Global sensitivity analysis used to interpret biological experimental results', *Journal of Mathematical Biology*. Springer Berlin Heidelberg, 71, pp. 151–170. doi: 10.1007/s00285-014-0818-3.

Jarrett, A. M., Bloom, M. J., *et al.* (2018) 'Mathematical modelling of trastuzumab-induced immune response in an in vivo murine model of HER2 + breast cancer', 2, pp. 1–30. doi: 10.1093/imammb/dqy014.

Jarrett, A. M., Lima, E. A. B. F., *et al.* (2018) 'Mathematical models of tumor cell proliferation: A review of the literature', *Expert Review of Anticancer Therapy*. Taylor & Francis,

18(12), pp. 1271–1286. doi: 10.1080/14737140.2018.1527689.

Jiang, Q. *et al.* (2019) 'Quorum Sensing : A Prospective Therapeutic Target for Bacterial Diseases', *BioMed Research International*, 2019. Available at: https://doi.org/10.1155/2019/2015978%0AReview.

Johnson, K. E. *et al.* (2019) 'Directional inconsistency between Response Evaluation Criteria in Solid Tumors (RECIST) time to progression and response speed and depth', *European Journal of Cancer*. Elsevier Ltd, 109, pp. 196–203. doi: 10.1016/j.ejca.2018.11.008.

Jolly, M. K. *et al.* (2017) 'Epithelial/mesenchymal plasticity: how have quantitative mathematical models helped improve our understanding?', *Molecular Oncology*, 11(7), pp. 739–754. doi: 10.1002/1878-0261.12084.

Junttila, M. R. and Sauvage, F. J. De (2013) 'Influence of tumour micro-environment heterogeneity on therapeutic response', *Nature*, 501, pp. 346–354. doi: 10.1038/nature12626.

Justman, Q. (2018) 'Editorial Splitting the World with Absolute Measurements : A Call for Collaborations in Physical Biology', *Cell Systems*. Elsevier, 6(4), pp. 395–396. doi: 10.1016/j.cels.2018.04.006.

Kalluri, R. *et al.* (2010) 'The basics of epithelial-mesenchymal transition', *The Journal of Clinical Investigation*, 119(May), pp. 1420–1428. doi: 10.1172/JCI39104.1420.

Kann, B. H. *et al.* (2019) 'Artificial Intelligence in Oncology: Current Applications and Future Directions', *Oncology Journal*, 33.

Kansal, A. R. *et al.* (2000) 'Cellular automaton of idealized brain tumor growth dynamics', *BioSystems*, 55(1–3), pp. 119–127. doi: 10.1016/S0303-2647(99)00089-1.

Kaznatcheev, A. *et al.* (2019) 'Fibroblasts and alectinib switch the evolutionary games played by non-small cell lung cancer', *Nature Ecology & Evolution*. Springer US, 3(March). doi: 10.1038/s41559-018-0768-z.

Ke, W. *et al.* (2011) 'MCF-7/ADR cells (re-designated NCI/ADR-RES) are not derived from MCF-7 breast cancer cells: A loss for breast cancer multidrug-resistant research', *Medical Oncology*, 28(SUPPL. 1), pp. 135–141. doi: 10.1007/s12032-010-9747-1.

Kimmel, G. J. *et al.* (2019) 'Neighborhood size-effects shape growing population dynamics in evolutionary public goods games', *Communications Biology*, 2(53), pp. 1–10. Available at: https://doi.org/10.1038/s42003-019-0299-4.

Klein, A. M. *et al.* (2015) 'Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells', *Cell*. Elsevier Inc., 161(5), pp. 1187–1201. doi: 10.1016/j.cell.2015.04.044.

Kobayashi, H. *et al.* (2017) 'A method for evaluating the performance of computer- aided detection of pulmonary nodules in lung cancer CT screening: detection limit for nodule size and density', *British Journal of Radiology*, 90.

Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, *Springer*. Springer. Available at: http://www.springer.com/978-0-387-71886-6.

Konstorum, A., Hillen, T. and Lowengrub, J. (2016) 'Feedback Regulation in a Cancer Stem Cell Model can Cause an Allee Affect', *Bulletin of Mathematical Biology*, 78(4), pp.

754–785. doi: 10.1007/s11538-016-0161-5.Feedback.

Korolev, K. S., Xavier, J. B. and Gore, J. (2014) 'Turning ecology and evolution against cancer', *Nature Reviews Cancer*. Nature Publishing Group, 14(5), pp. 371–380. doi: 10.1038/nrc3712.

Kowarz, E., Loescher, D. and Marschalek, R. (2015) 'Optimized Sleeping Beauty transposons enable robust stable transgenic cell lines', *Biotechnology Journal*, 41, pp. 647–53. doi: 10.1002/biot.201400821.

Kumar, M. P. *et al.* (2018) 'Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics', *Cell Reports*. ElsevierCompany., 25(6), pp. 1458-1468.e4. doi: 10.1016/j.celrep.2018.10.047.

L. Lun, A. T., Bach, K. and Marioni, J. C. (2016) 'Pooling across cells to normalize single-cell RNA sequencing data with many zero counts', *Genome Biology*, 17(1), p. 75. doi: 10.1186/s13059-016-0947-7.

Lavi, O. *et al.* (2013) 'The role of cell density and intratumoral heterogeneity in multidrug resistance', *Cancer Research*, 73(24), pp. 7168–7175. doi: 10.1158/0008-5472.CAN-13-1768.

Levitin, H. M., Yuan, J. and Sims, P. A. (2018) 'Single-Cell Transcriptomic Analysis of Tumor Heterogeneity', *Trends in Cancer*, 4(4), pp. 264–268. doi: 10.1016/j.trecan.2018.02.003.

Li, Q., Wennborg, A., Aurell, E., Dekel, E., Zou, J.-Z., *et al.* (2016) 'Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape', *Proceedings of the National Academy of Sciences*, 113(10), pp. 2672–2677. doi: 10.1073/pnas.1519210113.

Li, Q., Wennborg, A., Aurell, E., Dekel, E., Zou, J., *et al.* (2016) 'Dynamics inside the cancer cell attractor reveal cell heterogeneity , limits of stability , and escape'. doi: 10.1073/pnas.1519210113.

Lima, E. A. B. F. *et al.* (2016) 'Selection, calibration, and validation of models of tumor growth', *Mathematical Models in Applied Science*. doi: 10.1142/S021820251650055X.

Loos, C. *et al.* (2018) 'A Hierarchical, Data-Driven Approach to Modeling Single-Cell Populations Predicts Latent Causes of Cell-To-Cell Variability', *Cell Systems*, 6(5), pp. 593-603.e13. doi: 10.1016/j.cels.2018.04.008.

Luecken, M. D. and Theis, F. J. (2019) 'Current best practices in single-cell RNA-seq analysis : a tutorial', *Molecular Systems Biology*, 15(e8746). doi: 10.15252/msb.20188746.

Ma, K.-Y. *et al.* (2019) 'Single-cell RNA sequencing of lung adenocarcinoma reveals heterogeneity of immune response–related genes', *JCI Insight*. The American Society for Clinical Investigation, 4(4). doi: 10.1172/jci.insight.121387.

Maley, C. C. *et al.* (2017) 'Classifying the evolutionary and ecological features of neoplasms', *Nature Publishing Group*. Nature Publishing Group, 17(10), pp. 605–619. doi: 10.1038/nrc.2017.69.

Mani, S. A. *et al.* (2008) 'The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells', *Cell*, pp. 704–715. doi: 10.1016/j.cell.2008.03.027.

Manno, G. La *et al.* (2018) 'RNA velocity of single cells', *Nature*. Springer US. doi: 10.1038/s41586-018-0414-6.

Marusyk, A. *et al.* (2014) 'Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity', *Nature*. Nature Publishing Group, 514(7520), pp. 54–58. doi: 10.1038/nature13556.

Mátés, L. *et al.* (2009) 'Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates', *Nature Genetics*, 41(6), pp. 753–761. doi: 10.1038/ng.343.

Mcgregor, N., Axelrod, R. and Axelrod, D. E. (2008) 'Ecological Therapy for Cancer : Defining Tumors Using an Ecosystem Paradigm Suggests New Opportunities for Novel Cancer Treatments', *Translational Oncology*, 1(4), pp. 158–164. doi: 10.1593/tlo.08178.

McKenna, M. T. *et al.* (2017) 'A Predictive Mathematical Modeling Approach for the Study of Doxorubicin Treatment in Triple Negative Breast Cancer', *Scientific Reports*. Springer US, 7(1), pp. 1–14. doi: 10.1038/s41598-017-05902-z.

McKenna, Matthew T., Weis, J. A., Brock, A., *et al.* (2018) 'Precision Medicine with Imprecise Therapy: Computational Modeling for Chemotherapy in Breast Cancer', *Translational Oncology*. The Authors, 11(3), pp. 732–742. doi: 10.1016/j.tranon.2018.03.009.

McKenna, Matthew T *et al.* (2018) 'Variable Cell Line Pharmacokinetics Contribute to Non-Linear Treatment Response in Heterogeneous Cell Populations', *Annals of biomedical engineering*. 2018/02/26, 46(6), pp. 899–911. doi: 10.1007/s10439-018-2001-2.

McKenna, Matthew T., Weis, J. A., Quaranta, V., *et al.* (2018) 'Variable Cell Line Pharmacokinetics Contribute to Non-Linear Treatment Response in Heterogeneous Cell Populations', *Annals of Biomedical Engineering*, 46(6), pp. 899–911. doi: 10.1007/s10439-018-2001-2.

Meshkat, N., Sullivant, S. and Eisenberg, M. (2015) 'Identifiability Results for Several Classes of Linear Compartment Models', *Bulletin of Mathematical Biology*. Springer US, 77(8), pp. 1620–1651. doi: 10.1007/s11538-015-0098-0.

Miettinen, M. *et al.* (1999) 'Epithelioid Sarcoma : An Immunohistochemical Analysis of 112 Classical and Variant Cases and a Discussion of the Differential Diagnosis', pp. 934–942.

Mojtahedi, M. *et al.* (2016) 'Cell Fate Decision as High-Dimensional Critical State Transition', *PLoS Biology*, pp. 1–28. doi: 10.1371/journal.pbio.2000640.

Montagna, E. *et al.* (2014) 'Metronomic therapy and breast cancer: A systematic review', *Cancer Treatment Reviews*. Elsevier Ltd, 40(8), pp. 942–950. doi: 10.1016/j.ctrv.2014.06.002.

Mumenthaler, S. M. *et al.* (2013) 'The Need for Integrative Computational Oncology: An Illustrated Example through MMP-Mediated Tissue Degradation', *Frontiers in Oncology*, 3(July), pp. 9–12. doi: 10.3389/fonc.2013.00194.

Mumenthaler, S. M. *et al.* (2015) 'The Impact of Microenvironmental Heterogeneity on the Evolution of Drug Resistance in Cancer Cells', 14, pp. 19–31. doi: 10.4137/CIN.S19338.Received.

Neufeld, Z. *et al.* (2017) 'The role of Allee effect in modelling post resection recurrence of glioblastoma', *PLoS Computational Biology*, pp. 1–14.

Norton, L. (1988) 'A Gompertzian Model of Human Breast Cancer Growth', *Cancer Research*,

pp. 7067–7071.

Nowak, M. A. (2006) *Evolutionary Dynamics*. Cambridge, MA & London, England: The Belknap Press of Harvard University.

Pacheco, E. (2016) 'A review of models for cancer chemotherapy based on Optimal Control', *INESC-ID Technical report*, pp. 1–30.

Pan, J. and Johnson, K. (2019) '"Hacking" Our Way across Interdisciplinary Boundaries', *Cell Systems*. Elsevier Inc., 8(5), pp. 361–362. doi: 10.1016/j.cels.2019.04.006.

Panetta, J. C. (1997) 'A mathematical model of breast and ovarian cancer treated with paclitaxel', *Mathematical Biosciences*, 146(2), pp. 89–113. doi: 10.1016/S0025-5564(97)00077-1.

Panigrahy, D. *et al.* (2012) 'Epoxyeicosanoids stimulate multiorgan metastasis and tumor dormancy escape in mice', *Journal of Clinical Investigation*, 122(1), pp. 178–191. doi: 10.1172/JCI58128.

Patel, A. P. (2014) 'Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma', *Science*, 1396(June), pp. 1–10. doi: 10.1126/science.1254257.

Paudel, B. B. *et al.* (2018) 'A Nonquiescent "' Idling '" Population State in Drug-Treated , BRAF -Mutated Melanoma', *Biophysical Journal*. Biophysical Society, 114(6), pp. 1499–1511. doi: 10.1016/j.bpj.2018.01.016.

Pisco, A. O. *et al.* (2013) 'Non-Darwinian dynamics in therapy-induced cancer drug resistance', *Nature Communications*, 4(2467). doi: 10.1038/ncomms3467.Non-Darwinian.

Pisco, A. O. and Huang, S. (2015) 'Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse : " What does not kill me strengthens me "', *British Journal of Cancer*. Nature Publishing Group, 112(11), pp. 1725–1732. doi: 10.1038/bjc.2015.146.

Poleszczuk, J. *et al.* (2016) 'Cancer Stem Cell Plasticity as Tumor Growth Promoter and Catalyst of Population Collapse', *Stem Cells International*, 2016, pp. 1–12. doi: 10.1155/2016/3923527.

Poleszczuk, J. and Enderling, H. (2018) 'The Optimal Radiation Dose to Induce Robust Systemic Anti-Tumor Immunity', *International Journal of Molecular Sciences*, 19(11). doi: 10.3390/ijms19113377.

Poplawski, N. J. *et al.* (2010) 'Front Instabilities and Invasiveness of Simulated 3D Avascular Tumors', *PLOS ONE*. Public Library of Science, 5(5), p. e10641. Available at: https://doi.org/10.1371/journal.pone.0010641.

Prokopiou, S. *et al.* (2015) 'A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation', *Radiation Oncology*. Radiation Oncology, pp. 1–8. doi: 10.1186/s13014-015-0465-x.

Prosdocimi, F. *et al.* (2002) 'Bioinformatica: Manual do usuario', *Biotech. Cienc. Des.*, pp. 12–25.

Pyne, S. *et al.* (2009) 'Automated high-dimensional flow cytometric data analysis', *Proceedings of the National Academy of Sciences*, 106(21), pp. 8519–8524. Available at: www.pnas.org/cgi/doi/10.1073/pnas.0903028106.

Raftery, A. (1999) 'Bayes Factors and BIC', *Sociological Methods & Research*, 27(3), pp. 411–

427.

Ramis-Conde, I. *et al.* (2008) 'Modeling the influence of the E-cadherin-beta-catenin pathway in cancer cell invasion: a multiscale approach', *Biophysical journal*. 2008/03/13. The Biophysical Society, 95(1), pp. 155–165. doi: 10.1529/biophysj.107.114678.

Raue, A. *et al.* (2009) 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood', *Bioinformatics*, 25(15), pp. 1923–1929. doi: 10.1093/bioinformatics/btp358.

Ren, H. *et al.* (2016) 'TWIST1 and BMI1 in Cancer Metastasis and Chemoresistance', *Journal of Cancer*, 7(9), pp. 1074–1080. doi: 10.7150/jca.14031.

Rivera, E. and Gomez, H. (2010) 'Chemotherapy resistance in metastatic breast cancer: the evolving role of ixabepilone', *Breast Cancer Research*, 12(S2), p. S2. doi: 10.1186/bcr2573.

Robert, C. P. *et al.* (2011) 'Lack of confidence in approximate Bayesian computation model choice', *Proceedings of the National Academy of Sciences*, 108(37). doi: 10.1073/pnas.1102900108.

Rocha, H. L. *et al.* (2018) 'A HYBRID THREE-SCALE MODEL OF TUMOR GROWTH', *Mathematical models & methods in applied sciences : M3AS*. 2017/11/24, 28(1), pp. 61–93. doi: 10.1142/S0218202518500021.

Rockne, R. C. *et al.* (2019) 'The 2019 mathematical oncology roadmap The 2019 mathematical oncology roadmap'. IOP Publishing.

Rodriguez-brenes, I. A., Komarova, N. L. and Wodarz, D. (2013) 'Tumor growth dynamics: insights into evolutionary processes', *Trends in Ecology & Evolution*. Elsevier Ltd, 28(10), pp. 597–604. doi: 10.1016/j.tree.2013.05.020.

Rohrs, J. A., Makaryan, S. Z. and Finley, S. D. (2018) 'Constructing Predictive Cancer Systems Biology Models', pp. 10–12.

Saunders, N. A. *et al.* (2012) 'Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives', *EMBO molecular medicine*. 2012/06/25. WILEY-VCH Verlag, 4(8), pp. 675–684. doi: 10.1002/emmm.201101131.

Scheel, C. *et al.* (2011) 'Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast', *Cell*. Elsevier Inc., 145(6), pp. 926–940. doi: 10.1016/j.cell.2011.04.029.

Sewalt, L. *et al.* (2016) 'Influences of Allee effects in the spreading of malignant tumours', *Journal of Theoretical Biology*. Elsevier, 394, pp. 77–92. doi: 10.1016/j.jtbi.2015.12.024.

Shajahan-Haq, A. N., Cheema, M. S. and Clarke, R. (2015) 'Application of metabolomics in drug resistant breast cancer research', *Metabolites*. MDPI, 5(1), pp. 100–118. doi: 10.3390/metabo5010100.

Shin, V. Y. *et al.* (2019) 'Long non-coding RNA NEAT1 confers oncogenic role in triple-negative breast cancer through modulating chemoresistance and cancer stemness', *Cell Death and Disease*. Springer US, 10(4). doi: 10.1038/s41419-019-1513-5.

Shipitsin, M. *et al.* (2007) 'Molecular Definition of Breast Tumor Heterogeneity', *Cancer Cell*, 11(March), pp. 259–273. doi: 10.1016/j.ccr.2007.01.013.

Silva, A. S. and Gatenby, R. A. (2010) 'A theoretical quantitative model for evolution of cancer chemotherapy resistance', *Biology Direct*, 5(1), p. 25. doi: 10.1186/1745-6150-5-25.

Skipper, H. E. (1964) 'Perspectives in Cancer Chemotherapy: Therapeutic Design', *Cancer Research*, 24(8).

Smalley, I. *et al.* (2019) 'Leveraging transcriptional dynamics to improve BRAF inhibitor responses in melanoma', *EBioMedicine*. Elsevier B.V. doi: 10.1016/j.ebiom.2019.09.023.

Sontag, E. D. (2017) 'Dynamic compensation, parameter identifiability, and equivariances', *PLoS Computational Biology*, 13(4), pp. 1–17.

Sorace, A. G. *et al.* (2017) 'Quantitative F-FMISO-PET imaging shows reduction of hypoxia following trastuzumab in a murine model of HER2+ breast cancer', *Molecular Imaging Biology*, 19(1), pp. 130–137. doi: 10.1007/s11307-016-0994-1.Quantitative.

Speer, J. F. *et al.* (1984) 'A Stochastic Numerical Model of Breast Cancer Growth That Simulates Clinical Data', *Cancer Research*, (44), pp. 4124–4130.

Spiteri, I. *et al.* (2018) 'Evolutionary dynamics of residual disease in human glioblastoma', *Oxford University Press*.

Stukalin, E. B. *et al.* (2013) 'Age-dependent stochastic models for understanding population fluctuations in continuously cultured cells', *Journal of the Royal Society Interface*, 10(85). doi: 10.1098/rsif.2013.0325.

Stumpf, P. S. *et al.* (2017) 'Stem Cell Differentiation as a Non-Markov Stochastic Process', *Cell Systems*, 5(3), pp. 268-282.e7. doi: 10.1016/j.cels.2017.08.009.

Sun, S. (2015) 'Stochastic Models for Population Dynamics', *bioRxiv*, p. 31237. doi: 10.1101/031237.

Sun, S. X. (2015) 'Stochastic Models for Population Dynamics', *bioRxiv*. Available at: doi: http://dx.doi.org/10.1101/031237.

Suvà, M. L. and Tirosh, I. (2019) 'Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges', *Molecular Cell*, 75(1), pp. 7–12. doi: 10.1016/j.molcel.2019.05.003.

Syed, A. K. *et al.* (2019) 'Characterizing Trastuzumab- Induced Alterations in Intratumoral Heterogeneity with Quantitative Imaging and Immunohistochemistry in HER2 + Breast Cancer', *Neoplasia*. The Authors, 21(1), pp. 17–29. doi: 10.1016/j.neo.2018.10.008.

Thanos, D. *et al.* (2018) 'Stochastic Phenotype Switching Leads to Intratumor Heterogeneity in Human Liver Cancer', *Hematology*, 68(3), pp. 1–16. doi: 10.1002/hep.29679.

Thermo Fisher Scientific (no date) *Useful Numbers for Cell Culture*. Available at: https://www.thermofisher.com/us/en/home/references/gibco-cell-culture-basics/cell-culture-protocols/cell-culture-useful-numbers.html (Accessed: 11 February 2020).

Tian, X., Zhang, H. and Xing, J. (2013) 'Coupled Reversible and Irreversible Bistable Switches Underlying TGF b -induced Epithelial to Mesenchymal Transition', *Biophysj*. Biophysical Society, 105(4), pp. 1079–1089. doi: 10.1016/j.bpj.2013.07.011.

Tirosh, I. *et al.* (2019) 'Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq', *Science*, 352(6282).

Topol, E. (2019) *Deep Medicine: How artificial intelligence can make healthcare human again*.

Edited by D. of H. Basic Books, Inc. New York, NY.

Traag, V. A., Waltman, L. and van Eck, N. J. (2019) 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific Reports*, 9(1), p. 5233. doi: 10.1038/s41598-019-41695-z.

Tunio, I. A. *et al.* (2018) 'Face Recognition System by using Eigen Value Decomposition', *International Journal of Computer Science and Network Security*, 18(5), pp. 8–12.

Turner, S. and Sherratt, J. A. (2002) 'Intercellular adhesion and cancer invasion: A discrete simulation using the extended potts model', *Journal of Theoretical Biology*, 216(1), pp. 85–100. doi: 10.1006/jtbi.2001.2522.

UCSF (2014) *QCBNet*. Available at: https://qcbnet.ucsf.edu/ (Accessed: 14 February 2020).

Vazquez, F. *et al.* (2013) 'Article PGC1 a Expression Defines a Subset of Human Melanoma Tumors with Increased Mitochondrial Capacity and Resistance to Oxidative Stress', *Cancer Cell*. Elsevier Inc., 23(3), pp. 287–301. doi: 10.1016/j.ccr.2012.11.020.

Verli, H. (2014) *Bioinformatica: da Biologia a Flexibilidade Molecular*. Edited by H. Verli. SBBq, Sao Paulo.

Vickers, P. K. *et al.* (1988) 'A Multidrug-Resistant MCF-7 Human Breast Cancer Cell Line Which Exhibits Cross-Resistance to Antiestrogens and Hormone- Independent Tumor Growth in Vivo', *Molecular Endocrinology*, pp. 886–892.

Vieira, R., Ribeiro, F. L. and Souto, A. (2015) 'Models for Allee effect based on physical principles', *Journal of Theoretical Biology*. Elsevier, 385, pp. 143–152. doi: 10.1016/j.jtbi.2015.08.018.

Vieth, B. *et al.* (2019) 'A systematic evaluation of single cell RNA-seq analysis pipelines', *Nature Communications*, 10(1), p. 4667. doi: 10.1038/s41467-019-12266-7.

Waddington, C. H. (1940) 'Organisers and Genes'. Cambridge: Cambridge University Press.

Wagenmakers, E. and Farrell, S. (2004) 'AIC model selection using Akaike weights', *Psychonomic Bulletin & Review*, 11(1), pp. 192–196. doi: 10.3758/BF03206482.

Wang, K. *et al.* (2016) 'Hidden in the mist no more: physical force in cell biology', *Nature Methods*, 13(2), pp. 124–125. doi: 10.1038/nmeth.3744.Hidden.

Wang, Q., Cui, B. and Weng, J. (2012) 'Clinicopathological Characteristics and Outcome of Primary Sarcomatoid Carcinoma and Carcinosarcoma of the Liver', *Journal of Gastrointestinal Surgery*, 16, pp. 1715–1726. doi: 10.1007/s11605-012-1946-y.

Wang, Y. *et al.* (2019) 'iTALK : an R Package to Characterize and Illustrate Intercellular Communication', *bioRxiv*. Available at: https://dx.doi.org/10.1101/507871.

Wang, Y. and Sontag, E. D. (1989) 'On two definitions of observation spaces', *Systems & Control Letters*, 13, pp. 213–218. doi: 10.1080/02331888508801846.

Wangsa, D. *et al.* (2018) 'The evolution of single cell-derived colorectal cancer cell lines is dominated by the continued selection of tumor-specific genomic imbalances, despite random chromosomal instability', *Carcinogenesis*, (June), pp. 1–13. doi: 10.1093/carcin/bgy068.

Wei, S. C. *et al.* (2015) 'Matrix stiffness drives epithelial-mesenchymal transition and tumour metastasis through a TWIST1-G3BP2 mechanotransduction pathway', *Nature Cell*

Biology, 17(5), pp. 678–688. doi: 10.1038/ncb3157.

West, J. *et al.* (2016) 'An evolutionary model of tumor cell kinetics and the emergence of molecular heterogeneity driving Gompertzian growth', *SIAM Rev. Soc. Ind. Appl. Math*, 58(4), pp. 716–736. doi: 10.1137/15M1044825.AN.

West, J. and Newton, P. K. (2018) 'Cellular interactions constrain tumor growth', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1804150116.

Wilkinson, D. J. (2009) 'Stochastic modelling for quantitative description of heterogeneous biological systems', *Nature Reviews Genetics*, 10(2), pp. 122–133. doi: 10.1038/nrg2509.

Winsor, C. P. (1932) 'The Gompertz curve as a growth curve', *Proceedings of the National Academy of Sciences*, 18(1).

Wittmann, M., Gabriel, W. and Metzler, D. (2014) 'Genetic Diversity in Introduced Populations with an Allee Effect', *Genetics Society of America*, 198(September), pp. 299–310. doi: 10.1534/genetics.114.167551.

Wolf, F. A., Angerer, P. and Theis, F. J. (2018) 'SCANPY: large-scale single-cell gene expression data analysis', *Genome Biology*, 19(1), p. 15. doi: 10.1186/s13059-017-1382-0.

Wood, K. *et al.* (2012) 'Mechanism-independent method for predicting response to multidrug combinations in bacteria', *Proceedings of the National Academy of Sciences*, 109(30), pp. 12254–12259. doi: 10.1073/pnas.1201281109.

Wooten, D. J. and Quaranta, V. (2017) 'Mathematical Models of Cell Phenotype Regulation and Reprogramming: Make Cancer Cells Sensitive Again!', *BBA - Reviews on Cancer*. Elsevier B.V. doi: 10.1016/j.bbcan.2017.04.001.

Yankeelov, T. E. *et al.* (2013) 'Clinically Relevant Modeling of Tumor Growth and Treatment Response', 5(187), pp. 1–6.

Yankeelov, T. E. *et al.* (2015) 'Toward a science of tumor forecasting for clinical oncology', *Cancer Research*, 75(6), pp. 918–923. doi: 10.1158/0008-5472.CAN-14-2233.

Yankeelov, T. E. *et al.* (2016) 'Multi-scale Modeling in Clinical Oncology: Opportunities and Barriers to Success', *Ann Biomed Eng*, 44(9), pp. 2626–2641. doi: 10.1007/s10439-016-1691-6.

Ye, X. and Weinberg, R. A. (2015) 'Epithelial – Mesenchymal Plasticity : A Central Regulator of Cancer Progression', *Trends in Cell Biology*. Elsevier Ltd, 25(11), pp. 675–686. doi: 10.1016/j.tcb.2015.07.012.

Yu, Y. *et al.* (2015) 'Down-regulation of miR-129-5p via the Twist1-Snail feedback loop stimulates the epithelial-mesenchymal transition and is associated with poor prognosis in breast cancer', *Oncotarget*, 6(33).

Zhang, L. *et al.* (2009) 'Multiscale agent-based cancer modeling', *Journal of Mathematical Biology*, 58(4–5), pp. 545–559. doi: 10.1007/s00285-008-0211-1.

Zhang, Y. *et al.* (2019) 'Designing combination therapies with modeling chaperoned machine learning', *PLoS Computational Biology*, 15(9), pp. 1–17. doi: 10.1371/journal.pcbi.1007158.

Zhao, X. *et al.* (2019) 'Evaluation of single-cell classifiers for single-cell RNA sequencing data

sets', 00(July), pp. 1–15. doi: 10.1093/bib/bbz096.
Zhou, J. X. *et al.* (2014) 'Nonequilibrium population dynamics of phenotype conversion of cancer cells', *PLoS ONE*, 9(12), pp. 1–19. doi: 10.1371/journal.pone.0110714.
Zhou, J. X. *et al.* (2017) 'Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes', *Scientific Reports*, 7(8815), pp. 1–15. doi: 10.1038/s41598-017-09307-w.