Title

Mathematical tools for dissecting the heterogeneity in and cell cycle contributions of cancer therapy

Author

Mohammadi, Farnaz

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Mathematical tools for dissecting the heterogeneity in and cell cycle contributions of cancer therapy

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioengineering

by

Farnaz Mohammadi

2023

ABSTRACT OF THE DISSERTATION

Mathematical tools for dissecting the heterogeneity in and cell cycle contributions of cancer therapy

by

Farnaz Mohammadi

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2023

Professor Aaron S. Meyer, Chair

Cancer remains a formidable public health challenge, and identifying effective therapeutic strategies to prevent tumor cell proliferation is paramount to improving patient outcomes. Tumor cells exhibit remarkable phenotypic plasticity, enabling them to assume a diverse range of molecular and phenotypic states, and rapidly develop resistance to therapeutic or environmental stressors. This plasticity, however, presents unique opportunities to identify molecular programs that can be targeted for therapeutic purposes. Therefore, gaining a comprehensive understanding of how clinically relevant anti-cancer agents modulate cell cycle progression is pivotal to uncovering such strategies.

In this thesis, we present a suite of computational models that shed light on how drugs modulate the cell cycle, how quantifying drug effects on the cell cycle can inform drug combination recommendations, and how to analyze the heterogeneous response of single cells to cancer therapy. Specifically, Chapter 1 introduces a mathematical model that captures drug-induced dynamical responses, quantified cell cycle phase arrest, and cell death induction rates in cancer cells upon treatment using live-cell microscopy experiments. Leveraging this model, we predict drug combination effects and identify combination treatment strategies that can optimize therapeutic response in cancer, while accounting for specified cell cycle

effects. In Chapter 2, we expand the application of this modeling strategy by exploiting a newly introduced simplified experimental assay with fixed cell imaging, thereby broadening the scope of experimental data used for predicting drug combinations with our approach. This chapter also highlights the utility of a mathematical tool to discern general biological patterns within large-scale multi-dimensional data.

Finally, in the last chapter, we provide a computational approach to account for phenotypic heterogeneity in drug response observed at the single cell level. We develop a tree-based hidden Markov model that quantifies various drug-induced phenotypic cell states and transition rates between these states resulting from drug-induced cell cycle effects. This approach has potential for uncovering the relationship between molecular states and cellular phenotypes using end-point spatial transcriptomic profiles of cells under treatment.

In summary, this work presents a compelling case for how computational models can aid in understanding the effects of anti-cancer agents on the cell cycle and identifying optimal drug combinations. The models presented in this thesis provide an important foundation for further investigations into developing effective therapeutic strategies for cancer treatment.

The dissertation of Farnaz Mohammadi is approved.

Dino Di Carlo

Roy Wollman

Stephanie Kristin Seidlits

Willy Hugo

Aaron S. Meyer, Committee Chair

University of California, Los Angeles

2023

This dissertation is dedicated to my mother, Saba

Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 2:

Chapter 3:

Acknowledgements

I would like to express my deepest appreciation and gratitude to my supervisor, Dr. Aaron Meyer, for their exceptional guidance and unwavering support throughout my Ph.D. journey at UCLA. I am incredibly grateful for the invaluable lessons, patience, and kindness extended to me, which have helped me develop into the person I am today. Dr. Meyer's mentorship was instrumental in navigating the challenges and ups and downs of research and graduate school life. I am humbled by how much I have grown and learned over the years, and I cannot overstate how much I owe Dr. Meyer for their instrumental role in shaping my professional life. I could not have asked for a better mentor than Dr. Meyer.

I would also like to extend my sincere thanks to the members of my committee, Dr. Roy Wollman, Dr. Stephanie Seidlits, Dr. Dino Di Carlo, and Dr. Willy Hugo, for their insightful feedback and continued support.

I would like to extend my sincere gratitude to Dr. Laura Heiser who has been a mentor to me during the past five years with our shared projects. I would like to thank Dr. Marc Hafner who was my supervisor during my summer internship at Genentech who then later became our collaborator on continuing the internship project. I learned so much working with Dr. Hafner and I am very grateful to him for helping me grow as a scientist.

Furthermore, I would like to express my gratitude to the UCLA Graduate Division for their financial support through the Dissertation Year Fellowship during the last school year, and to the Department of Bioengineering for their support through departmental fellowship programs, and most importantly, the Meyer lab that supported me in the past 5 years.

I would also like to acknowledge the support and encouragement I received from my labmates, all members of the Meyer lab, and friends.

Lastly, I would like to extend my heartfelt appreciation to my family, Ahoora and Saba, who have always been my rock and a constant source of inspiration and motivation, eventhough I was far from home.

I would like to thank Lili Bulhoes and Daphne-Jane Dizon from the department of Bioengineering who helped me through many processes during the past years.

Last but not least, I am grateful for Jackie, my dog, who kept me company and kept me sane during the past two years!

Vita

Education

University of California, Los Angeles (UCLA) 2018-2020

M.Sc. in Bioengineering

University of Tehran 2013-2018

B.Sc. in Electrical Engineering

Internships

Genentech – Oncology-Bioinformatics Summer 2022

Teaching Assistantship

Machine Learning and Data-driven Modeling in Bioengineering Winter 2022

Mentorship

Capstone Bioengineering Winter 2023

Bruins in Genomics (B.I.G.) Summer Program Summer 2020

J.C. Lagarde: undergraduate student

Aryak Rekhi: undergraduate student

Fellowships and Awards

Dissertation Year Fellowship 2022-2023

Selected Publications

Farnaz Mohammadi, John Moffat, Steffan Vartanian, Aaron S Meyer, and Marc Hafner, "Accounting for Cell Cycle Effects Identifies CDK Therapy Rational Combinations", in prep.

Farnaz Mohammadi, Shakthi Visagan, Sean M Gross, Luka Karginov, JC Lagarde, Laura

M Heiser, Aaron S Meyer, "A lineage tree-based hidden Markov model quantifies cellular heterogeneity and plasticity", *Communication Biology*, 2022

Farnaz Mohammadi*, Sean M Gross*, Crystal Sanchez-Aguila, Paulina J Zhan, Tiera A Liby, Mark A Dane, Aaron S Meyer, and Laura M Heiser, "Analysis and Modeling of Cancer Drug Responses Using Cell Cycle Phase-Specific Rate Effects", accepted in *Nature Communications*

Marc Creixell, Hyuna Kim, **Farnaz Mohammadi**, Shelly R Peyton, Aaron S Meyer, "Systems approaches to uncovering the contribution of environment-mediated drug resistance", *Current Opinion in Solid State and Materials Science*, 2022

Hyuna Kim, Anna Wirasaputra, **Farnaz Mohammadi**, Aritra Nath Kundu, Jennifer AE Esteves, Laura M Heiser, Aaron S Meyer, Shelly R Peyton, "Live Cell Lineage Tracing of Dormant Cancer Cells", *Advanced Healthcare Materials*, 2022

Chapter 1

Analysis and Modeling of Cancer Drug Responses Using Cell Cycle Phase-Specific Rate Effects

Farnaz Mohammadi*, Sean Gross*, Crystal Sanchez-Aguila, Paulina J Zhan, Tiera A Liby, Mark A Dane, Aaron S. Meyer, Laura M. Heiser

Abstract

Identifying effective therapeutic treatment strategies is a major challenge to improving outcomes for patients with breast cancer. To gain a comprehensive understanding of how clinically relevant anti-cancer agents modulate cell cycle progression, we used genetically engineered breast cancer cell lines to track drug-induced changes in cell number and cell cycle phase, which revealed drug-specific cell cycle effects that vary across time. We developed a linear chain trick (LCT) computational model, where the cell cycle is partitioned into subphases that can faithfully capture drug-induced dynamic responses. The model correctly infers drug effects and also localizes them to specific cell cycle phases. We used our LCT model to predict the effect of unseen drug combinations and experimentally confirmed the effectiveness of predicted combination treatment strategies. Our integrated experimental and modeling approach opens avenues to assess drug responses, predict effective drug combinations, and identify optimal drug sequencing strategies.

Introduction

Developing transformative anti-cancer therapies requires drug combinations [1], however rational identification of effective combination therapy regimens remains challenging [2–5]. Many anti-cancer agents are designed to impact cell proliferation and viability, which suggests that incorporating information about how individual drugs impact cell cycle behavior can lead to improved predictions about drug combination effects. The mammalian cell cycle is typically separated into four linked phases (G_1, S, G_2, and M) with multiple checkpoints (restriction point, DNA damage checkpoint, and the spindle assembly checkpoint) [6–9], each relying on distinct molecular programs and resulting in minimal correlation between cell cycle phase durations in individual cells [10]. This independence between phases and checkpoints has implications for cancer treatment because many cancer drugs directly target different aspects of the cell cycle; for example, CDK4/6 inhibitors block progression out of G_1 phase [11], while

the nucleoside analog gemcitabine activates the DNA damage checkpoint by targeting DNA synthesis during S-phase [12]. Together, these findings imply that drug-induced changes to cell numbers can be achieved through distinct cell cycle-dependent molecular mechanisms. For example, these observations suggest that combing two drugs that each reduce the rate of G1 progression will lead to deeper reductions in the rate of G_1 progression, rather than an increase in cell death. Further, this framework predicts dose-dependent impacts: at sub-saturating doses, G_1 effects will add together to reduce cell numbers while at higher saturating doses the cell number will peak at the maximum cytostatic effect. This general idea of drug combination efficacy was recently explored in a study of the multi-drug CHOP protocol , which showed that the effectiveness of this drug combination for treatment of non-Hodgkin Lymphoma could be attributed to the fact that each agent had non-overlapping cytotoxic effects [13]. The effectiveness of the CHOP protocol also demonstrates the benefit of drug combinations to improve patient outcomes. Considering both cell cycle and cell death effects in greater detail, therefore, has the potential to significantly improve drug combination predictions.

The classic approach to quantifying drug response assumes that cells are undergoing exponential growth at the time of drug treatment and then calculates the number of cells 72 hours after drug addition [14–18]. Other approaches to quantify drug response include compartmental models such as pharmacokinetic and pharmacodynamic (PK-PD) models that consider drug uptake and population dynamics [19]. Recent advances in methodological and quantitative approaches enable assessment of the impact of therapies on cell growth rates, rather than static cell counts20, which yields more robust correlations between molecular features and drug sensitivity [20, 21]. However, while growth rate approaches significantly improve quantification, they provide limited information about cell cycle effects. A related approach, fractional proliferation, which models the number of cycling, quiescent, and dying cells in a drug-treated population, incorporates growth rates and assumes that cells irreversibly exit the cell cycle into quiescence [22]. Recent studies demonstrate that cells

may not irreversibly exit the cell cycle and instead may extend the duration of a specific cell cycle phase before restarting progression through the cell cycle [23]. These prior findings motivate our interest to deeply assess the influence of drugs on specific cell cycle phases and progression through the cell cycle.

In this work, we quantify and incorporate cell cycle phase effects in an analysis of drug responses. We use live-cell imaging of a panel of molecularly diverse breast cancer cells engineered to express a cell cycle reporter and tracked the dynamics of cell number and cell cycle phase in response to single drugs and drug combinations. Across single drugs, we observed distinct cell cycle effects that lead to similar final cell numbers, with phase-specific responses that are oscillatory over time due to the temporal impacts on the cell cycle. To describe these responses, we developed a computational model that uses a linear chain trick (LCT) to account for the delay from cell cycle phase transit time upon drug treatment. The LCT model correctly infers single drug responses across time as well as the drug-induced oscillatory cell cycle dynamics. We used this model to predict the effect of unseen combinations of drugs that impact different aspects of the cell cycle. Experimentally testing several drug combinations validates that responses were primarily determined by the specific cell cycle effects of each drug pair. These studies reveal the complexity of cell behavior underlying drug responses, provide mechanistic insights into how individual drugs modulate cell numbers, and yield a framework to rationally model and predict drug combinations.

Results

1. Drug treatments induce distinct changes in cell number and cell cycle phasing

To track drug responses in individual cells, we genetically engineered HER2+ AU565 breast cancer cells to stably express the HDHB cell cycle reporter [23] and a nuclear-localized red fluorescent protein (**Figure 1A,B**). Cells were treated with escalating doses of five

clinically relevant breast cancer drugs, each targeting different cell cycle phases or apoptotic mechanisms (**Figure 1C**). Cells were imaged every 30 minutes for 96H and the number of cells in each cell cycle phase and total cell numbers were quantified. Each drug effectively reduced cell numbers in a dose-dependent manner (**Figure 1D, S1**). As expected, paclitaxel, gemcitabine, and doxorubicin led to cytotoxic effects indicated by the final cell numbers dropping below the starting cell numbers (**Figure 1D**) [24, 25]. In contrast, at the highest doses of palbociclib and lapatinib, final cell numbers were approximately equal to the starting cell numbers, suggesting cytostatic effects. For each drug, the pattern of cell counts varied across time; at high doses responses tended to reach a peak and then decline as the duration of drug exposure increased—an effect most marked for 30 nM gemcitabine where the relative cell number declined from 1.1 at 48H to 0.5 at 96H (**Figure 1D**) [21,26]. The fraction of S-G2 cells varied over time and showed both drug- and dose-specific effects (**Figure 1E,F**). For example, lapatinib and palbociclib initially reduced the fraction of cells in $S-G_2$ phase in a dose-dependent manner, whereas gemcitabine and doxorubicin initially increased this fraction. Of note, intermediate doses of lapatinib (50 nM) and paclitaxel (3 nM) induced oscillating cell cycle responses, with an initial $S-G_2$ reduction near 30H, followed by a second $S-G_2$ reduction at 84H. In sum, this approach revealed drug-specific cell cycle changes across time, which confirms that these drugs yield similar final numbers through distinct impacts on the cell cycle.

2. A dynamical model captures drug-induced changes to cell cycle behavior

A common approach to model drug effects assumes exponential growth that varies as a function of drug dose [27]. This approach, although informative, cannot explain the cell cycle dynamics described above and motivated the development of a dynamical model to capture the observed behavior. As an initial model, we defined a system of ordinary differential equations (ODEs) with transitions between G_1 and $S-G_2$. The parameters of the ODE

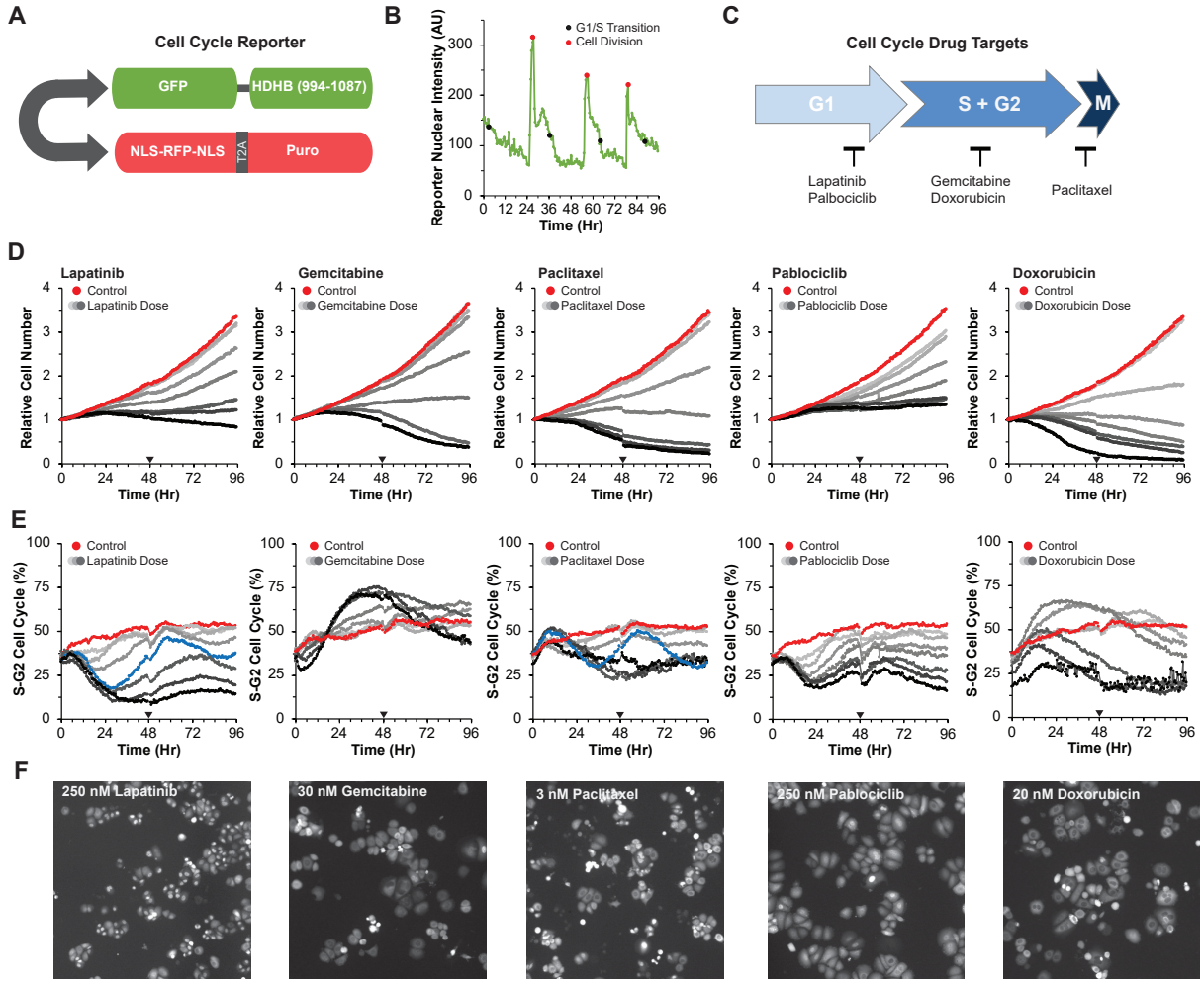


Figure 1: Drugs induce dose- and time-dependent changes in cell cycle behavior. **A.** Schematic of reporter with a bidirectional promoter driving expression of human DNA Helicase B (HDHB) fused to the green fluorescent protein clover, and a second transcript coding for NLS-RFP-NLS, a ribosome skipping domain (T2A), and a puromycin resistance protein. **B.** Quantification of nuclear intensity of the cell cycle reporter in a cell and its progeny across time. The time of G1-S transition and cell division are demarcated with black and red circles respectively. **C.** Schematic of the five drugs tested and the cell cycle phase they target. **D.** Average growth curves of AU565 cells tracked every 30 min for 96H across an 8-point dose response for lapatinib, gemcitabine, paclitaxel, palbociclib, and doxorubicin. The null dose is colored red. Line traces show the average from three independent experiments. The black triangle indicates the addition of fresh drug and media. **E.** Percentage of cells in S-G2 phase of the cell cycle across doses. 50 nM Lapatinib and 3 nM paclitaxel are colored blue. **F.** Representative GFP images at 39.5H for 250 nM lapatinib, 30 nM gemcitabine, 3 nM paclitaxel, 250 nM palbociclib, 20 nM doxorubicin from data plotted in D.

model were the cell cycle phase progression and death rates, which were assumed to follow a Hill function with respect to drug concentration. This model failed to fit the experimental data of G_1 and S-G2 cell numbers (**Figure S2**); furthermore, dynamical systems theory dictates that this model is unable to oscillate under any reasonable parameterization [28].

To address these limitations and capture the observed oscillatory temporal dynamics, we incorporated into the model the observations that phase durations follow a gamma distribution and are uncorrelated [10] (**Figure S3A**). Gamma and related distributions model each cell cycle phase as a series of steps, with the key feature that they can model processes wherein there is always some measurable duration before a system (e.g., a cell progressing through the cell cycle) can move to the next state.

The number of steps in each phase were determined by estimating the shape parameter of the gamma distributions fitted to single cell measurements of G_1 and $S-G_2$ phase durations measured in the untreated control condition [29]. This resulted in partitioning the G1 phase into 8 and $S-G_2$ phase into 20 steps (**Figure 2A**). We incorporated a linear chain trick into our model, which creates similarly-distributed time delays in the cell cycle phase durations through a mean-field system of ODEs [30]. The model was further simplified by sharing parameters that were not drug specific, such as the number of cell cycle subphases and the initial fraction of cells in G_1 phase. We then fit all five drug dose responses, varying the drug-specific and shared parameters, simultaneously. Incorporation of this component enabled the model to capture the experimentally observed oscillatory cell cycle behavior and cell cycle phase-specific drug effects. We computed the fitting error of the two modeling frameworks by calculating the sum of squared error of the difference between the data and model predictions across all concentrations and observed that the LCT model had lower error terms (**Figure 2B**). The fits to lapatinib and palbociclib were particularly improved by the model refinement. Examples of dose-response curves and model fits for lapatinib and gemcitabine are shown in Figure 2C-H. Importantly, the model captured the dose-dependent changes to G_1 and $S-G_2$ populations as well as the oscillatory dynamics. Estimating the cell

cycle phase progression and death rates also enabled calculation of the accumulated amount of cell death across time using inferred cell counts at each phase (**Figure 2E,H**). The LCT model also performed well for each of the other drugs (**Figure S4A-F**).

The phase durations and cell death probabilities inferred from the LCT model varied with drug treatment (**Figure 2I-L**). Comparison of inferred effects at the half maximum concentration (EC50) revealed that lapatinib and palbociclib treatments lead to longer average G_1 phase durations compared to untreated cells (**Figure 2I-J**), a 10% higher chance of cell death in G_1 phase for lapatinib treated cells, and a slight chance of cell death in $S-G_2$ after palbociclib treatment (**Figure 2K-L**). The model also inferred that gemcitabine induces an increase in $S-G_2$ durations and greater chance of cell death in $S-G_2$ phase as compared to untreated cells (**Figure 2G-H**). Finally, a 10% chance of cell death at the EC_{50} concentration (2.4 nM) was inferred in late G_2 phase for cells treated with paclitaxel as compared to untreated controls (**Figure 2J** and **Figure S3J**).

3. Analysis of single cell responses confirms model inferences and reveals drug-specific cell cycle phase effects

We developed model parameters from the average population response at each timepoint, which facilitates robust model development by leveraging information from a large number of cells. Importantly, as described above, the LCT model infers aspects of drug responses that can be quantified at the individual cell level—including cell cycle phase duration and cell cycle-specific death. We therefore tracked single cells in the image time course data to quantify cell cycle phase durations and also cell death events associated with specific drug treatments and concentrations (**Figure S3B**). Quantification of cell death events also enables direct assessment of whether drug effects are cytotoxic or cytostatic. The first complete cell cycle was analyzed to examine early drug effects. We also quantified the relative fate outcomes for the progeny of cells observed at time 0H (relative to drug addition) that later underwent division, which provides insights into drug treatment effects observed at later

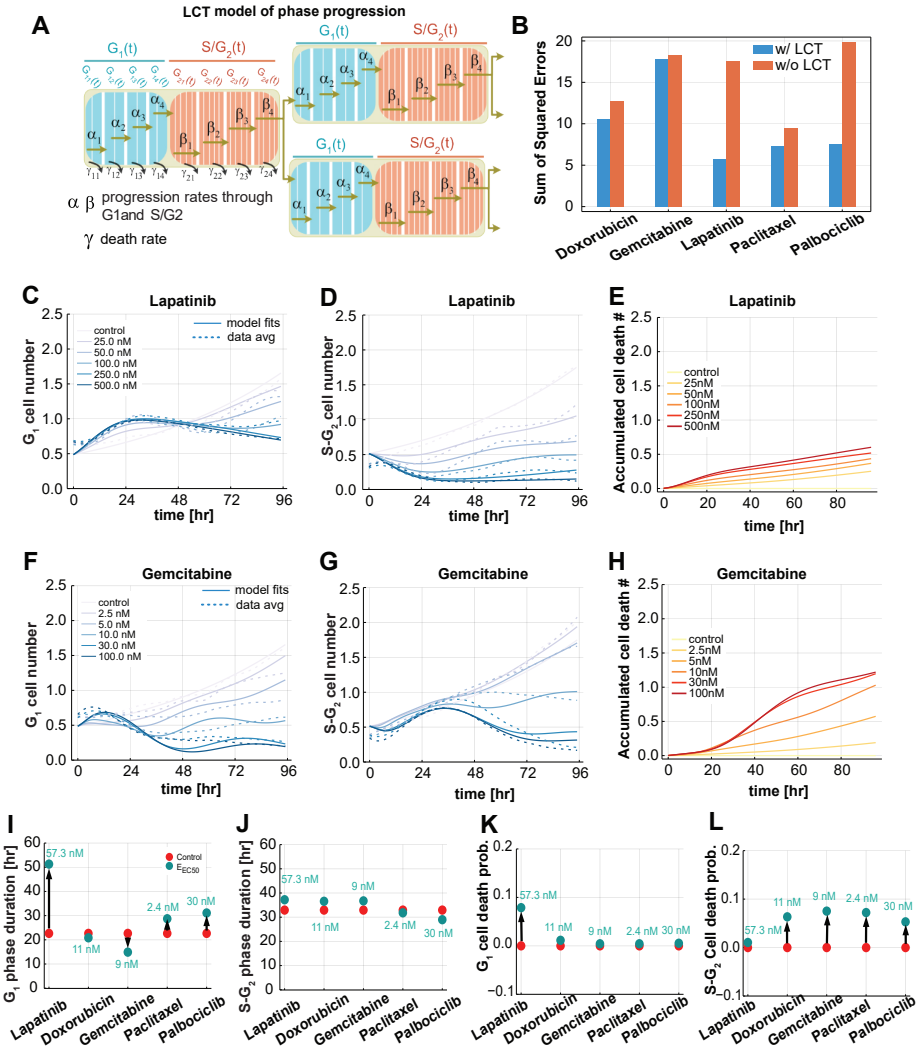


Figure 2: A computational model of the cell cycle captures the dynamics of drug response. A. Diagram of the phase transitions in the linear chain trick (LCT) model. α_1,..., α_4 are the progression rates through G_1 phase; β_1,..., β_4 are the progression rates through $S - G_2$ phase. Similarly, γ_{11},..., γ_{14} are the death rates within the G_1 phase parts, and γ_{21},..., γ_{24} are death rates within $S - G_2$ phase parts. **B.** The sum of squared errors for the fits of each of five drugs over all concentrations with (blue) and without (orange) the LCT modification. **C-H.** G_1 (**C, F**) and $S - G_2$ (**D, G**) cell numbers over time, respectively, for lapatinib and gemcitabine treated cells at 5 concentrations and untreated control (solid lines), respectively, overlayed with the average of three experimental replicates (dashed lines). The predicted accumulated dead cells over time for lapatinib (**E**) and gemcitabine (**H**). **I-J.** The average phase durations in G_1 and $S - G_2$ phases for all five drug treatments. **K-L.** The overall probability of cell death in G_1 and $S - G_2$ phase, respectively, for all five drug treatments. The arrow shows the shift from the control condition to the drug effect at the half maximum concentration ($E_{EC_{50}}$) for G_1 and $S - G_2$ phases.

timepoints (**Figure S3C**). As expected, in the untreated condition, most cells (93%) present at 0H underwent cell division. In contrast, at the highest lapatinib and gemcitabine doses, 32% and 61% of the cells present at time 0H failed to divide. Additionally, of the cells that did divide in these two conditions, only 10% underwent a second division. For both drugs, lower doses showed more modest changes in the fraction of cells that divided as compared to untreated. As described below, we compared these experimentally observed drug-induced cell cycle effects to those inferred by the LCT model.

The model inferred that the predominant lapatinib effect was to extend G_1 durations from 22.3H in the untreated condition to 33.6H and 47.4H for 25 nM and 50 nM lapatinib, respectively (**Figure 3A**). Experimentally, G_1 durations increased after lapatinib (mean 26.2H and 32.5H with 25 nM and 50 nM lapatinib, respectively) (**Figure 3A,B**). We also quantified an increase in the G_1 duration variance showing that cells varied in their responsiveness to lapatinib (**Figure 3B**). The model inferred only modest changes to $S - G_2$ durations or cell death, which was consistent with experimentally observed $S - G_2$ durations and cell death associated with lapatinib treatment (**Figure 3C,D**).

The model inferred that oscillations in the percentage of G1 cells after lapatinib treatment arise from waiting time effects in cell cycle progression (see **Figure 2**). Waiting times, which can be modeled through distributions such as the gamma distribution, refer to the delay effect created by processes that are comprised of many sequential steps. To confirm the mechanism underlying this behavior at the single cell level, we examined various cell cycle measures and found a reduction in the fraction of cells undergoing their first division beginning around 24H (**Figure 3E**). This observation, together with the lengthening of the subsequent G_1 duration following cell division (**Figure 3B**), can explain the cell cycle synchronization observed in the experimental data (see **Figure 1**) and in the LCT model. At the start of the assay, cells in G1 are delayed in their time to division, while cells in $S - G_2$ only become delayed at the onset of G_1 following division. In effect, this creates two populations of cells with distinct timing in the induction of drug effects. We observed a similar effect after treatment with

palbociclib (**Figure S3D**).

For gemcitabine, the model inferred a slight acceleration of G_1 phases, which was recapitulated experimentally (**Figure 3A,F**). The model inferred that $S-G_2$ durations were extended following gemcitabine treatment, which we confirmed experimentally: $S-G_2$ durations were extended from 22.3H to 34.5H with 5 nM and to 38H with 10 nM gemcitabine (**Figure 3G**). Lastly, the model inferred an increase in the number of cell death events relative to the starting cell number, from 0 in control to 0.57 with 5 nM gemcitabine. At 10 nM gemcitabine, the model predicted 1.0 relative cell death events such that the number of cell death events across 96H was the same as the initial starting cell number (**Figure 3A**). The experimentally observed values showed similar trends, though with more modest changes in cell numbers (0.14 and 0.41 relative cell numbers for 5 and 10 nM gemcitabine, respectively) (**Figure 3H**). Overall, we observed similar trends in each of the parameters for gemcitabine treated cells as inferred by the model; modest differences were that the model inferred higher cell death and shorter extensions to $S-G_2$ than we observed experimentally.

We also tested an assumption of the model that G_1 and $S-G_2$ phases are independent variables, which captures the idea that these cell cycle phases are independently regulated at the molecular level. We analyzed G_1 versus $S-G_2$ durations for individual cells in the untreated, 10 nM gemcitabine, and 50 nM lapatinib conditions, and found a minimal correlation between G1 and $S-G_2$ durations (**Figure 3I**). These experimental observations confirm the implicit assumption of the model that G_1 and $S-G_2$ durations are uncorrelated.

Lastly, we evaluated model inferences for paclitaxel treatment. Consistent with our experimental observations, the model inferred minimal changes to G1 and $S-G_2$ durations following treatment (**Figure 3A,J,K**). At 2 nM paclitaxel, the model inferred 0.56 cell deaths relative to the starting cell numbers, and at 3 nM inferred 0.90 relative cell deaths (**Figure 3A**, Methods). Experimentally, our observations were consistent with the values inferred by the model: we observed 0.54 and 1.00 relative cell deaths for 2 nM and 3 nM paclitaxel (**Figure 3L**). To summarize the mechanisms that account for the observed changes

in cell numbers due to paclitaxel treatment, we compared the number of cell death events against final cell counts for each of the other drugs. These data show the relative bias of paclitaxel toward inducing cell death, especially at 2 nM, as compared to 5 nM gemcitabine and 50 nM lapatinib, which both result in similar final cell numbers (**Figure 3M**). Overall, the LCT model captures key observations about the cell cycle effects of each drug, which were confirmed by in-depth single-cell tracking of the experimental data.

4. Drug-induced changes to cell cycle behavior generalize across a molecularly diverse panel of breast cancer cell lines

To assess the generalizability of our computational framework and experimental observations, we generated and tested three additional breast cancer cell lines from diverse molecular backgrounds [31]: 21MT1 (Basal subtype, HER2+), HCC1143 (Basal subtype, HER2-) and MDAMB157 (Claudin-low subtype, HER2-) (**Figures S5-7**). Because these cell lines do not uniformly overexpress HER2, we additionally tested BEZ235 and trametinib, which respectively target PI3K and MEK, two growth factor pathways downstream from HER2. We observed dose-dependent reductions in cell numbers and also modulation of the percent of G_1 cells following drug treatment. Importantly, similar to our findings for AU565 cells, we observed dynamic responses not captured by terminal endpoint readouts of cell viability (**Figures S5-7, panels A-B**). Unique response patterns observed include: a delayed G_1 enrichment from trametinib in 21MT1 cells (**Figure S6**), a lack of G1 enrichment from palbociclib and BEZ235 in MDAMB157 cells (**Figure S7**), and a dose-dependent bifurcation in G1 enrichment for doxorubicin in all three of the cell lines (**Figures S5-7**).

Next, we tested our LCT model on each of the new cell lines. Comparison of model fits to experimental observations confirmed that our model could capture the dynamic responses observed across this panel of molecularly distinct cell lines, indicating the generalizability of our computational framework (**Figures S5-7, panels C-E**). We analyzed the output of the LCT model, which inferred changes to cell cycle phase durations and cell death probabilities

A

Treatment	G1 Duration		S/G2 Duration		Cell Death	
	Predicted (Hr)	Observed (Hr)	Predicted (Hr)	Observed (Hr)	Predicted (Relative Number)	Observed (Relative Number)
Control	22.7	19.4	32.9	22.3	0.00	0.14
Lapatinib 25nM	33.6	26.2	33.2	24.0	0.25	0.06
Lapatinib 50nM	47.4	32.5	36.1	25.2	0.37	0.20
Gemcitabine 5 nM	15.1	18.0	33.3	34.5	0.57	0.14
Gemcitabine 10nM	15.1	20.5	37.8	38.0	1.00	0.41
Paclitaxel 2nM	26.7	21.8	31.4	24.6	0.56	0.54
Paclitaxel 3nM	32.0	22.9	33.7	26.9	0.90	1.00

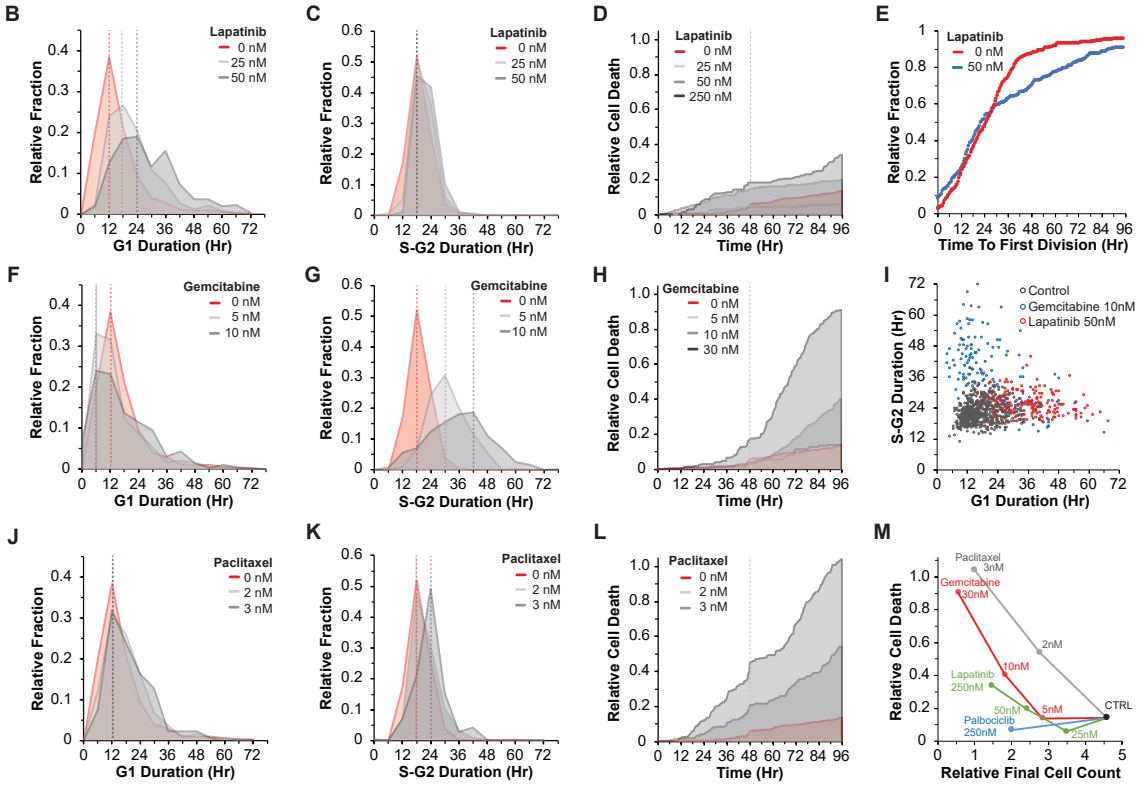


Figure 3: Analysis of single cell responses confirms model inferences and reveals drug-specific cell cycle phase effects. **A.** Quantification of cell cycle parameters as inferred by the model and observed experimentally (G1 and S-G2 durations and cell death). **B.** Distributions of G1 durations for cells that underwent one division in response to 0, 25, and 50 nM lapatinib. **C.** Distributions of S-G2 durations. **D.** Accumulated cell death across time. **E.** Time to first division for cells in the untreated condition (red line) compared to 50 nM lapatinib (gray line). **F-H.** G1 and S-G2 distributions, and cell death accumulation in response to gemcitabine. **I.** G1 and S-G2 durations for the first complete cell cycle for all cells tracked in the control condition (black dots), in response to 100 nM lapatinib (red dots), and 10 nM gemcitabine (blue dots). **J-L.** G_1 and $S-G_2$ distributions, and cell death accumulation in response to paclitaxel. **M.** Observed cell counts against cell deaths per drug.

for drug-cell line pairs at the EC50 concentration (**Figure S8**). The model inferred cell-line-specific changes to both G_1 and G_2 phases (**Figure S8A,B**). For instance, 21MT1 were inferred to preferentially undergo G_1 cell death after doxorubicin and paclitaxel treatments, at probabilities of 60% and 15%, respectively (**Figure S8C**). The model inferred that HCC1143 cells arrest and die in $S - G_2$ following paclitaxel or palbociclib treatment (**Figure S8B,D**). MDAMB157 cells were inferred to become growth-arrested by drug treatment and to preferentially die in G_1 phase (**Figure S8C,D**). Overall, we confirmed that our computational framework generalizes across several drugs and cell lines and can infer a range of drug treatment response behaviors.

5. Responses to drug combinations are dependent on drug specific cell cycle and cell death effects

Durable and effective cancer treatments frequently require administration of multiple drugs; however, identification of the principles underlying optimal drug combinations have been challenging [32]. Here, we tested the idea that our LCT model, which incorporates cell cycle effects, can be used to predict the impact of different drug combinations on cell cycle behavior and final cell numbers. We compared two strategies in accounting for drug combination effects. In the first, we combined drug effects on the rates of G_1 and $S - G_2$ progression using Bliss additivity and assumed the rates of cell death additively combined. In the second, we assumed an additive combination through use of the drug effects on overall cell numbers. To explore these predictions, we varied the dose of one drug in the two drug combination pair and analyzed responses to drug combinations that targeted either the same cell cycle phase (G_1 and G_1, or $S - G_2$ and $S - G_2$) or different cell cycle phases (G_1 and $S - G_2$).

We tested combining the rates for two G1 targeted drugs, in this case lapatinib and palbociclib. The model predicted that effects on cell number would saturate around the initial starting cell number, indicating cytostatic effects of this drug combination (**Figure 4A**). In contrast, drug combination effects based on cell numbers alone predicted a cytotoxic effect at

higher drug concentrations, resulting in a reduction in cell numbers relative to the starting cell numbers. We tested these drug combinations experimentally and found a cytostatic effect at higher doses, which matches the model prediction based on combining rates of cell cycle progression (**Figure 4A**). We also analyzed predictions of gemcitabine combined with doxorubicin, which both extend $S - G_2$ durations and induce cell death (see **Figure 3A**). Combination predictions based on rates and cell counts both predicted a reduction in cell numbers relative to each drug on its own, which we also observed experimentally (**Figure 4B**).

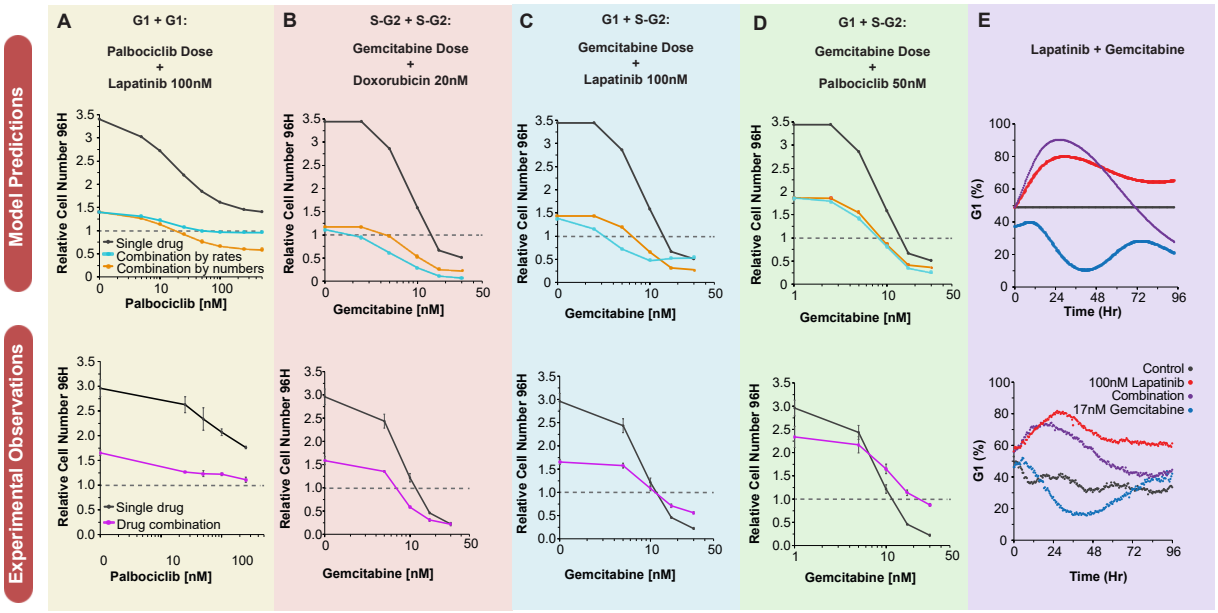


Figure 4: Responses to drug combinations are dependent on drug-specific effects on the cell cycle and cell death. A-D. Comparison between model predictions (top rows) and experiments (bottom rows). Comparing model predicted single drug responses (black), with drug combinations of Bliss additivity using cell numbers (orange) or single drug rates (cyan). Comparing single drug (black) and drug combinations (purple) from experiments. **A.** Single drug responses for increasing doses of palbociclib and in combination with 100 nM lapatinib. **B-D.** Single drug responses for increasing doses of gemcitabine and in combination with 20 nM doxorubin, 100 nM lapatinib, 50 nM palbociclib. **E.** Model predictions for the percentage of cells in G1 phase for the control condition, 100 nM lapatinib, 17 nM gemcitabine, or the combination of lapatinib and gemcitabine. **E-H.** Comparison between the model predictions and experimental observations for the single drug responses and the drug combinations as described in panels A-D. Error bars show the standard error of the mean for three biological replicates.

Finally, we used the LCT model to examine the impact of combining two drugs that target different cell cycle phases, which mimics lapatinib (G_1 effect) combined with gemcitabine (S phase effect). The cell cycle model predicted an antagonistic effect at higher doses, such that 30 nM gemcitabine combined with 100 nM lapatinib is expected to yield a similar final cell number as 30 nM gemcitabine on its own (**Figure 4C**). Experimentally, we observed that three of the four lapatinib and gemcitabine combination doses show an antagonistic impact on cell number as compared to gemcitabine alone indicating that combining these two drugs is counterproductive. These antagonistic effects of the combination held when lapatinib was replaced by palbociclib, which also impacts G1 durations (**Figure 4D**). We examined the model predictions in more detail to gain insights into the underlying biological mechanisms driving these drug combination responses. The LCT model predicted that the G_1 effect of lapatinib initially dominates over the S-phase effects of gemcitabine, leading to an increased G_1 proportion for the population. We confirmed this prediction experimentally, indicating that lapatinib co-treatment can mitigate the $S - G_2$ effects of gemcitabine (**Figure 4E**). In summary, these data indicate that the cell cycle phase and cell death impacts of each drug in a pair are critical for determining the influence of single drugs on cell cycle behavior and that this information can be used for rational identification of drug combinations likely to be therapeutically beneficial.

Discussion

In this report, we link cell cycle regulatory mechanisms with drug-specific cell cycle effects to gain insights into cancer cell responses to individual drugs and drug combinations. To meld these ideas, we developed a combined experimental and modeling approach to measure cell dynamics and infer cell behavior. This approach revealed that assessment of temporal dynamics and cell behavior is critical to interpret and model drug-induced effects. Importantly, assessment of the impacts of single agents on cell cycle behavior could be used to identify drug combinations likely to yield therapeutic benefits.

Recently, an in-depth analysis revealed that cell cycle phases in individual cells are uncorrelated and have durations that can be accurately modeled as an Erlang distribution, which is a special case of a gamma distribution [33]. This observation indicates that the cell cycle can be viewed as a series of uncoupled, memoryless phases rather than a single process [10, 34]. In this work, we found similar uncorrelated patterns in cell cycle phase responses after treatment with different anti-cancer drugs. This revealed multiple implications for assessing and modeling drug responses. First, viewing the cell cycle as a single process implies that cell behavior is immediately impacted upon drug treatment; however, we and others have reported that drug effects are often not observed until individual cells enter or approach a specific phase or checkpoint [35, 36]. For instance, we found that cells were initially distributed across all phases of the cell cycle and that the addition of lapatinib, a G_1-targeting drug, did not initially affect cells in $S - G_2$ phase. This led to a partial cell cycle synchronization across the population and required the incorporation of a linear chain trick into our model to account for this dwell time. Additionally, the temporal dynamics of the therapeutic response were an important consideration for co-treatment with gemcitabine and lapatinib. If both drugs had immediate effects on cell behavior, we would expect that the G_1 and $S - G_2$ effects of each drug would counteract each other and lead to a constant ratio of cells in G_1 phase. Instead, both experimentally and through model predictions, we found an initial G_1 enrichment. This likely induced a secondary effect of reducing the relative time that each cell spent in S-phase, which further reduced gemcitabine sensitivity. This finding could explain the antagonistic effects on cell numbers that we and others have observed when combining gemcitabine with lapatinib or palbociclib [37, 38]. We speculate that a synergistic effect on cell numbers could also arise by combining two drugs that target $S - G_2$ phases, where each drug acts to extend the relative duration in which the other is effective. This general strategy could be used to identify optimal temporal scheduling of other drug combinations that induce different effects on the cell cycle.

A second implication of multiple independently regulated cell cycle processes relates

to the concept of effect equivalence in drug combinations. This concept—that two drugs with independent targets can be used to identify drug synergy or drug antagonism—has predominantly focused on the cell number effect of each drug [2–5]. Our current work suggests that equivalence in effect may be better applied to rates of cell cycle phase progression and cell death. In our work, we found that lapatinib and palbociclib primarily impacted G_1 phase with limited effects on cell death. In contrast, doxorubicin and gemcitabine extended $S - G_2$ durations and induced cell death. These cell cycle and cell death effects were critical for gaining insights into the effect of drug combinations. For example, two cytostatic drugs, lapatinib and palbociclib, were additive up to doses that reached the maximum cytostatic effect, with further dose increases leading to only minor effects on cell numbers. In contrast, combining the two cytotoxic drugs led to increasingly cytotoxic responses across the full dose range. These results suggest that considering the cell cycle and cell death impacts of each drug is necessary to make predictions about the effects of their combinations and implies that this information could be used for the rational identification of effective drug combinations [39, 40].

Drug response measurements evaluated in the context of a mechanistic cell cycle model can reveal insights about the nature of drug response and resistance not immediately apparent from purely data-driven analyses. For instance, a model for the proliferation dynamics of cancer cells can separate the contribution of dividing, non-dividing, and dying cells [22], revealing that the rates of cell death and entry into quiescence change with drug treatment. Previous computational models of cell cycle behavior have explored various ways in which cell cycle behavior might impact drug response but have struggled to identify experimental data amenable for model fitting and evaluation. For instance, others have appreciated that drugs do not affect the cell cycle uniformly and have therefore proposed computational models that partition the cell cycle into several independent steps, both with [10] and without [34] cell death effects. Modeling cell lifetimes as being hypo-exponentially distributed helps to explain the distribution of cell lifetimes within a population but does not connect these

observations to known cell cycle stages [41]. In this study, we demonstrate that partitioning known cell cycle phases to account for their dwell time effects—and including experimentally observed drug effects like cell death—results in a modeling framework that can faithfully and mechanistically capture experimentally observed anti-cancer drug effects.

We applied our experimental approach and computational framework to examine dynamic drug-induced responses in a molecularly diverse set of breast cancer cell lines. In all cases, we observed that therapeutic inhibition induces a wide array of responses, indicating that the influence of therapies on cell cycle dynamics is a generalizable mechanism operable in a wide array of molecular backgrounds. Cancer cells treated with therapies may adopt new molecular programs associated with adaptive and acquired resistance, and indeed previous studies have demonstrated this principle in both model systems and patient samples [42]. We hypothesize that cells with acquired resistance may show distinct drug-induced cell cycle programs as compared to naïve cells and that our approach could be used to uncover the molecular mechanisms associated with adaptive resistance. The approach outlined here is built around the concept that therapies perturb cell cycle behavior and is agnostic of the exact type of cellular perturbation. Our study therefore provides a blueprint for studying responses of diverse cell types—both normal and diseased—to a wide array of perturbations, including diverse panels of therapeutic inhibitors, growth factors, or genetic manipulation with CRISPRi/a. The resultant data could be used to adapt our computational framework to identify mechanisms of cell cycle control in different cellular contexts, microenvironmental conditions, or disease states.

While our model could explain many of the key observations in our experimental data, extensions of the model could further improve its generalizability and robustness. We partitioned the cell cycle into two observed phases, G_1 and $S - G_2$, which were further subdivided to explain the dwell time behavior of each phase. With improved reporter strategies [43], we may be able to further subdivide these phases into constituent parts, which could help to localize the effect of a drug to a more specific portion of one cell cycle phase.

Generalizations of the linear chain trick could be used to account for both subphases of varying passage rates, as well as heterogeneity in the rates of passage, which would arise through cell-to-cell heterogeneity [30]. While the subdivisions within each cell cycle phase are phenomenological, it is tempting to imagine they represent mechanistic steps within each phase. Identifying how effects connect to actual biological events in the cell cycle would help identify opportunities for drug combinations. A practical challenge when using the model for drug combinations has been normalization between experiments. While cell number measurements are routinely normalized by dividing by a control, experiment-to-experiment variation in inferred rates requires additional consideration. A wider panel of experiments, across multiple cell lines, may help to tease apart variations associated with drugs, cell lines, or experiments. A final potential extension is considering the existence of phenotypically diverse subpopulations [44]. At the cost of additional complexity, one could employ several instances of the current model with transition probabilities between these states when the cells divide to simulate a heterogeneous population of cells.

Methods

Creation of Stable Cell Lines

AU565 (ATCC CRL 2351) and MDAMB157 (ATCC HTB 24) cells were grown in DMEM supplemented with 10% FBS, HCC1143 (ATCC CRL 2321) cells were grown in RPMI supplemented with 10% FBS, and 21MT1 (generous gift from Kornelia Polyak) cells were grown in DMEM/F12 supplemented with 5% horse serum, 20 ng/ml rhEGF, 0.5 µg/ml hydrocortisone, 100 ng/ml cholera toxin, and 10 µg/ml insulin. The coding fragment for clover-HDHB was cloned in frame into a transposase expression plasmid modified to also express a nuclear localized mCherry [45]. The stable cell lines were created as previously described [45] and selected for 7 days with 0.75 µg/ml puromycin. To mitigate a range of fluorescent signals from transfection, HCC1143 and 21MT1 cells were sorted at OHSU's Flow Cytometry Core and cells with a medium intensity clover-HDHB signal and a high intensity

NLS-mCherry signal were selected for drug dose response experiments. In all cases, cells were validated by STR profiling (LabCorp) and tested negative for mycoplasma.

Drug Dose Response Protocol

AU565 cells were plated at a density of 25,000 cells per well into 24-well Falcon plates (Corning #353047). 24H after plating the media was exchanged with Fluorobrite media supplemented with 10% FBS, glutamine, and penicillin-streptomycin. Cells were then treated with dose-escalation: l apatinib (Selleckchem #S1028), gemcitabine (#S1149), paclitaxel (#S1150), doxorubicin (#S1208), palbociclib (#S1116), BEZ235 (#S1009), and trametinib (#S2673). After drug addition, plates were imaged every 30 minutes for 96H using phase, GFP, and RFP imaging channels with an IncuCyte S3. For single drug treatments of AU565 cells only, at 48H the media was replaced in all wells including the control wells, and fresh media and drug were added. Four equally-spaced image locations per well and three biological replicates were collected.

MDAMB157, HCC1143, and 21MT1 cell lines were transferred to and maintained in a base of either Fluorobrite media and 1x GlutaMAX or mixed Fluorobrite/F12 media and 0.5x GlutaMAX along with their corresponding supplements for no less than one week before performing the drug dose response protocol. MDAMB157 and HCC1143 cells were plated at a density of 25,000 cells per well, while the larger 21MT1 cells were plated at a density of 5,000 cells per well into 24-well Falcon plates (Corning #353047). 24H after plating the media was exchanged with fresh Fluorobrite media as indicated per cell line. Cells were then treated with dose-escalation: BEZ235, gemcitabine, paclitaxel, doxorubicin, palbociclib, and trametinib. After drug addition, plates were imaged every 2 hours for 96H using phase, GFP, and RFP imaging channels with the IncuCyte S3. Four equally-spaced image locations per well and three biological replicates were collected.

Image Analysis

To analyze AU565 image data, phase, GFP, and RFP images were overlaid and collated into single files using FIJI [46], then segmented into three classes (nuclei, background, debris) using a manually trained classifier in Ilastik [47]. The segmented nuclear masks from Ilastik and the IncuCyte GFP images were used to count the number of nuclei in each image with Cell Profiler [48]. Additionally, using the same images (nuclear masks from Ilastik and GFP cell cycle reporter images) cell cycle phase was determined by taking the mean fluorescence in the nucleus compared to the mean fluorescence in a 5-pixel ring surrounding the nucleus, excluding background pixels. A threshold was then manually set for the ratio of nuclear fluorescence to cytoplasmic fluorescence and cells with values below the threshold were defined as being in G_1 and cells with values above the threshold were defined as being in $S - G_2$ phase [48].

To manually track AU565 cells and identify drug-induced changes operable in single cells, GFP image sequences were registered using the FIJI plug-in 'StackReg'. Individual cells present in the first image and their progeny were followed to identify the time of G1 transition, cell death, and cell division using the plug-in mTrackJ [49]. We excluded cells that were binucleated, had abnormally large nuclei, or were near the image border where complete lineages could not be tracked. The G1 transition was defined as the last frame before the nuclear intensity of the cell cycle reporter was below the level of the cytoplasm. Assessment of cell death enabled disentangling of cytostatic and cytotoxic drug effects.

We used the following approach for automated analysis of HCC1143, 21MT1 and MDAMB157 cell lines. Image registration was performed on the red channel nuclear marker image stack using the python skimage phase_cross_correlation function to correct translations. Image stacks were cropped to their common areas and individual cells were segmented with the Cellpose LC2 model trained on phase and nuclear images from the untreated and highest drug concentration treatments [50]. Nuclei were segmented with the Cellpose cyto2 model on the nuclear channel. To associate nuclei across the image stack, to identify progeny

after mitosis, and to identify cell death events we used Loeffler tracking [51] with the default parameters of delta_t = 3 and roi_size = 2. We created cytoplasm masks by subtracting the nuclear masks from the cell masks and applied these masks to the green channel cell cycle reporter images using the python skimage function regionprops_table. To assign cells to G_1 or $S-G_2$ states, we computed the ratios between the cytoplasm and nuclear cell cycle reporter. k-means clustering of the ratios observed in cells in the untreated condition was used to establish a per-plate threshold between cell cycle states.

The quantified cell-level data was mean summarized to the population level for each image and to assess cell counts and G_1 cell cycle state proportion. The cell counts were normalized to the mean of the counts of the first three images. The cell count dose response curves were normalized to the control by dividing each drug cell count by the control cell count at the same time slice.

Core Model

To identify the dynamics of the AU565 cancer cell population in response to compounds, we built a system of ordinary differential equations (ODEs) with two states: G_1, and $S-G_2$. Cells transition from G_1 to $S-G_2$ phase, and then vice versa when doubling. Cell death can occur in either phase with phase-specific death rates. S and G_2 phases are combined as our reporter cannot distinguish them. From single-cell tracking, we identified that G1 and $S-G_2$ phase time-intervals are gamma-distributed. Based on this observation, we employed the linear chain trick (LCT) [29] to capture these waiting time distributions. We broke down each phase into a series of sequential sub-phases and derived the system of mean-field ordinary differentials. Each sub-phase is represented as a single state variable within the differential equation system. The total number of cells in each phase is the sum of the cell numbers in each sub-phase. Furthermore, to account for the non-uniform effect of the drugs over each cell cycle phase, we divided G_1 and $S-G_2$ into 4 parts each, such that the effect of a drug can be distinguishable at the beginning, middle, or the end of the phases.

The mean-field system of ODEs is:

$$\frac{dG_{11,1}}{dt} = 2\beta_4 G_{24,5} - (\alpha_1 + \gamma_{1,1})G_{11,1} \tag{1.1}$$

$$\frac{dG_{1k,1}}{dt} = \alpha_{k-1} G_{1k-1,2} - (\alpha_k + \gamma_{1,k})G_{1k,1} \tag{1.2}$$

$$\frac{dG_{1k,2}}{dt} = \alpha_k G_{1k,1} - (\alpha_k + \gamma_{1,k})G_{1k,2}, \quad 1 \leq k \leq 4 \tag{1.3}$$

$$\frac{dG_{21,1}}{dt} = \alpha_4 G_{14,2} - (\beta_1 + \gamma_{2,1})G_{21,1} \tag{1.4}$$

$$\frac{dG_{2i,j}}{dt} = \beta_i G_{2i,j-1} - (\beta_i + \gamma_{2,i})G_{2i,j}, \quad 2 \leq j \leq 5, 1 \leq i \leq 4 \tag{1.5}$$

The parameters of the model include progression rates through G_1 phase, α, and $S - G_2$ phase, β, and death rates in each of the G_1 phase, γ_1, and $S - G_2$ phase, γ_2. Cells at the end of the $S - G_2$ phase divide and give birth to two cells at G_1 phase. Because each phase is divided into 4 parts, each part of G1 contains 2 sub-phases, and each part of $S - G_2$ contains 5 sub-phases.

The model was implemented in Julia v1.5.3. The differential equations were solved by the matrix exponential. As the data was measured with equal spacing, we pre-calculated the transition matrix between timesteps.

Dose Response Relationship

We assumed that the progression and death rates in G_1 and $S - G_2$ that form the quantified drug effects on the population follow a Hill function:

$$Hill(C) = E_{min} + \frac{E_{max} - E_{min}}{(1 + \frac{EC_{50}}{C})^k} \tag{1.6}$$

where the EC_{50} indicates the half-maximal drug effect concentration, E_{min} the value of the rate parameter in the absence of drug, E_{max} the rate parameter at infinite concentration, and

k the steepness of the dose-response curve. Given these parameters and the drug concentration C we then calculated the specific rate parameters for that treatment.

Exponential Model

To show the benefit of our LCT model, we employed a commonly used exponential model to fit to the G1 and $S - G_2$ cell numbers and showed that the exponential model cannot capture the dynamics of the data. The parameters were the same as the mean-field model.

$$\frac{dG_1}{dt} = 2\beta G_2 - (\alpha + \gamma_1)G_1 \tag{1.7}$$

$$\frac{dG_2}{dt} = \alpha G_1 - (\beta + \gamma_1)G_2 \tag{1.8}$$

Model Fitting

The data included the percentage of cells in G_1 phase and the total number of cells normalized to the cell numbers at the initial time point. We assumed 1 starting cell at 0H and calculated the number of cells in G_1 and $S - G_2$ phase over time. The Savitzky-Golay filter was used to smooth the data. Three replicates of each experiment were averaged, and the average was used for the purpose of fitting.

The number of G_1 and $S - G_2$ subphases, and the parameters in the absence of drug were shared across all drugs and concentrations. The sum of squared error was used as the cost function value and was calculated between the cell numbers predicted by the model and the average cell numbers of three replicates, over all time points, concentrations, and drugs tested. This cost function was then minimized using the default adaptive differential evolution optimizer from the BlackBoxOptim.jl Julia package, version 0.5.0.

To characterize the identifiability of our fit parameters we conducted a local sensitivity analysis. To do so, we calculated the cost function while varying each parameter from 0.1 to 10 times the optimal value, holding all the other parameters at their optimum. All parameters

were identifiably constrained by this analysis.

Calculating relative number of cell deaths and average phase durations. We evaluated the number of dead cells at 96H relative to the starting cell number at 0H. This formed the observed relative cell death numbers reported in **Figure 3A**. To calculate the corresponding cell death values inferred from the model, we calculated the predicted number of cells at each phase part (G_{11}, G_{12}, G_{13}, G_{14}, G_{21}, G_{22}, G_{23}, G_{24}) separately, and multiplied them by their individual death rates at all time points. This provides the number of dead cells at each phase part at each time point. The sum of cell numbers died in each phase part provides the total cell death counts at each time point, $n(t)$. **Figure 2C-D** show the accumulated dead cells across time for lapatinib and gemcitabine treatments which was calculated by summing over the cell death counts, $n(t)$, across time from 0 to each timepoint, T. Calculating for 96H results in the total cell death normalized by the initial cell numbers, 1, this value refers to the relative predicted cell death number reported in **Figure 3A**.

$$n(t) = \sum_{i=1}^{2} G_{i,j}(t) \times \gamma_{i,j} \tag{1.9}$$

$$N(T) = \sum_{t=0}^{T} n(t) \tag{1.10}$$

The average phase durations $\bar{G_1}$,$S - \bar{G_2}$ from the model were calculated using the progression rates. The G_1 phase has 8 subphases which is divided into 4 parts that results in 2 phases per part. $S - G_2$ phase has 20 subphases divided into 4 parts that results in 5 subphases in each part. Each phase part has a unique parameter for cell death and phase progression rate. The average phase duration will be given by the following expressions, derived by recognizing that the time in each part is gamma-distributed with a shape parameter

equal to the number of subphases.

$$\bar{G_1} = \sum_{j=1}^{4} \frac{2}{\alpha_{1j}} \tag{1.11}$$

$$S - \bar{G_2} = \sum_{j=1}^{4} \frac{5}{\beta_{2j}} \tag{1.12}$$

Predicting Drug Combinations

Bliss independence was used to calculate the predicted effect of drug combinations. Assuming E_a and E_b to be the saturable, quantified effects of drugs a and b, the expected combined effect would be:

$$E_{ab} = E_a + E_b - E_a.E_b \tag{1.13}$$

For death effects, we added the effects of each drug to find the death effect of the drug combination:

$$D_{ab} = D_a + D_b \tag{1.14}$$

The Bliss relationship requires that data first be scaled to be between 0 and 1, and then scaled back after the interaction calculation:

$$\hat{X} = \frac{X_{control} - X}{X_{control}} \tag{1.15}$$

This measure is usually used as a baseline to decide whether the combination of two drugs is synergistic or antagonistic. Here we used Bliss in two ways: (1) on the progression rate parameters to simulate the model predictions of drug combinations; and (2) on cell numbers to serve as a baseline approach to calculate drug combination effects, as is commonly used. In the first case, we use Bliss additivity on the cell cycle progression rates (*) to find the

set of progression parameters representing the combined treatment and assume that the death effects are only additive because the cell death process is not saturable (**). The combination parameters for all the eight concentrations for all pairs of drugs were calculated and then converted back to their original units. Next, we simulated the cell numbers using these parameters. In the baseline case, we used the cell numbers in the control condition to normalize the cell number measurements and then converted the cell numbers back to their original scale. This was used as a benchmark reference.

SUPPLEMENTARY MATERIALS

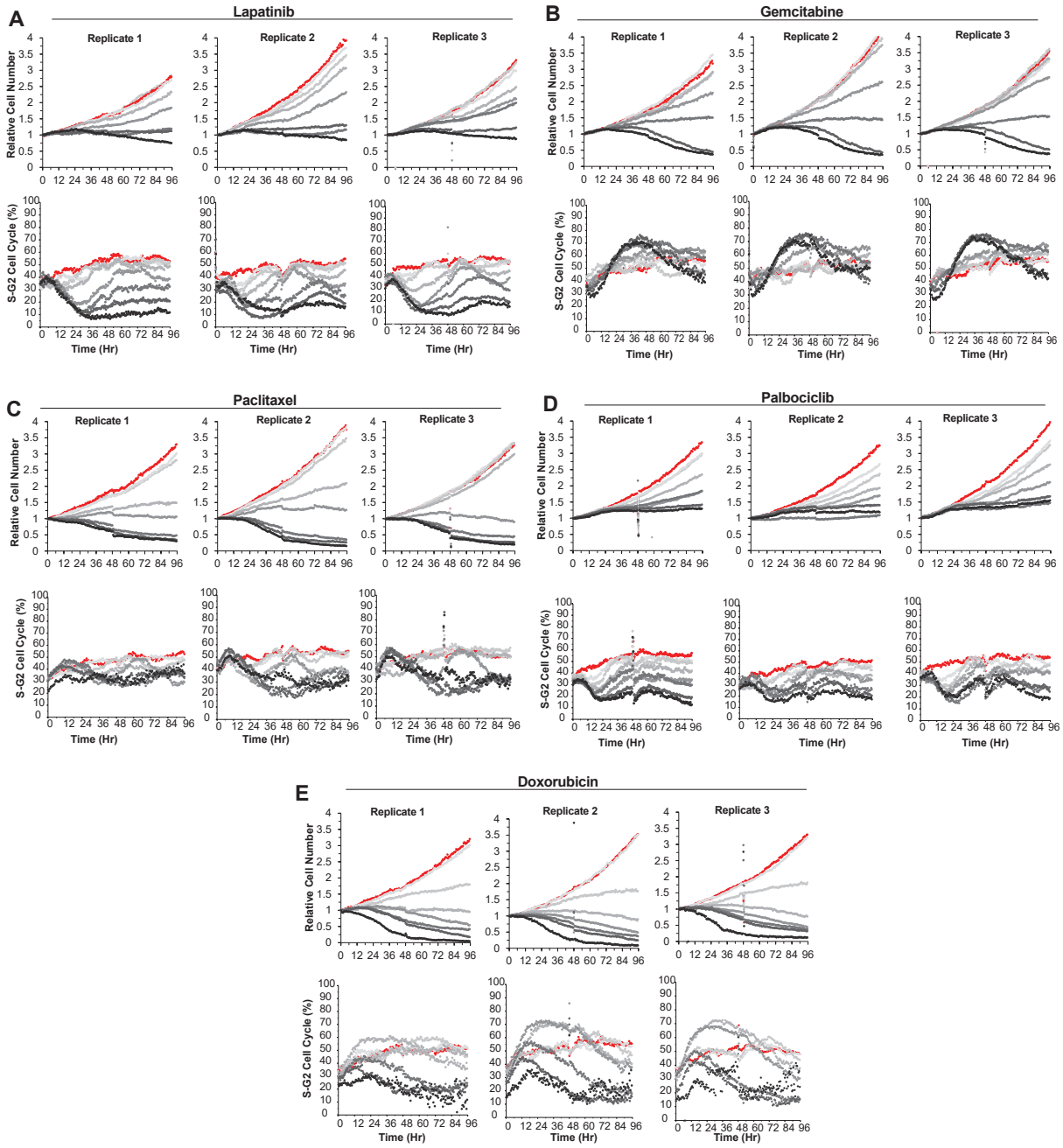


Figure S1: Individual replicates for AU565 drug responses show similar temporal dynamics and drug-induced changes to cell cycle. Panels show relative cell numbers and $S - G_2$ normalized cell numbers for lapatinib (**A**), gemcitabine (**B**), paclitaxel (**C**), palbociclib (**D**), and doxorubicin (**E**) treatments for three biological replicates. Five drug concentrations (gray lines) and untreated control (red line) are plotted.

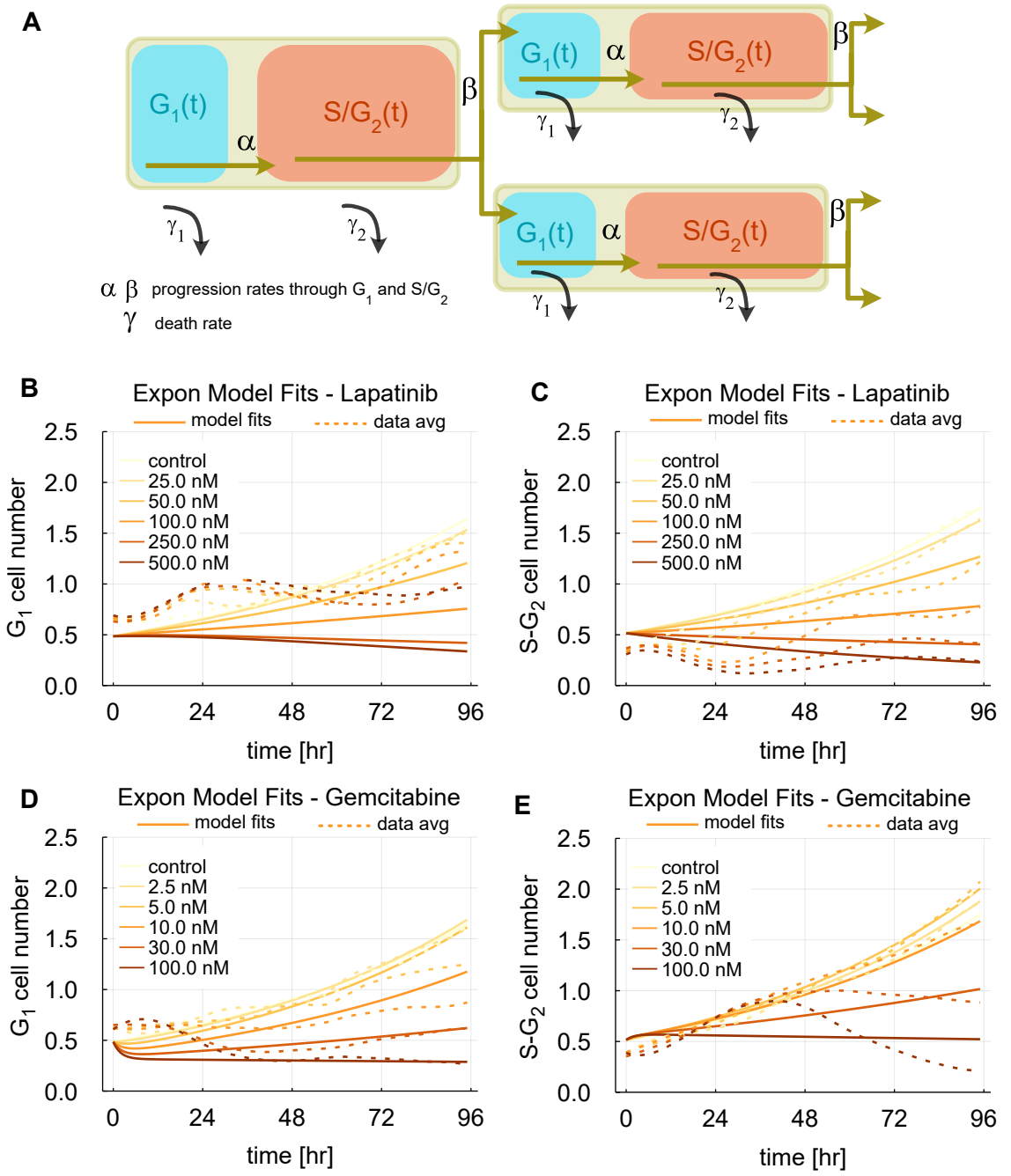


Figure S2: An exponential cell cycle model without incorporating delay times fails to capture the dynamics of drug response. A. The transition diagram for a simple dynamical model with 2 phases (G_1 and $S-G_2$) and without the LCT. α and β are the transition rates from G_1 to $S-G_2$ and vice versa, γ_1 and γ_2 are the death rates in G_1 and $S-G_2$, respectively. **B-E.** Exponential cell cycle model simulations of G_1 and $S-G_2$ cell numbers over time for control and 5 concentrations of lapatinib (**B-C**) and gemcitabine (**D-E**) (solid lines), respectively, overlayed with the average of three experimental replicates (dashed lines).

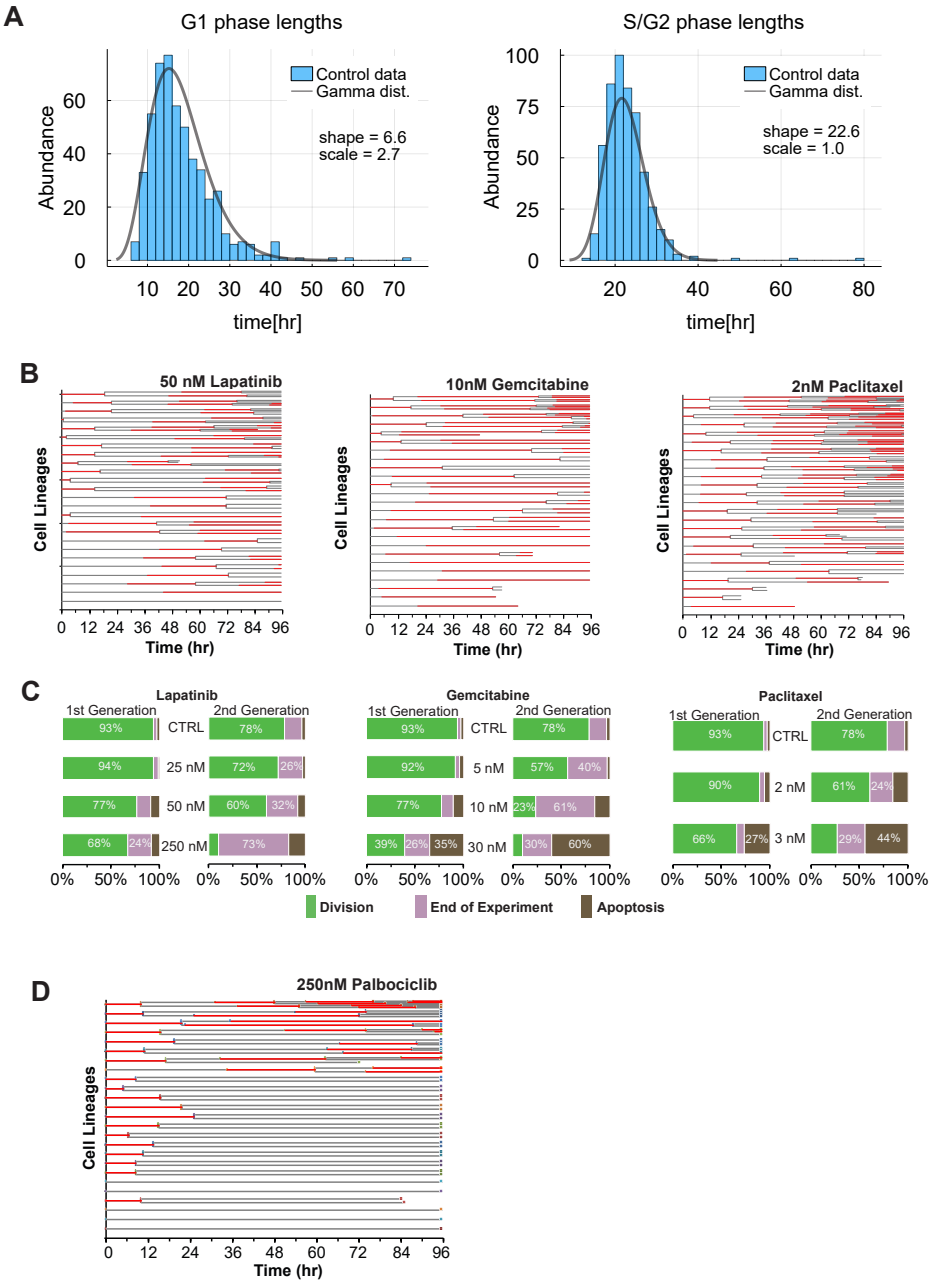


Figure S3: Analysis of single cell tracking data reveals drug-specific cell cycle phase effects in AU565 cells. A. Lineage trees of 25 lineages across 96H for various drug treatments. Tracks are colored coded based on cell cycle phase: gray indicates G_1 and red indicates $S - G_2$ phase. Track splitting indicates mitosis, and track ending prior to 96H corresponds to apoptosis. **B.** Quantification of cell outcomes (division, apoptosis, still present at end of experiment) for cells from the first and second generations treated with lapatinib, gemcitabine, or paclitaxel. **C.** Gamma distribution of G_1 and $S - G_2$ phase durations for cells in control condition with sample size of 520 and 514 for G_1 and $S - G_2$ phases, respectively. **D.** Lineage trees for 25 lineages across 96H after treatment with Palbociclib.

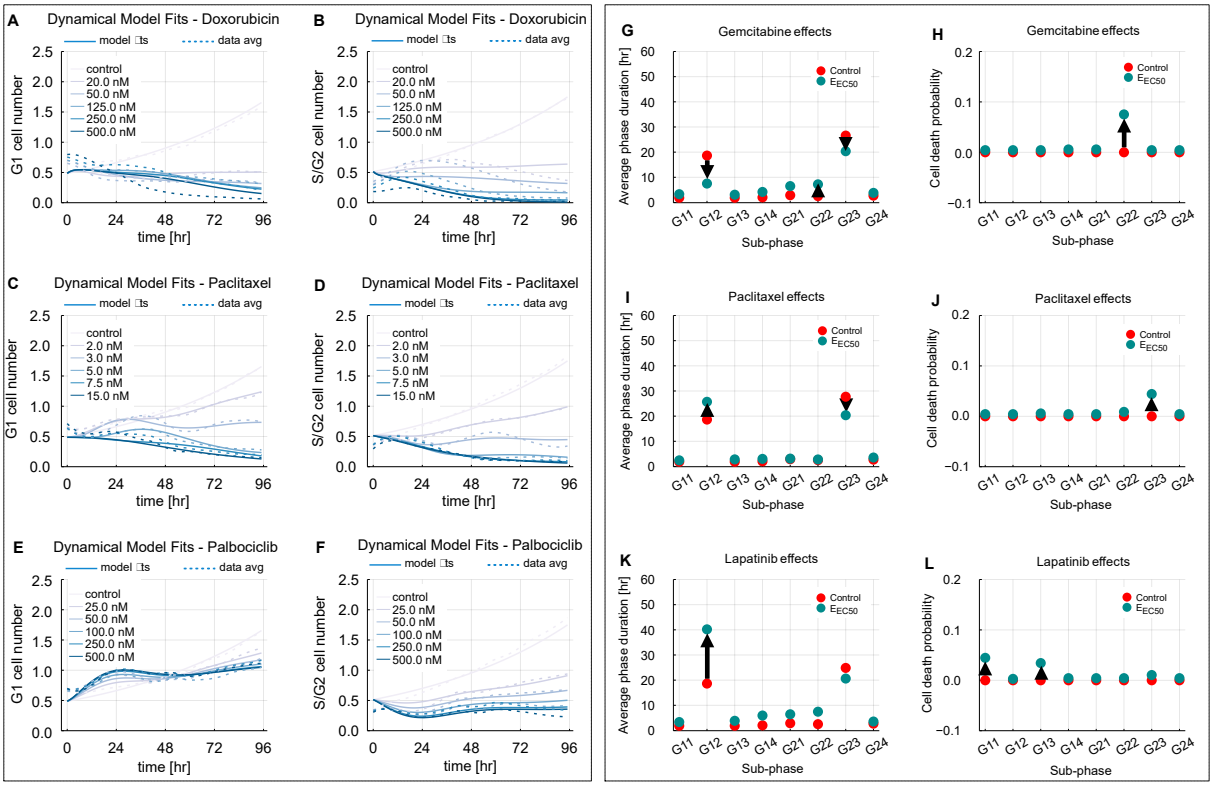


Figure S4: A dynamical model of the cell cycle captures the dynamics of drug response. A-F. $G1$ and $S - G_2$ cell numbers overtime, respectively, for the control and treatment at 5 concentrations (solid lines) for doxorubicin (**A-B**) paclitaxel (**C-D**), and palbociclib (**E-F**) overlayed with the average of three corresponding experimental replicates (dashed lines). **G-L.** The average phase durations in $G1$ and $S - G_2$ phases for selected drug treatments. The arrow shows the shift from the control condition to the drug effect at the half maximum concentration (E_{EC50}).

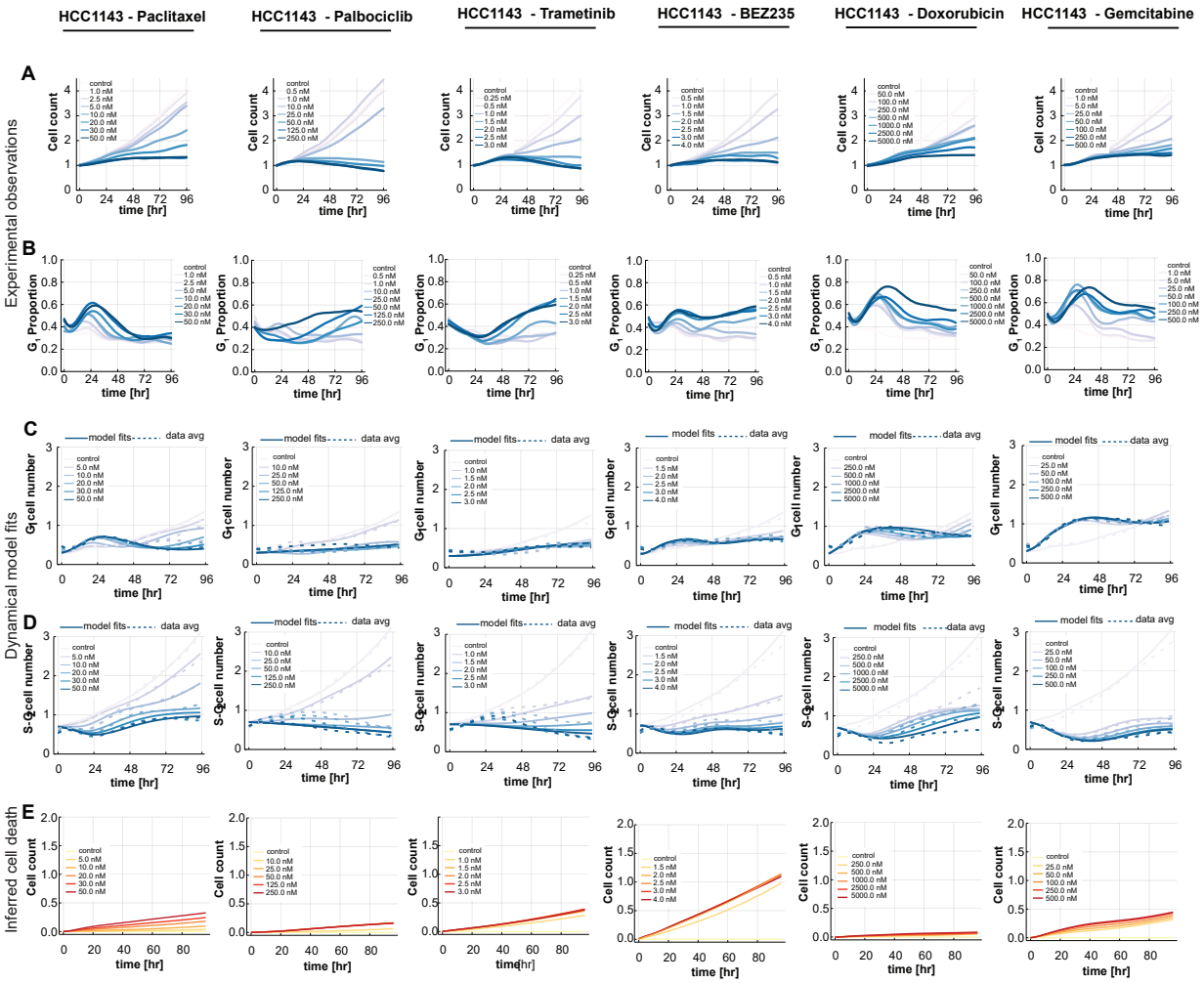


Figure S5: The introduced dynamical model captures the cell cycle dynamics of drug response in TNBC cell line HCC1143. A,B. Experimentally observed drug-induced changes to cell numbers (A) and G1 cell cycle phase proportion (B) after dose-escalation treatment with a panel of inhibitors. **C,D.** G1 and $S - G_2$ fits overtime, respectively, for the untreated and treatment at 5 concentrations (solid lines) overlayed with the average of three corresponding experimental replicates (dashed lines) for 6 drug treatments. **E.** Inferred accumulated dead cells over time for 6 drug treatments.

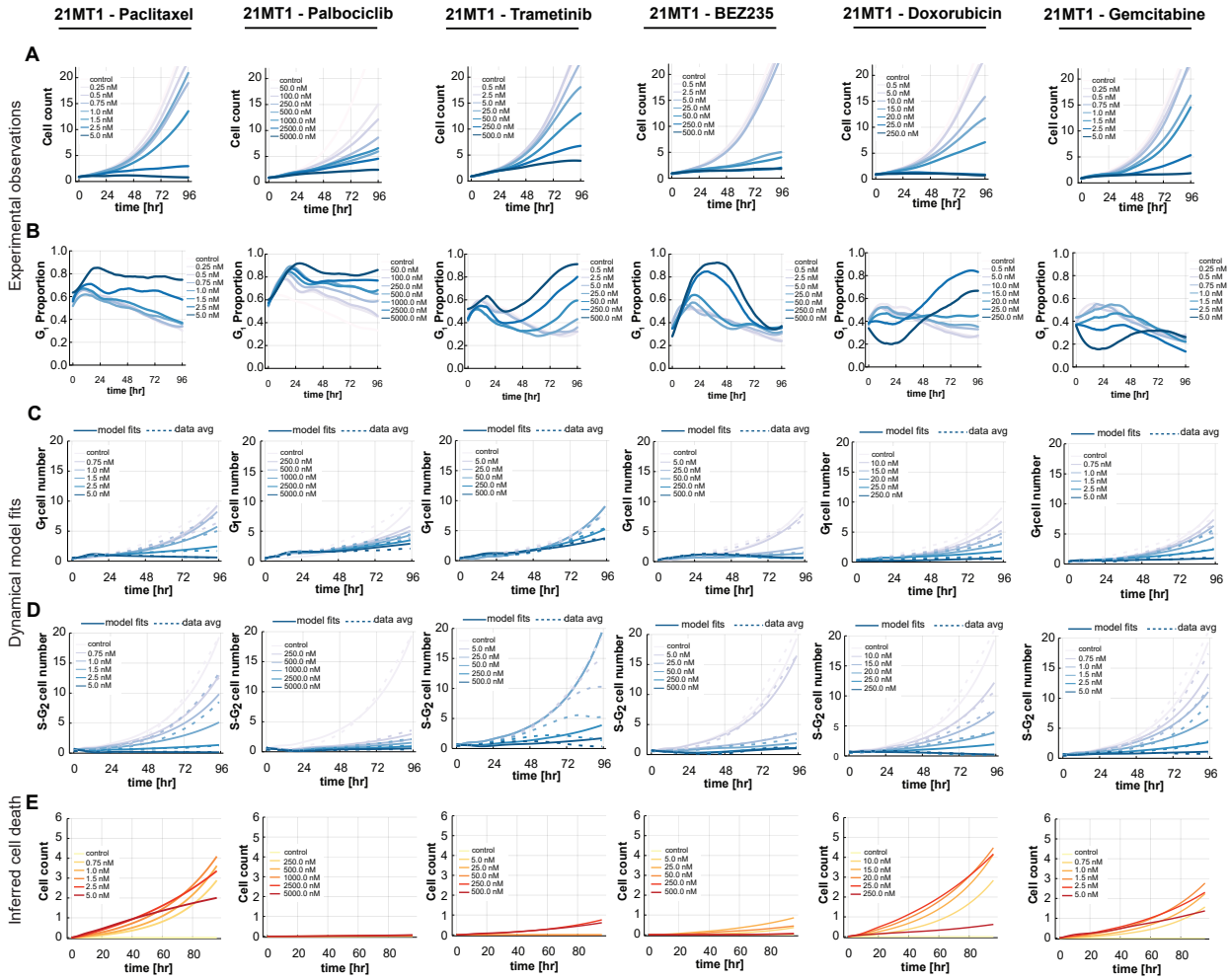


Figure S6: The introduced dynamical model captures the cell cycle dynamics of drug response in 21MT1 cell line. A, B. Experimentally observed drug-induced changes to cell numbers (**A**) and G1 cell cycle phase proportion (**B**) after dose-escalation treatment with a panel of inhibitors. **C,D.** G1 and S-G2 fits overtime, respectively, for the untreated and treatment at 5 concentrations (solid lines) overlayed with the average of three corresponding experimental replicates (dashed lines) for 6 drug treatments. **E.** Inferred accumulated dead cells over time for 6 drug treatments. **A,B.** Experimentally observed drug-induced changes to cell numbers (**A**) and G1 cell cycle phase proportion (**B**) after dose-escalation treatment with a panel of inhibitors. **C,D.** G_1 and $S - G_2$ fits overtime, respectively, for the untreated and treatment at 5 concentrations (solid lines) overlayed with the average of three corresponding experimental replicates (dashed lines) for 6 drug treatments. **E.** Inferred accumulated dead cells over time for 6 drug treatments.

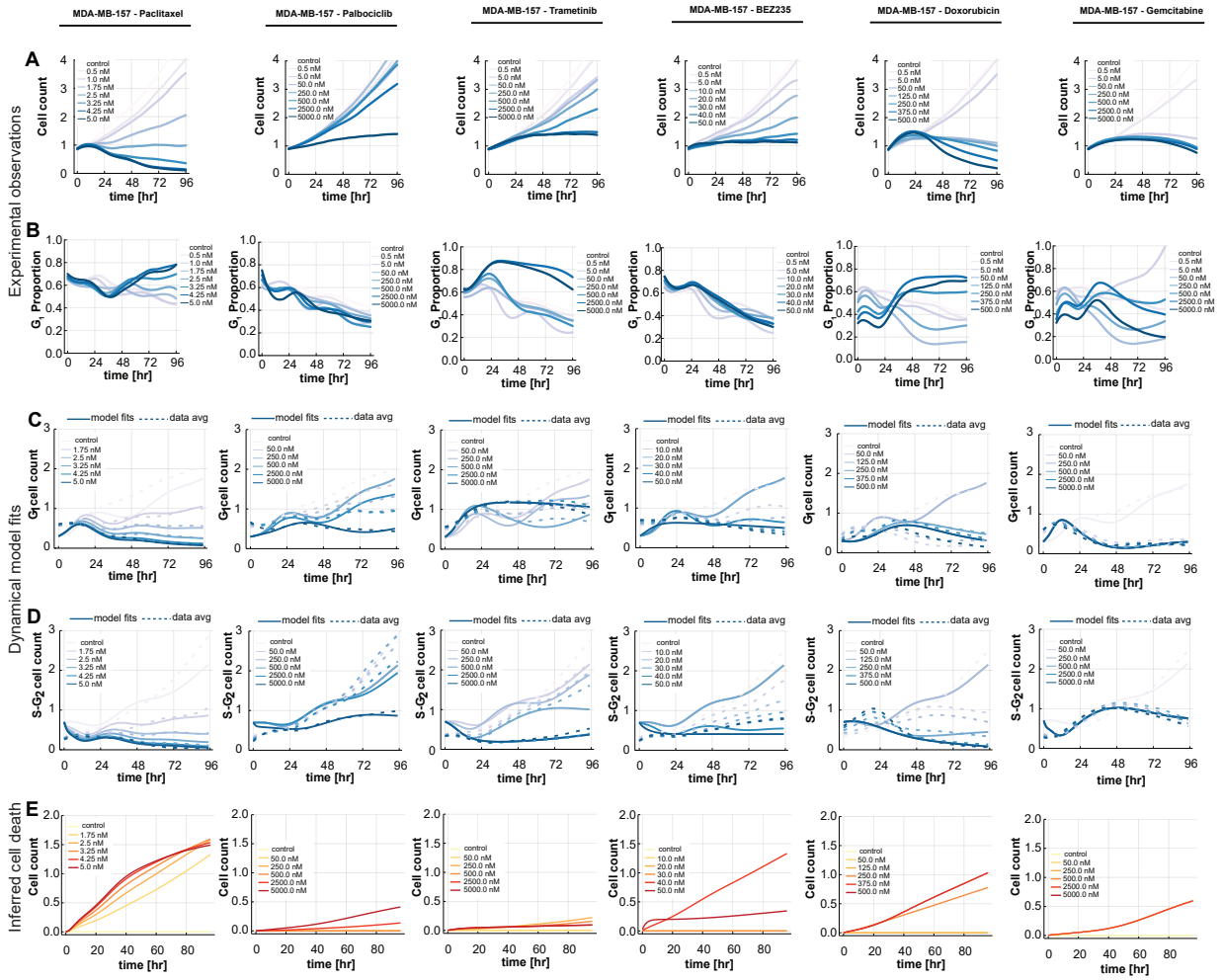


Figure S7: The introduced dynamical model captures the cell cycle dynamics of drug response in TNBC cell line MDA-MB-157. A,B. Experimentally observed drug-induced changes to cell numbers (**A**) and G_1 cell cycle phase proportion (**B**) after dose-escalation treatment with a panel of inhibitors. **C,D.** G_1 and $S-G_2$ fits overtime, respectively, for the untreated and treatment at 5 concentrations (solid lines) overlayed with the average of three corresponding experimental replicates (dashed lines) for 6 drug treatments. **E.** Inferred accumulated dead cells over time for 6 drug treatments.

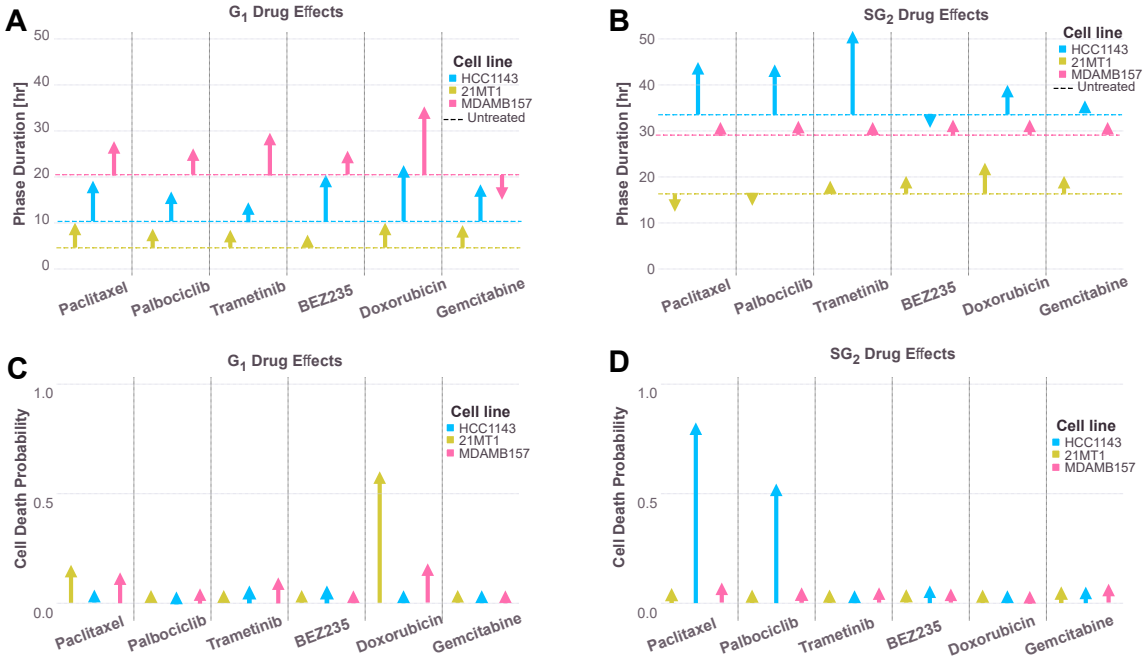


Figure S8: Summary of inferred cell cycle drug effects at half maximum concentration compared to untreated. A-B. The average phase durations in G1 (**A**) and $S - G_2$ (**B**) phases for HCC1143 (blue), 21MT1 (olive) and MDA-MB-157 (pink) treated with paclitaxel, palbociclib, trametinib, BEZ235, doxorubicin, and gemcitabine. The dashed lines show the average phase duration at untreated for each cell line. **C-D.** The cell death probability in $G1$ (**C**) and $S - G_2$ (**D**) phases for HCC1143 (blue), 21MT1 (olive) and MDA-MB-157 (pink) treated with the same panel of drugs. The arrows show the quantity of increase or decrease in the effects from untreated to the half maximal concentration (E_{EC50}).

Bibliography

[1] Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022, 2017.

[2] Jennifer O'Neil, Yair Benita, Igor Feldman, Melissa Chenard, Brian Roberts, Yaping Liu, Jing Li, Astrid Kral, Serguei Lejnine, Andrey Loboda, et al. An unbiased oncology compound screen to identify novel combination strategies. *Molecular cancer therapeutics*, 15(6):1155–1162, 2016.

[3] Susan L Holbeck, Richard Camalier, James A Crowell, Jeevan Prasaad Govindharajulu, Melinda Hollingshead, Lawrence W Anderson, Eric Polley, Larry Rubinstein, Apurva Srivastava, Deborah Wilsker, et al. The national cancer institute almanac: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activitynci almanac of approved cancer drug combinations. *Cancer research*, 77(13):3564–3576, 2017.

[4] Michael P Menden, Dennis Wang, Mike J Mason, Bence Szalai, Krishna C Bulusu, Yuanfang Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature communications*, 10(1):2674, 2019.

[5] Christian T Meyer, David J Wooten, Carlos F Lopez, and Vito Quaranta. Charting the fragmented landscape of drug synergy. *Trends in pharmacological sciences*, 41(4):266–280,

2020.

[6] Aziz Sancar, Laura A Lindsey-Boltz, Keziban Ünsal-Kaçmaz, and Stuart Linn. Molecular mechanisms of mammalian dna repair and the dna damage checkpoints. *Annual review of biochemistry*, 73(1):39–85, 2004.

[7] Leland H Hartwell and Ted A Weinert. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246(4930):629–634, 1989.

[8] Kevin J Barnum and Matthew J O'Connell. Cell cycle regulation by checkpoints. *Cell cycle control: mechanisms and protocols*, pages 29–40, 2014.

[9] Pablo Lara-Gonzalez, Frederick G Westhorpe, and Stephen S Taylor. The spindle assembly checkpoint. *Current biology*, 22(22):R966–R980, 2012.

[10] Hui Xiao Chao, Randy I Fakhreddin, Hristo K Shimerov, Katarzyna M Kedziora, Rashmi J Kumar, Joanna Perez, Juanita C Limas, Gavin D Grant, Jeanette Gowen Cook, Gaorav P Gupta, et al. Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Molecular systems biology*, 15(3):e8604, 2019.

[11] Richard S Finn, Judy Dering, Dylan Conklin, Ondrej Kalous, David J Cohen, Amrita J Desai, Charles Ginther, Mohammad Atefi, Isan Chen, Camilla Fowst, et al. Pd 0332991, a selective cyclin d kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Research*, 11(5):1–13, 2009.

[12] Peng Huang and William Plunkett. Fludarabine-and gemcitabine-induced apoptosis: incorporation of analogs into dna is a critical event. *Cancer chemotherapy and pharmacology*, 36(3):181–188, 1995.

[13] Adam C Palmer, Christopher Chidley, and Peter K Sorger. A curative combination cancer therapy achieves high fractional cell killing through low cross-resistance and drug additivity. *Elife*, 8:e50036, 2019.

[14] Aziz Sancar, Laura A Lindsey-Boltz, Keziban Ünsal-Kaçmaz, and Stuart Linn. Molecular mechanisms of mammalian dna repair and the dna damage checkpoints. *Annual review of biochemistry*, 73(1):39–85, 2004.

[15] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

[16] Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012.

[17] Robert H Shoemaker. The nci60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, 2006.

[18] Caitlin E Mills, Kartik Subramanian, Marc Hafner, Mario Niepel, Luca Gerosa, Mirra Chung, Chiara Victor, Benjamin Gaudio, Clarence Yapp, Ajit J Nirmal, et al. Multiplexed and reproducible high content screening of live and fixed cells using dye drop. *Nature Communications*, 13(1):6918, 2022.

[19] Matthew T McKenna, Jared A Weis, Stephanie L Barnes, Darren R Tyson, Michael I Miga, Vito Quaranta, and Thomas E Yankeelov. A predictive mathematical modeling approach for the study of doxorubicin treatment in triple negative breast cancer. *Scientific reports*, 7(1):1–14, 2017.

[20] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature methods*, 13(6):521–527, 2016.

[21] Leonard A Harris, Peter L Frick, Shawn P Garbett, Keisha N Hardeman, B Bishal Paudel, Carlos F Lopez, Vito Quaranta, and Darren R Tyson. An unbiased metric of

antiproliferative drug effect in vitro. *Nature methods*, 13(6):497–500, 2016.

[22] Darren R Tyson, Shawn P Garbett, Peter L Frick, and Vito Quaranta. Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. *Nature methods*, 9(9):923–928, 2012.

[23] Sabrina L Spencer, Steven D Cappell, Feng-Chiao Tsai, K Wesley Overton, Clifford L Wang, and Tobias Meyer. The proliferation-quiescence decision is controlled by a bifurcation in cdk2 activity at mitotic exit. *Cell*, 155(2):369–383, 2013.

[24] Tzu-Hao Wang, Hsin-Shih Wang, and Yung-Kwei Soong. Paclitaxel-induced cell death: where the cell cycle and apoptosis come together. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 88(11):2619–2628, 2000.

[25] Peng Huang and William Plunkett. Induction of apoptosis by gemcitabine. In *Seminars in oncology*, volume 22, pages 19–25, 1995.

[26] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature methods*, 13(6):521–527, 2016.

[27] Shea N Gardner. A mechanistic, predictive model of dose-response curves for cell cycle phase-specific and-nonspecific drugs. *Cancer research*, 60(5):1417–1425, 2000.

[28] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[29] Paul J Hurtado and Adam S Kirosingh. Generalizations of the 'linear chain trick': incorporating more flexible dwell time distributions into mean field ode models. *Journal of mathematical biology*, 79(5):1831–1883, 2019.

[30] Paul J Hurtado and Cameron Richards. Building mean field ode models using the generalized linear chain trick & markov chain theory. *Journal of Biological Dynamics*, 15(sup1):S248–S272, 2021.

[31] Anneleen Daemen, Obi L Griffith, Laura M Heiser, Nicholas J Wang, Oana M Enache, Zachary Sanborn, Francois Pepin, Steffen Durinck, James E Korkola, Malachi Griffith, et al. Modeling precision treatment of breast cancer. *Genome biology*, 14:1–14, 2013.

[32] Keith T Flaherty, Jeffery R Infante, Adil Daud, Rene Gonzalez, Richard F Kefford, Jeffrey Sosman, Omid Hamid, Lynn Schuchter, Jonathan Cebon, Nageatte Ibrahim, et al. Combined braf and mek inhibition in melanoma with braf v600 mutations. *New England Journal of Medicine*, 367(18):1694–1703, 2012.

[33] Hui Xiao Chao, Randy I Fakhreddin, Hristo K Shimerov, Katarzyna M Kedziora, Rashmi J Kumar, Joanna Perez, Juanita C Limas, Gavin D Grant, Jeanette Gowen Cook, Gaorav P Gupta, et al. Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Molecular systems biology*, 15(3):e8604, 2019.

[34] Sean T Vittadello, Scott W McCue, Gency Gunasingh, Nikolas K Haass, and Matthew J Simpson. Mathematical models incorporating a multi-stage cell cycle replicate normally-hidden inherent synchronization in cell proliferation. *Journal of the Royal Society Interface*, 16(157):20190382, 2019.

[35] Tatsiana Ryl, Erika E Kuchen, Emma Bell, Chunxuan Shao, Andrés F Flórez, Gregor Mönke, Sina Gogolin, Mona Friedrich, Florian Lamprecht, Frank Westermann, et al. Cell-cycle position of single myc-driven cancer cells dictates their susceptibility to a chemotherapeutic drug. *Cell systems*, 5(3):237–250, 2017.

[36] Hui Xiao Chao, Cere E Poovey, Ashley A Privette, Gavin D Grant, Hui Yan Chao, Jeanette G Cook, and Jeremy E Purvis. Orchestration of dna damage checkpoint dynamics across the human cell cycle. *Cell systems*, 5(5):445–459, 2017.

[37] A Kathleen McClendon, Jeffry L Dean, Dayana B Rivadeneira, Justine E Yu, Christopher A Reed, Erhe Gao, John L Farber, Thomas Force, Walter J Koch, and Erik S Knudsen. Cdk4/6 inhibition antagonizes the cytotoxic response to anthracycline therapy. *Cell cycle*, 11(14):2747–2755, 2012.

[38] Jeffry L Dean, A Kathleen McClendon, and Erik S Knudsen. Modification of the dna damage response by therapeutic cdk4/6 inhibition. *Journal of Biological Chemistry*, 287(34):29075–29087, 2012.

[39] Guan N. Yan R. Warner K. Taylor S. D. Bae, S. Y. and A. S. Meyer. Measurement and models accounting for cell death capture hidden variation in compound response. *Cell Death and Disease*, 11(8), 2020.

[40] Ryan Richards, Hannah R Schwartz, Megan E Honeywell, Mariah S Stewart, Peter Cruz-Gordillo, Anna J Joyce, Benjamin D Landry, and Michael J Lee. Drug antagonism and single-agent dominance result from differences in death kinetics. *Nature chemical biology*, 16(7):791–800, 2020.

[41] Christian A Yates, Matthew J Ford, and Richard L Mort. A multi-stage representation of cell proliferation as a markov process. *Bulletin of mathematical biology*, 79:2905–2928, 2017.

[42] Marilyne Labrie, Joan S Brugge, Gordon B Mills, and Ioannis K Zervantonakis. Therapy resistance: opportunities created by adaptive responses to targeted therapies in cancer. *Nature reviews Cancer*, 22(6):323–339, 2022.

[43] Bryce T Bajar, Amy J Lam, Ryan K Badiee, Young-Hee Oh, Jun Chu, Xin X Zhou, Namdoo Kim, Benjamin B Kim, Mingyu Chung, Arielle L Yablonovitch, et al. Fluorescent indicators for simultaneous reporting of all four cell cycle phases. *Nature methods*, 13(12):993–996, 2016.

[44] Aaron S Meyer and Laura M Heiser. Systems biology approaches to measure and model phenotypic heterogeneity in cancer. *Current opinion in systems biology*, 17:35–40, 2019.

[45] Sean M Gross, Mark A Dane, Elmar Bucher, and Laura M Heiser. Individual cells can resolve variations in stimulus intensity along the igf-pi3k-akt signaling axis. *Cell systems*, 9(6):580–588, 2019.

[46] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

[47] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods*, 16(12):1226–1232, 2019.

[48] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A Cimini, Kyle W Karhohs, Minh Doan, Liya Ding, Susanne M Rafelski, Derek Thirstrup, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, 16(7):e2005970, 2018.

[49] Erik Meijering, Oleh Dzyubachyk, and Ihor Smal. Methods for cell and particle tracking. *Methods in enzymology*, 504:183–200, 2012.

[50] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[51] Katharina Löffler, Tim Scherr, and Ralf Mikut. A graph-based cell tracking algorithm with few manually tunable parameters and automated segmentation error correction. *PloS one*, 16(9):e0249257, 2021.

Chapter 2

Accounting for Cell Cycle Effects Identifies CDK Therapy Rational Combinations

Farnaz Mohammadi, John Moffat, Steffan Vartanian, Aaron S Meyer, and Marc Hafner

Abstract

The effectiveness of cancer therapies is limited by incomplete response to treatment, which enables propagation of the tumor and selection of resistant cells. One potential approach to overcoming this limitation is identification of combination therapies that can deepen therapeutic response. At the core of cancer's unchecked proliferation is regulation of the cell cycle; therefore, the mechanism and timing by which therapies modulate progression through the cell cycle is critical for developing effective therapeutic combinations. Most pharmacologic responses are measured only by counting viable cells as an endpoint assay, which provides insufficient information to infer cell cycle effects. However, recent experimental approaches have provided high-throughput techniques to collect robust and accurate drug response data while incorporating markers to quantify cell cycle state. We apply a modeling-based and data-driven approach to use this data in identifying effective therapeutic combinations and patterns of response resolving effects specific to certain cell cycle phases. Our mechanistic model was able to identify off-target effects of drugs with respect to cell cycle phases and predict the combination effects specific to each cell cycle phase. In addition, with our data-driven approach we found associations between cell cycle progression rates and BRCA mutation in breast cancer cell lines. In total, this demonstrates the value in quantifying and directly accounting for therapeutic effects incorporating cell cycle information.

Introduction

Targeted therapies have been developed to modulate various components of the cell cycle to inhibit cell growth or induce cell death. The targeting molecules are often used in combination with chemotherapy [1] and immunotherapy [2] for a more effective response. Among common cell cycle inhibitors are cyclin-dependent kinases (CDKs) inhibitors [3, 4], which are enzymes that have an essential role in cell cycle regulation and progression. Other targeted molecules include apoptosis inducers [5] and DNA damage response inhibitors [6]. Cells in different phases of the cell cycle exhibit varying degrees of sensitivity to different drugs and each small

molecule inhibitor effects a specific part of the cell cycle; for example, CDK4/6 inhibitors such as palbociclib inhibit the activity of CDK4/6-cyclin D which then disrupts the phosphorylation of Rb1 protein, and result in cell cycle arrest primarily in the G1 phase of the cell cycle [7]. On the other hand, chemotherapy drugs also often have a specific target within the cell cycle; for example, doxorubicin disrupts DNA repair in the S phase and induces cell cycle arrest and death in the S phase of the cell cycle [8]. Hence, it is crucial to accurately understand the impact of drugs on the cell cycle for the development of effective cancer treatments, and also to understand how the combination of drugs work mechanistically. However, not many studies have focused on investigating the drug response with cell cycle phase resolution.

Cancer cell lines from the same type with varying genotypic features often respond differently to the same treatment. For example, there has been extensive studies highlighting heterogeneous response of different triple negative breast cancer cell lines, despite the fact that they all are from the same nominal subtype [9]. Incorporating information on the cellular genotype can also provide insights into the drug's mechanism of action and help to identify patient subgroups that are most likely to benefit from specific therapies [10]. CDK inhibitors usually force cells to exit the cell cycle by triggering senescence, quiescence, or apoptosis. As such, CDK4/6-cyclin D plays a role in preventing cell cycle exit, hence, the inhibition of CDK4/6 can force cells to exit the cell cycle [11]. It has been shown that amplification of cyclin D levels which results in increased activity of CDK4/6 confers sensitivity to CDK4/6 inhibitors [12]. Also patients with cancer-associated mutations in proteins involved in cell cycle exit pathways, such as p16 (CDKN2A mutation), are more likely to benefit from CDK4/6 inhibitors [13], and those with mutations in pathways driving the cell cycle entry, such as cyclin E or Rb1, are less likely to benefit from it [14]. It is important to considering the cellular genotype in cancer therapy studies, as they demonstrate how specific genetic alterations can impact the efficacy of different drugs and personalizing medicine.

Quantifying drug response is essential to understanding mechanism of action of drugs, discovery of new targeted molecules, and drug combination screening. A common approach

for studying drug response is using single snapshot of total cell viability several hours after drug exposure [15], and employing a sigmoidal function to estimate the parameters such as concentration at half maximal effect (EC_{50}), asymptotic fractions of viable cells (E_{max}, and E_{min}), and some measure of potency (Hill slope). However, depending on the target pathway, each anti-cancer agent can arrest cells in a specific cell cycle phase which results in a non-uniform response in each cell cycle phase. Classical models of drug response that use total cell viability cannot distinguish the specific part of the cell cycle that is perturbed. Previously, we have developed a mathematical model that can capture the drug-induced dynamics of drug response using time series data of breast cancer cells harboring a cell cycle reporter to identify G1 and S/G2 cell cycle phases [16]. This model improved prediction of drug combination in various scenarios such as with two drugs arresting cells at the same or differing phases of the cell cycle, and also shed light in dynamical behavior of cells across time which showed a pseudo-synchronization effect as a result of various treatments. However, the data acquisition process, including cell line development that incorporates the nuclear reporter to distinguish G1 from S/G2 phase, cell segmentation and tracking across time are technically challenging and difficult to scale.

Advances in high-throughput screening technologies and imaging techniques have enabled the development of more sophisticated drug response assays that can provide more detailed insights into the mechanisms of drugs and the cellular responses [17]. Multiplexed approaches such as the dye drop assay developed by Mills et al. [18], includes sequential density displacement and microscopy which eliminates the mix and wash step. They utilize fluorescent dyes to label the amount of total and newly synthesized DNA content and based on that, distinguish cell cycle phases, making it a promising tool for studying drug response with cell cycle phase resolution in a high throughput manner. This approach also allows for follow-on assays of fixed cells by the use of immunofluorescence [19].

Dimensionality reduction techniques are popular approaches in studying large scale multimodal data, especially where there are measurements can be arranged in several dimensions,

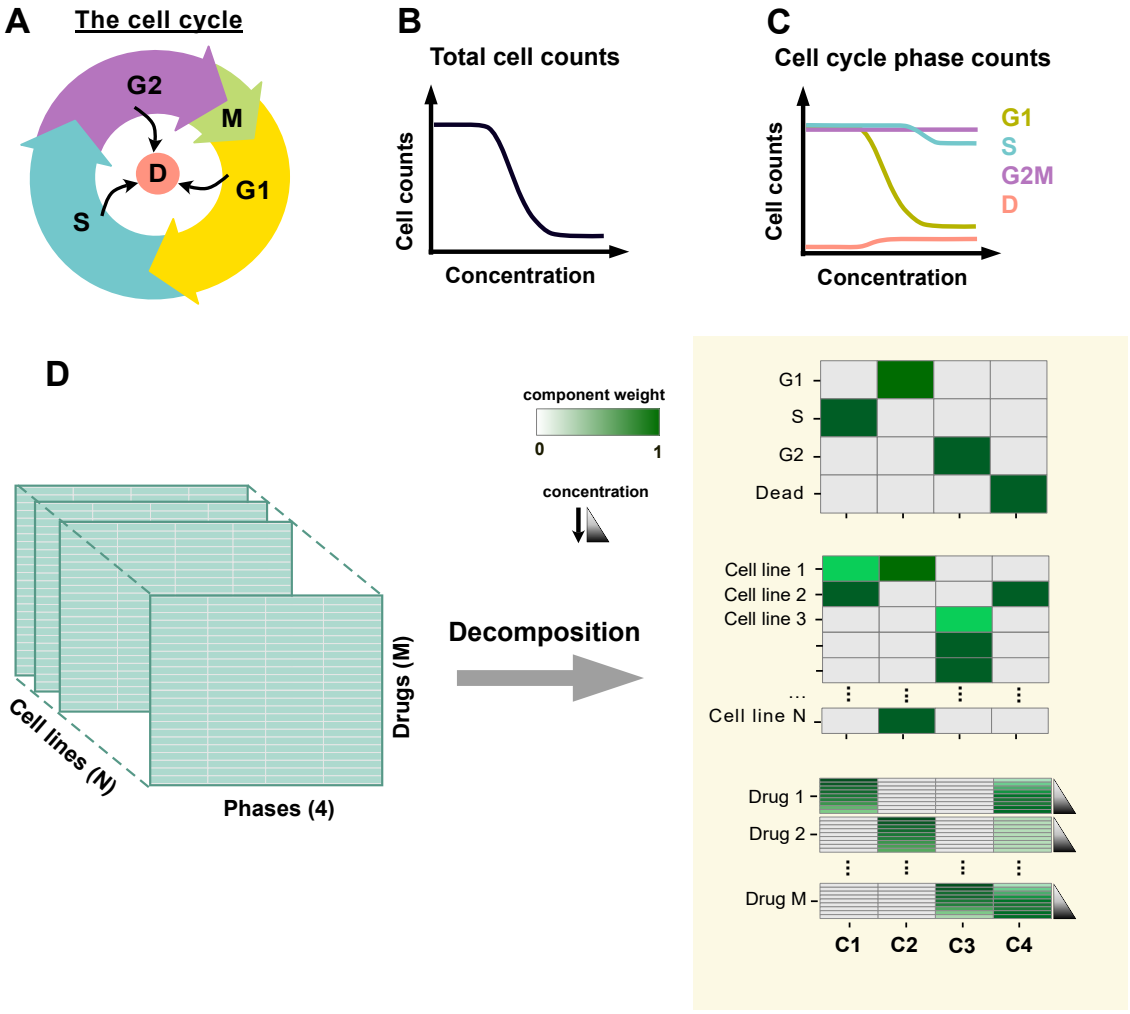


Figure 1: Cell cycle phase specific data reveals in-depth information about drug effects compared to total cell counts. A. A schematic of the cell cycle. **B.** Traditional drug response data providing the total cell counts across a range of concentrations. **C.** Decomposing the total cell counts into cell cycle phases in a high throughput manner using the dye drop assay. **D.** The data structure acquired from the dye drop assay and using tensor decomposition for pattern recognition across cell cycle phases, cell lines, and drugs.

such as, cell lines, drugs, time, and concentration. Tensor decomposition approaches, such as CP decomposition (also known as PARAFAC) reduce the multi-dimensional data into a series of 1-dimensional vectors that each correspond to one of the dimensions in the original dataset [20]. In contrast to matrix decomposition where all of the features are flattened along one dimension, tensor decomposition techniques reduce the data efficiently while preserving the structure of the data. Generally, dimensionality reduction techniques are useful in noise

reduction, data visualization, and data compression [20].

In this study, we explore the effects of various anti-cancer agents on breast cancer cell lines at different cell cycle phases through two complementary approaches; first, using unsupervised tensor decomposition technique to reveal patterns throughout the large scale data which otherwise would be difficult to obtain from the raw data, and second, using an ordinary differential equation model to quantify cell cycle phase specific rates of transition and cell death that can be used to predict drug combinations. We hypothesize that a systematic approach to dissecting dynamical cell cycle phase-specific effects not only uncovers the mechanism of action of the targeting molecules under exploration individually, but also can promote well-informed decision-making for screening drug combinations. To this end, we explore the dataset introduced in Mills et al [18] (the HMS dataset) and extract some of the patterns within the dataset; and then, we fit a system of ordinary differential equations (ODEs) to the data and further investigate the quantified drug effects at different cell cycle phases. Employing this model we demonstrate the advantage of using our model for drug combination analysis and explore the combination of CDK inhibitors with chemotherapy.

Results

1. A tensor-based strategy to analyze cell cycle-dissolved drug response measurements

The cell cycle is a tightly regulated series of events in a sequential manner that ultimately results in cell division (**Figure 1A**). The cell cycle includes four main stages that are G1, S, G2, and M. Progression through the cell cycle depends on the interaction of various enzymes that activate or deactivate signalling pathways at each phase. Small molecules that can manipulate the activity of essential elements of the cell cycle have been proposed as strategies for anti-cancer therapy [21]. Among important examples are the inhibitors of cyclin dependent kinases (CDK inhibitors) [22–24]. To study how various anti-cancer agents modulate the

cell cycle, we selected a subset of treatments and cell lines from the dataset introduced in Mills et al [18]. The selected dataset includes response measurements of 54 cell lines treated with 33 anti-cancer drugs at 9 concentrations. As opposed to traditional viability assays (**Figure 1B**), the dye drop assay allows one to dissect the total cell counts into individual cell cycle phases within each condition robustly and in a high-throughput manner (**Figure 1C**). In Mills et al [18] authors employed a cell death marker and an M phase marker to further segregate the cell cycle states into a total of 8 phases and subphases. The cell lines within this dataset include a variety of clinical subtypes of breast cancer, such as triple negative, ER positive, HER2 amplified, and HR positive. The panel of drugs includes small molecule inhibitors that target different pathways such as PI3K/AKT/mTOR, the DNA damage response pathway, and those inhibiting the cell cycle machinery elements such as CDKs or disrupting the transcriptional events that regulate the cell cycle.

To explore and visualize the patterns existing within this large dataset, we employed a tensor decomposition technique, called CP decomposition (or PARAFAC), which captures a low-rank representation of the original data (**Figure 1D**). In this way, we reduced the dataset while preserving the relationship between the dimensions (here cell lines, phases, and drugs). Tensor decomposition is capable of capturing patterns specific to each dimension and the associations between dimensions. For example, patterns can be associated with a specific cell cycle phase and variably represented across the cell lines and drug treatments (**Figure 1D**).

2. Dataset structure and a comparison between CP decomposition and PCA

Considering all the dimensions, the HMS dataset is a $54 \times 33 \times 10 \times 8$ tensor. To reduce the amount of missing data we combined the drugs and their associated concentrations into one dimension which restructured the tensor to a $54 \times 324 \times 8$, where in total we have close to 12000 data points (**Figure 2A**). CP decomposition decomposes the tensor

into a set of 1-dimensional vectors that each correspond to one mode in the original tensor (**Figure 2B**). To select the optimum number of components, we examined the accumulated explained variance and a core consistency measure to ensure no over-fitting by evaluating how consistently the selected components capture the true structure of the data (**Figure 2C, D**). The R2X measure showed that more than 95% of the total variance can be explained by only 6 components (**Figure 2D**). This tensor decomposition approach preserved the relationship between different dimensions in this multi-modal dataset whereas, with other dimensionality reductions techniques such as PCA, we would be forced to flatten the input tensor into a two dimensional matrix which eliminates the relationship between the dimensions [25]. Moreover, CP decomposition reduced the data more efficiently than PCA (**Figure 2E**).

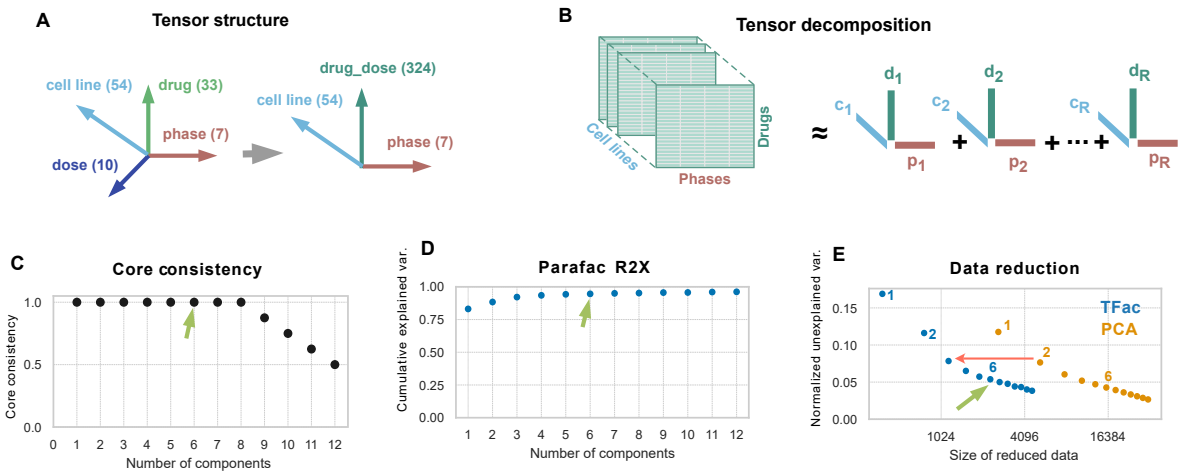


Figure 2: Tensor factorization compresses data efficiently. A. Reshaping the dataset from 4 dimensions to 3 dimensions. **B.** Tensor decomposition for the 3-dimensional data. **C.** Core consistency metric shows that with 8 components we are still not overfitting. **D.** R2X measure or the explained variance with increasing number of components. **E.** Comparing PCA with CP decomposition in their data reduction power.

3. Decomposition of the dataset revealed cell line- and drug-specific patterns and association with cellular genotype

The CP decomposition analysis revealed that each factorization component corresponds to one or two cell cycle phases, and the drug dimensions exhibited concentration-dependent

trends (**Figure 3**). Component 1 predominantly captured G1 accumulation, and exhibited strong associations with the majority of the cell lines and drugs (**Figure 3A-C**). Component 2 was associated with G2 arrest, with CDK1/2, CDK7, CDK12/13, and CDK9 inhibitors in addition to paclitaxel demonstrating highest associations. Component 3 represented accumulation in the S phase, that has a decreasing trend across concentrations for a majority of agents meaning that with higher concentrations, fewer and fewer of the cells are found in S phase. Component 4 was positive in the beyond G2 phase, highly weighted in SUM185PE, PDXHCI002, and HCC1187 cell lines which represented multinucleated cells as a result of failure in cytokinesis [26]. Component 5 and 6 reflected cell death, and exhibited a concentration-dependent effect in components of transcriptional regulators active mostly in the M phase, including CDK12 and CDK13 inhibitors, which affect transcriptional regulation and DNA damage response, as well as CDK9 inhibitors that are crucial for cancer cell survival through aiding in transcriptional elongation, and CDK14 inhibitors that can result in mitotic defects and cell death.

CDK4/6 inhibitors displayed similar patterns across components, except for abemaciclib at its highest concentration, which induced additional effects related to cell death and G2 phase arrest, consistent with off-target effects observed in a previous study by Hafner et al [27]. Agents that target the PI3K/Akt/mTOR pathway, such as AZD5363, AZD2014, LY3023414, and everolimus also exhibited a similar pattern across all components.

We sought to investigate whether the factor matrices of the drug response data can predict cellular subtypes, such as triple negative (TNBC), HER2 receptor positive (HER2+), Estrogen or Progesterone receptor positive (HR+), or genetic mutations such as PTEN, PIK3CA, BRCA, or CDKN2A. To do so, we used the 54 by 6 matrix of the cell lines factors from CP decomposition and implemented logistic regression with L1-norm regularization. The target outcome was a binary vector which indicated whether each cell line is positive for the subtype/genetic mutation (1) or not (0). Since most of the target categories were under-represented, we used SMOTE (synthetic minority oversampling technique) [28, 29] to

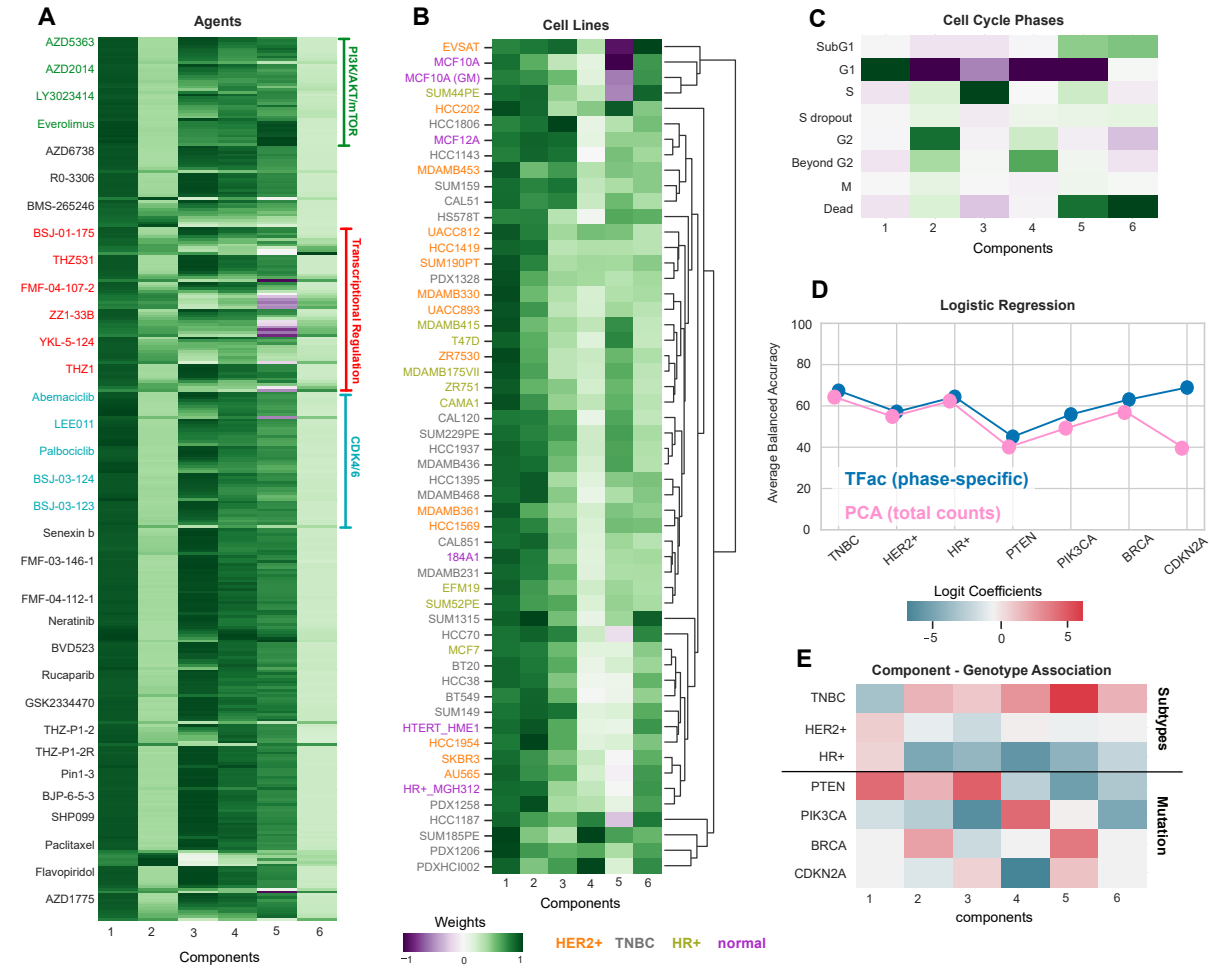


Figure 3: The cell cycle-specific drug responses reveal specific patterns across drugs and cell lines with ability to predict cellular genotypes. A-C. Heatmap of agents (**A**), cell lines (**B**), and cell cycle phases (**C**) for 6 components. The color bar shows normalized weight of components between 0 and 1. Drugs in **A** are ordered by concentration from untreated to the highest. Cell lines in **B** are clustered using the average Eucledian distance and color-coded based on cellular subtype. **D.** The balanced accuracy of the logistic regression for each cellular subtype or genotype using factor matrices of the cell cycle fraction CP decomposition (blue) or from PCA analysis of total cell counts (pink). **E.** The coefficients of the logistic regression model corresponding to each component to identify associations between factor matrices and cellular genotypes (cellular subtypes and genetic mutations).

adjust the imbalance within the dataset. To demonstrate whether there is an advantage in using cell cycle specific measurements over the total cell counts in predicting the genotypes, we conducted PCA analysis for the matrix of total cell numbers across drug treatments (a matrix of 54 by 324). With 6 components the PCA recovered 92% of the total variance. It

appeared that the cell cycle measurements versus total cell counts had a slight improvement in dissecting the cellular genotype (**Figure 3D**). The coefficients of the logistic regression model revealed associations between the genotypes and each component (**Figure 3E**) with TNBC, HR+ and CDKN2A having close to 70% accurate to identify the correct classes (**Figure 3D**).

4. A dynamical ODE model quantifies cell cycle phase-specific drug effects

To quantitatively investigate the cytostatic and cytotoxic effects of drugs on the cell cycle, we developed an ordinary differential equation (ODE) model that represents transitions through the cell cycle phases. We simplified the cell cycle to three phases, G1, S, and G2M; the G2 and M phases were combined due to the short duration of the M phase and lack of distinct markers. Additionally, we assumed that the cell death rate was shared among the phases, as there was no way to distinguish the number of dead cells arising from each cell cycle phase. The Markov chain Monte Carlo approach (MCMC) [30] was used to estimate the transition rates between cell cycle phases and the cell death rate. To improve the interpretability of the results, we imposed a Hill function on the cell cycle rates, in effect enforcing that the behavior of the transition and death rates be monotonic with respect to concentration (**Figure 4A**).

Our ODE model accurately fit the HMS dataset at each cell cycle phase and provided estimates of the cell cycle transition rates and cell death rate (**Figure 4B-D**). We confirmed the convergence of the MCMC by diagnostics such as Gelman-Rubin and the effective sample size diagnostics(**Tables S1, S2**). The results of our estimations confirmed our observation of the subtle differences in the mechanism of action of CDK4/6 inhibitors abemaciclib and palbociclib, particularly the higher potency of abemaciclib in G1 arrest and additional off-target effect of this agent on the G2 phase [27], (**Figure 4B**)

The live/dead markers that are used to identify dead cells in experiments are not completely reliable because dead cells may detach or wash out even before staining for live/dead markers.

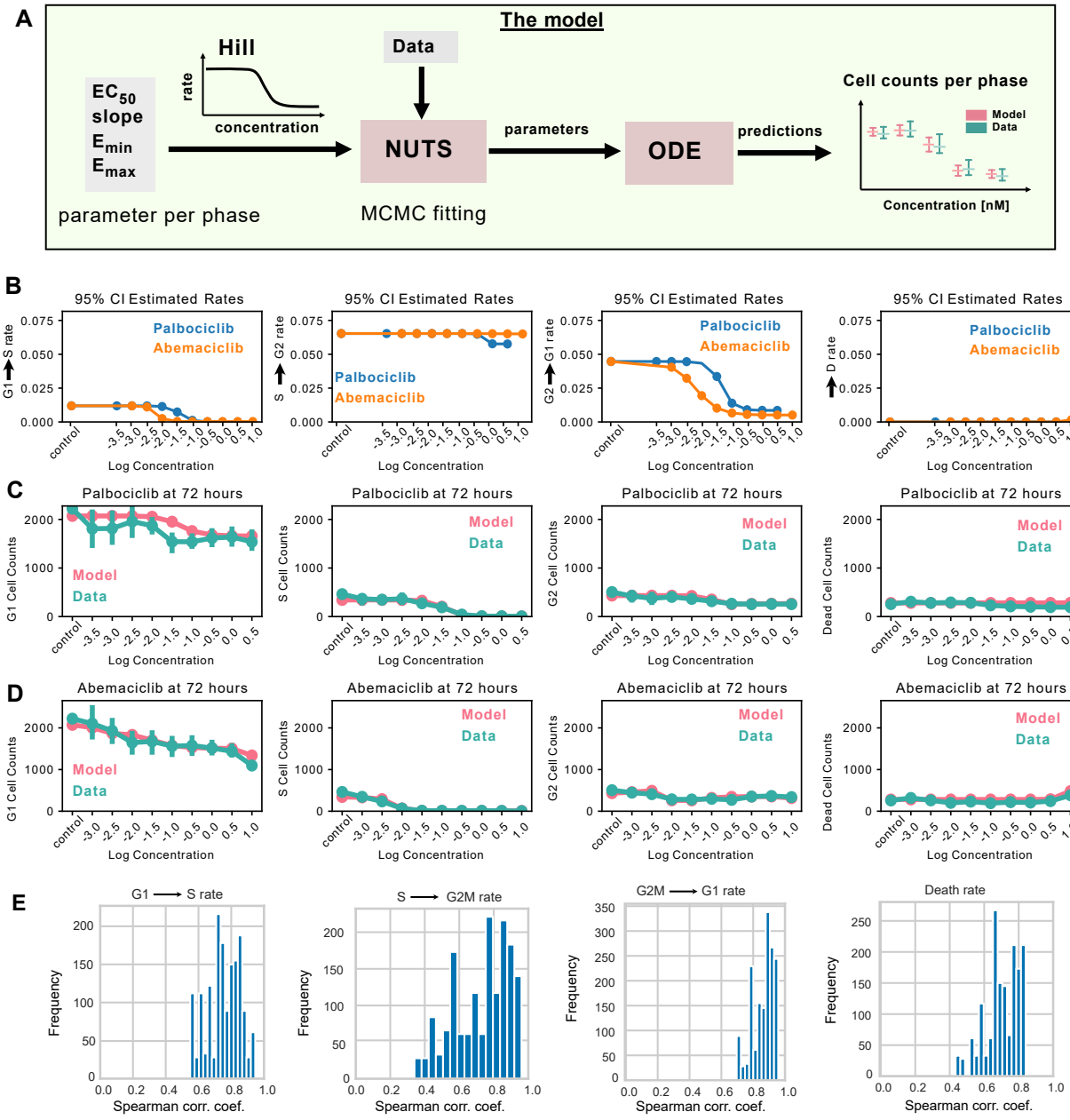


Figure 4: The modeling approach to estimate cell cycle transition and death rates. A. Overview of the modeling approach. The Hill parameters specific for each rate are estimated by a Bayesian inference approach (NUTS). The estimated Hill parameters are converted to ODE parameters and used for model predictions. **B-D.** An example of the fitting results from MDAMB175VII, an HR+ breast cancer cell line treated with palbociclib and abemaciclib. **B.** Estimated rates across concentrations for abemaciclib (orange) and palbociclib (blue) with 95% confidence interval. **C-D.** Predicted (pink) versus experimental cell counts (green) at each cell cycle phase for abemaciclib and palbociclib, respectively. **E.** The Spearman correlation coefficient of comparing estimations with and without having the number of dead cells, across 54 cell lines.

Consequently, the number of dead cells reported from experiments could be an under-estimate of the true count. Therefore, we sought to investigate how well we can estimate the cell death rate without having the dead cell counts from the experiments. To do so, we compared the estimated rates while including and excluding the cell death counts from the model. The Spearman correlation coefficient was used to compare the estimated rates between the two configurations (**Figure 4E, S1**). Overall the average Spearman correlation coefficient was consistently above 0.7 confirming we can reliably estimate the cell death rates even without directly having the number of dead cells. In most cases, we found that removing the number of dead cells from the equation resulted in slightly higher death rates, matching our expectation of under-estimation for dead cell counts from the experiments.

5. Cell cycle rates provide further details for phase-specific and cell line-specific drug effects and associate with cellular genotypes

We used our factorization technique on estimated rates from the model to gain further insights into the drugs' effects and to examine whether the estimated progression and cell death rates are better predictors of cell types than the data itself (**Figure 5**). The core consistency metric and R2X showed that with 3 components we can retrieve almost 70% of the total variance without overfitting (**Figure S2**). Among the cell lines, the majority of HER2-positive cell lines were associated with component 1 and most of the triple-negative cell lines were associated with component 2 (**Figure 5A**). Component 3 represented higher cell death and it was highly associated with one of the HR+ cell lines (**Figure 5B**). We hypothesized that the progression and estimate rates can predict the receptor status or genotype of the cell lines. To test this, we used the factor matrices of the cell lines resulting from the factorization, and performed the logistic regression for each cellular genotype. The balanced accuracy from the 5-fold cross validation showed that the factor matrices from the rates are better at predicting the BRCA mutation than the cell cycle data itself (**Figure 5C**). BRCA1 and BRCA2 genes are tumor suppressor genes that normally repair DNA damage

and prevent genomic instability. The coefficients of the logistic regression model suggested that an increase in component 1 results in an increase in the prevalence of cells with BRCA mutations. Component 1 has a positive association with S to G2M transition rate, which could mean that those with BRCA mutation miss the DNA repair during the S phase and bypass the checkpoint, hence, they are rushed out of the S phase (**Figure 5B, D-E**).

6. Tensor factorization captures subtle differences in drug combination data

We aimed to explore the combination effects of chemotherapy agents with CDK inhibitors on different cell lines using the dye drop assay – we will refer to this dataset as the GNE dataset from now on. Specifically, we conducted an experiment that involved the use of 12 individual and combinations of drugs, including CDK inhibitors (a CDK2 inhibitor (PF-07104091) [31], a CDK4/6 inhibitor (palbociclib and abemaciclib), and a CDK2/4/6 inhibitor (PF-06873600) [32]) and chemotherapy agents (dinaciclib, paclitaxel, and doxorubicin), across six cell lines with various genetic mutations or amplified receptor status as detailed in **Table 1**.

Cell Line	Receptor Status	Mutations
MDAMB468	Triple-negative	RB1 LOF, PTEN loss
BT-549	Triple-negative	RB1 LOF, PTEN loss
MDAMB175VII	HR+	
HCC1143	Triple-negative	CCND1 high
HCC1806	Triple-negative	CCNE1 high
OVCAR3	–	CCNE1 high, PIK3R loss

Table 1: Cell line genotypes of the GNE dataset. OVCAR3 is an ovarian cancer cell line.

As initial analysis we applied tensor decomposition on this dataset. Two components explained over 95% of the variance, with about 15 times more data reduction compared to PCA (**Figure S4**). The first component identified patterns associated with G1 phase accumulation, while the second component captured effects associated with the S and G2 phase of the cell cycle. The weights associated with each cell line comes from the cell counts;

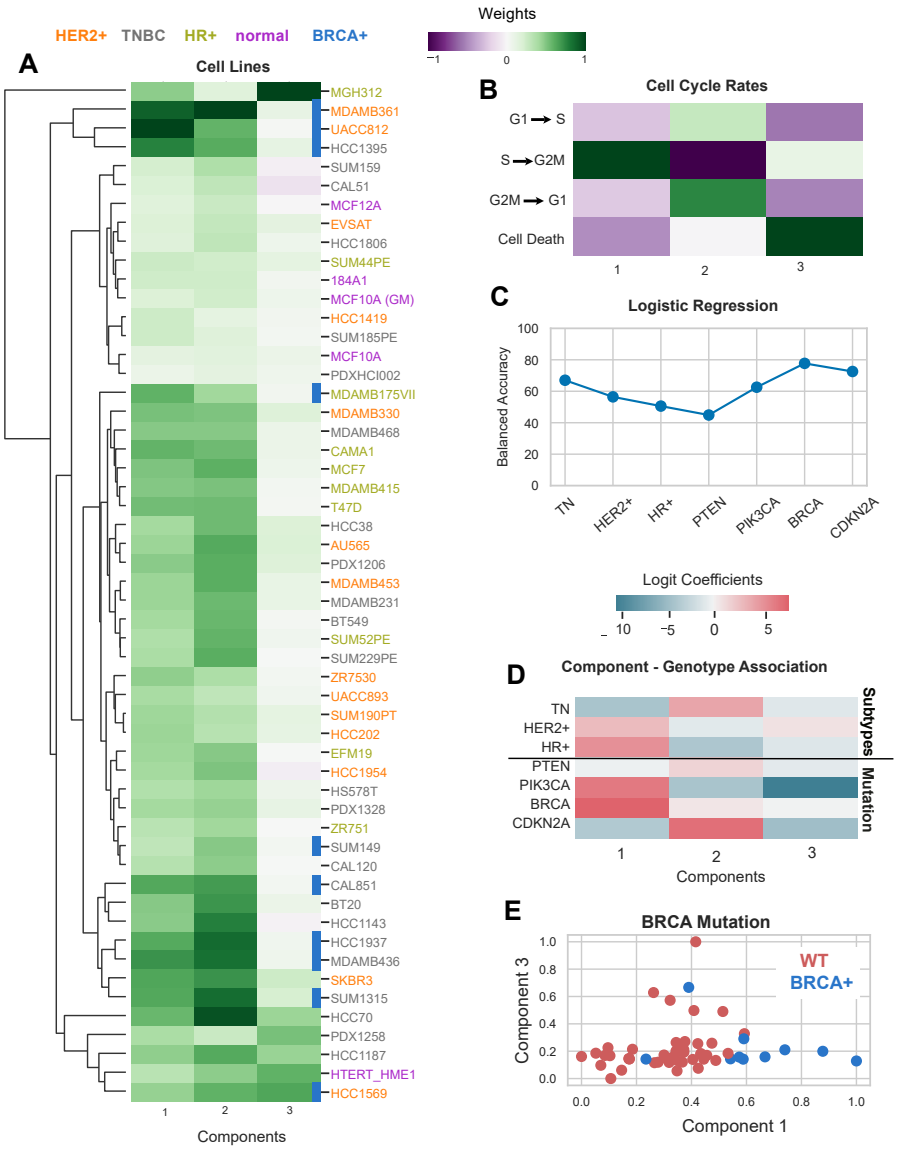


Figure 5: Factorization of estimated rates predicts association with cellular genotypes. The heatmap of cell lines (**A**) and cell cycle rates (**B**) factor matrices for 3 components after CP decomposition. Cell lines in **A** are clustered using the average Eucledian distance. The cell lines are color-coded as orange (HER2+), gray (triple negative), olive (HR+), and normal (purple). The BRCA-mutated cell lines are color coded as dark blue. The color bar shows normalized component weights between 0 and 1. **C.** The balanced accuracy of the logistic regression for each cellular genotype using factor matrices of the estimated cell cycle rates from the ODE model. **D.** The coefficients of the logistic regression model corresponding to each component to identify association between factor matrices and cellular genotypes (cellular subtypes and genetic mutations). **E.** Component 1 versus component 3 of the decomposition for 54 cell lines, color coding the cell lines based on BRCA mutation: dark blue (BRCA-mutated), red (wild type).

for instance, MDAMB468 and to a lesser extent OVCAR3 cells have higher growth rates than other cell lines. HCC1806 cell line was distinguished in a sense that the nominal abundance of cells at the S phase was relatively higher than all other cell lines (**Figure 6A-B**), which could be associated with its amplification of CCNE1. Notably, CDK2i alone and in combination with 0.3 μM CDK4/6i (palbociclib) have similar trends, implying that the palbociclib has no additional effects in the G1 arrest. This is also observed in the combinations of paclitaxel (taxol) with palbociclib and doxorubicin (dox) with palbociclib (**Figure 6C**). In the combinations of doxorubicin and paclitaxel with CDK2i, however, accumulations in the G1 (component 1) and the S or G2 phase (component 2) have started to drop at lower concentrations.

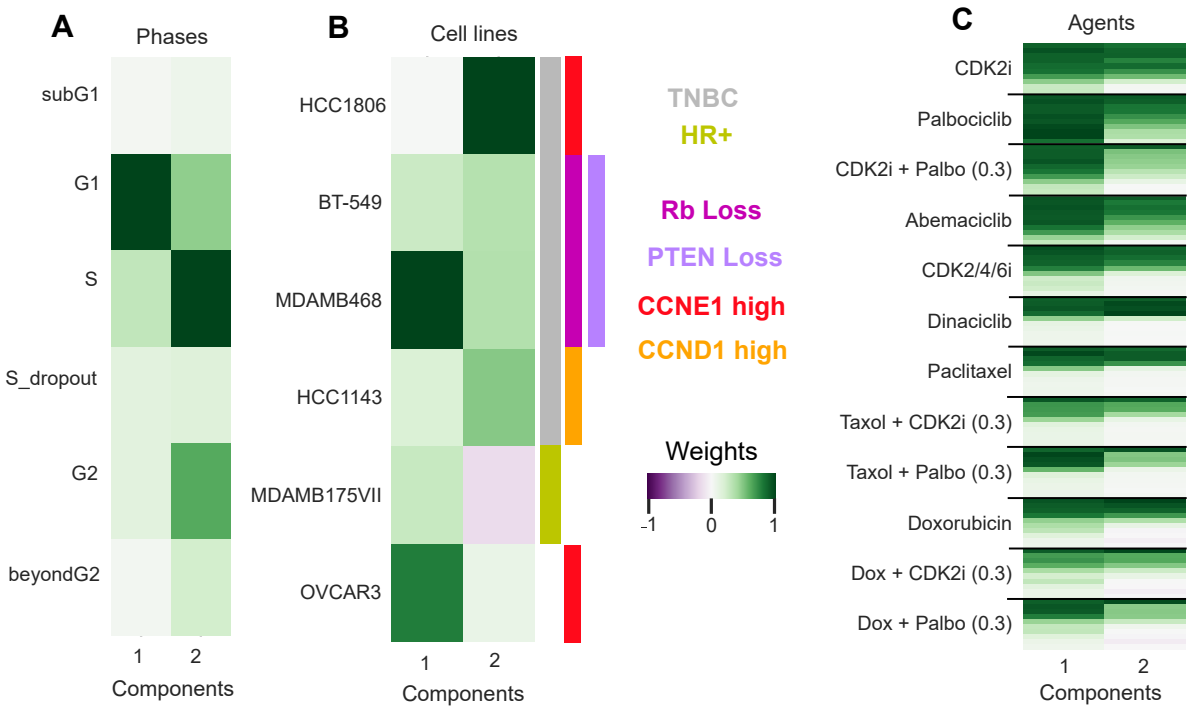


Figure 6: Tensor factorization captures subtle differences in drug combination responses. A. R2X plot summarizing the cummulative explained variance with increasing number of components. **B.** Data reduction plot comparing efficiency of PCA with CP decomposition. **C.** Core consistency metric to determine the number of optimal components. The heatmaps for two components for the cell cycle phases (**D**), cell lines (**E**), and agents (**F**). The color bar shows normalized component weights.

To examine the drug combinations, we estimated the progression and death rates as a

result of fitting this dataset to our ODE model. The cell line list includes four triple negative breast cancers, one HR+ breast cancer, and an ovarian cancer cell line with various genotypes. As expected, the cell lines responded heterogeneously (**Figure S5**). BT549 and MDAMB468 with RB loss of function were not responsive to CDK4/6 inhibitors (**Figure S5A**). Although these two cell lines both are triple negative and in addition to RB loss of function also have a PTEN mutation, they differed in their response. One reason for that could be that MDAMB468 cells have a higher doubling time (2.5 days versus 1.7 days). Also, the PTEN mutation in BT-549 is heterozygous (one normal copy and one mutated copy of the PTEN gene) and in MDAMB468 is homozygous (both copies mutated). Homozygous mutation in the PTEN gene has been associated with more severe phenotypic consequences [33]. MDAMB468 has higher cell death rates compared to BT-549 (**Figure S5D**) possibly due to having more cells in S phase at the initial time of drug administration (**Figure S6**). BT-549 and HCC1143 cell lines both show no arresting effects with paclitaxel alone or in combination at G1, but they get arrested in G2M phase and undergo cell death. With doxorubicin alone and in combination, HCC1806 had a significant S phase arrest and cell death (**Figure S5B, D**), also possibly due to a higher initial number of cells in S phase (**Figure S6**). The high G1 arrest in this cell line could be caused by DNA damaging effect of doxorubicin treatment at G1 which stalls for DNA repair (**Figure S5A**). Based on our estimation of EC50 for palbociclib and the CDK2 inhibitor at each cell line, 0.3 μM is less than EC50 of G1 phase in most of the cell lines (**Table S3**), hence, we think that is the reason the effects at combinations are not significantly different from single drug treatments. Overall, we can conclude that the growth rate, the initial composition of cells across the cell cycle, and cellular genotype can influence the response to drugs individually or in combination, and we can capture these details in the mechanisms of drugs with our ODE model.

7. The ODE model can accurately predict drug combination response.

We incorporated the Bliss independence framework into our ODE model to predict the drug combination responses of CDK inhibitors with chemotherapy. To do so, we applied the

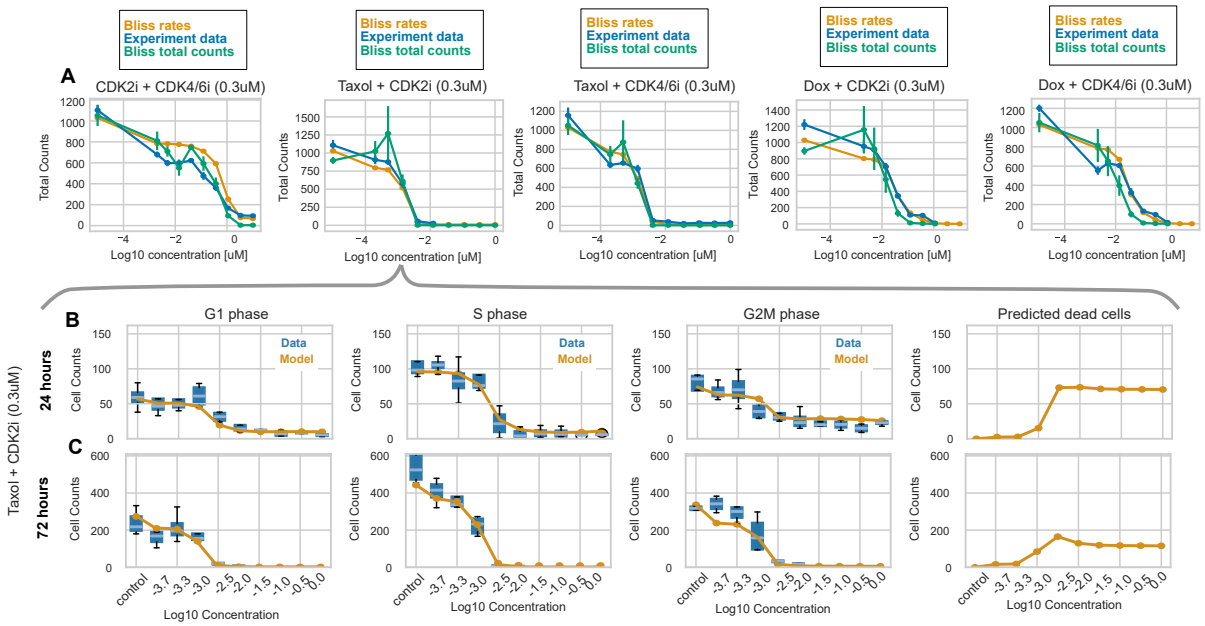


Figure 7: The ODE model predicts combination outcomes at each cell cycle phase. A. Comparison between classic drug response predictions by Bliss (Bliss total counts, green), our proposed approach which is Bliss applied on transition rates (Bliss rates, yellow) overlayed with the experimental data of the combination (Experiment data, blue) for five cases including CDK2i + CDK4/6i, paclitaxel + CDK2i, paclitaxel + CDK4/6i, doxorubicin + CDK2i, and doxorubicin + CDK4/6i, respectively for HCC1806 cell line. **B-C.** Drug combination predictions for each phase of the cell cycle separately for paclitaxel + 0.3 μM palbociclib at 24 hours (**B**) and 72 hours (**C**). Box plots (blue) represent the 4 replicates of the experimental data, and the yellow solid lines are the average and standard error of model predictions.

Bliss framework to the single drug rate estimations and, with that, calculated the rates corresponding to drug combinations. We additionally applied the Bliss framework directly on the total cell counts of individual drug treatments – the common approach to study drug combinations. Bliss additivity applied on the cell counts and with the rates both matched closely with the true total cell counts (**Figure 7**). Our proposed approach being slightly more accurate in higher concentrations of chemotherapy with CDK inhibitors (**Figure 7B-D**) with exception of CDK2 inhibitor and CDK4/6 inhibitor combination, which could imply that since CDK2 and CDK4/6 are not completely independent, this framework is not entirely suitable for the combination of this class of drugs. Another advancement of our approach compared to the Bliss-on-counts approach is that, with the ODE cell cycle model, we can

predict the combination effects at each cell cycle phase, separately. As shown in **Figure 7B-C** the model predictions matched perfectly with the cell cycle phase-specific counts from the experiments at both 24 hour and 72 hour timepoints. Additionally, we were able to predict the number of dead cells, without knowing the true dead cell counts from the experiment (**Figure 7B-C**).

Discussion

In this study, we sought to investigate the advantage of using cell cycle phase-specific measurements in contrast to total cell viability (**Figure 1**) through analyzing drug responses of a panel of breast cancer cell lines with respect to each cell cycle phase. We explored the existing patterns within datasets using a data-driven approach (**Figure 3**) and implemented a system of ODEs to quantitatively study the drug effects on cell cycle phases (**Figure 4**). We demonstrated that we can estimate the rate of cell death, even in the absence of experimental data from cell death counts (**Figure 4, S1**). In addition, we showed that with limited accuracy drug response data can predict breast cancer subtypes such as triple negative and genotypes such as BRCA mutation (**Figure 4-5**). To further analyze cell cycle phase-specific insights in drug combinations, we conducted an experiment including the combination of chemotherapy drugs (paclitaxel and doxorubicin) with CDK4/6 and CDK2 inhibitors (**Figure 6**). Finally, by incorporating the Bliss framework, we demonstrated the ability of our model to predict drug combinations specific to each cell cycle phase (**Figure 7**). Our analysis showed that cell cycle phase-specific measurements improve dissecting the cytostatic and cytotoxic drug responses and allow drug combination predictions with cell cycle phase-specific resolution.

Chemotherapy agents usually target the S phase of the cell cycle and disrupt DNA replication or repair [34], and the combination of chemotherapy drugs with CDK inhibitors has been the focus of recent studies [35]. It has been shown that the order of treatments with chemotherapy and CDK4/6 inhibition matters such that chemotherapy should be before

CDK4/6 inhibition [36,37]. The rationale driving this statement is that DNA damaging drugs sensitize cancer cells and this way, when CDK4/6 inhibition happens, cells would permanently exit the cell cycle into a senescent or apoptotic state. In our data of chemotherapy and CDK4/6i combinations, we observed that in specific cases such as HCC1143 cell line, the addition of 0.3 μM palbociclib to some extent antagonizes the arresting effects of doxorubicin in the S phase, and also in the G1 phase (**Figure S5A-B**). This is possibly due to the co-administration of the two drugs simultaneously, and the fact that cells existed at various phases of the cell cycle at the time of treatment (**Figure S6**). In contrast, the combination of palbociclib or PF-07104091 with paclitaxel in OVCAR3, an ovarian cancer cell line, showed a significant synergy in G1, S, and G2 arrest (**Figure S6**). In other cell lines where the combination of CDKi with doxorubicin or paclitaxel had no clear additional effect relative to chemotherapy alone, we believe that the concentration of the CDKi has not been high enough to cause any effects.

High-throughput drug screening is a resource-intensive and costly process. Computational models capable of generating reliable predictions hold immense value in reducing the need for extensive experimentation in drug discovery and combination evaluations. In this study, we have developed a mechanistic model that facilitates the analysis of large-scale drug response data and enables exhaustive testing of all potential drug combinations to identify the most promising candidates. A notable advancement of this model is its ability to provide predictions specific to different phases of the cell cycle, thereby offering insights into a drug's mechanism of action as well as potential off-target effects. Recently, novel experimental techniques such as the dye drop assay have emerged, providing exciting opportunities to gather extensive data on drug responses. The dye drop assay enables efficient and robust measurement of cell cycle progression, allowing for rapid data collection on the response of numerous cells to a wide array of drugs within a short time frame. Such data can then be utilized to train and validate mechanistic models, which subsequently aid in the identification of novel drug combinations. The integration of mechanistic modeling with high-throughput experimental

approaches holds significant promise in the realm of drug discovery.

Methods

GNE data collection

The dye drop assay, developed by Mills et al [18], was used to collect responses of 6 cell lines across a range of CDK inhibitors and chemotherapy drugs. The cell lines were seeded in 384-well black CellCarrier Ultra plates (Perkin Elmer #6057300) using a MultiDrop combi liquid dispenser (Thermo), at 50μl per well. The seeding density of each cell line is reported in **Table 2**. All cells were cultured in RPMI+ 10%FBS +2mM L-glutamine with the exception of MDA-MB-175 VII, which was cultured in DMEM + 10%FBS + 2mM L-glutamine. Four replicate plates were seeded per cell line, in addition to 20 "Day 0" control wells per cell line which were seeded on a separate plate. EdU staining was performed using the Click-iT EdU Alexa Fluor 488 HCS assay kit (Thermo #C10351), and included the LIVE/DEAD Fixable Far Red Dead Cell Stain (Thermo #L34974). Briefly, EdU (10μM final) and LIVE/DEAD stain (1:2000 final) were diluted in warm culture media and added to each well. The plates were placed in a humidified incubator at 370C / 5% CO2 for 30min. Paraformaldehyde (catalog) was then added to a final concentration of 4%, and plates were incubated for 15min at RT. Wells were then washed with PBS using a ELx405 plate washer (Bio-tek). Cells were then permeabilized with Triton-X 100 in PBS (0.2% final) at RT for 15min. Wells were then washed 2X with PBS, followed by addition of 30μl Click reaction mix, prepared as per manufacturer's protocol (Thermo #C10351). Wells were again washed 2X with PBS, followed by addition of 50μl NuclearMask Blue stain (Thermo #H10325) in PBS at 1:2000 final. Plates were sealed with foil and incubated at 4C overnight. Plates were then washed 2X with PBS, followed by a final addition of 50μl PBS. Plates were sealed with foil tape and held at 4C until imaging. Staining was then performed either 24hr or 72hr post-dosing (2 replicate plates per timepoint). The day 0 control plate was stained at time of test plate dosing. The compounds used and their concentrations is reported in **Table 3**.

Cell line	Seeding density per well
HCC1806	500
BT-549	1500
MDA-MB-468	2000
HCC1143	2500
MDA-MB-175 VII	2500
OVCAR3	3500

Table 2: Seeding density of the cell lines.

drug 1	drug 1 dose range	drug 2	drug 2 dose
PF-07104091 (CDK2i)	0.0015 - 10 μM		
palbociclib (CDK4/6i)	0.0015 - 10 μM		
abemaciclib (CDK4/6i)	0.0015 - 10 μM		
PF-06873600 (CDK2/4/6i)	0.0015 - 10 μM		
dinaciclib (pan-CDKi)	0.0015 - 10 μM		
PF-07104091 (CDK2i)	0.0015 - 10 μM	palbociclib (CDK4/6i)	0.3 μM
paclitaxel	0.00015 - 1 μM		
paclitaxel	0.00015 - 1 μM	PF-07104091	0.3 μM
paclitaxel	0.00015 - 1 μM	palbociclib	0.3 μM
doxorubicin	0.0015 - 10 μM		
doxorubicin	0.0015 - 10 μM	PF-07104091	0.3 μM
doxorubicin	0.0015 - 10 μM	palbociclib	0.3 μM

Table 3: Compounds and their concentrations used in the GNE experiment for each cell line.

Image processing and cell segmentation

Images were acquired with a Phenix Opera imaging system (Perkin Elmer) using a 20x high-NA water immersion objective in non-confocal mode. Flatfield correction, image segmentation using the Hoechst 33342 channel, and per-object quantitation of fluorescence intensity for each channel were carried out with Signals Image Artist (Perkin Elemer). Values for individual objects were exported as .csv files for further processing.

Cell gating for cell cycle phase quantification

In order to perform cell cycle gating and quantify the number of cells in different cell cycle phases for the data collected at Genentech, a dynamical gating approach was adopted. Briefly, we used a combination of total DNA intensity (Hoechst) and EdU incorporation to

discriminate cells in different phases of the cell cycle. The LDR and EdU intensities were log-transformed and smoothed using a kernel density function. To classify cells into live or dead (from the LDR), and positive or negative for the S phase, a peack finding algorithm was used to find the minima of the intensity distributions. The total DNA content is used to differentiate cells in the G1 and the G2 phase. Cells with negative EdU and intermediate DNA content were classified as "S dropout". Those cells with the DNA content lower than G1 were classified as "sub G1", and those with higher than G2 DNA content were classified as "beyond G2". **Figure S3** shows the detailed gating of cells based on the EdU intensity and DNA content. The results of cell gating were manually verified for quality assurance. The live/dead marker was not reliable to identify dead cells but allowed for proper filtering live cells.

The model

To model the cell cycle dynamics while quantifying the transition rates, we implemented a system of ordinary differential equations. The equations represent the G1 phase (G1), the S phase (S), the G2 is combined with the M phase (G2M), and the total number of dead cells (D). The constant coefficients (k_{G1-S}, k_{S-G2M}, and k_{M-G1}) refer to the transition rates from each phase to the next, and (d) refers to the cell death rate which is shared among all phases.

$$\frac{dG1}{dt} = 2k_{M-G1} \times G2M - (k_{G1-S} + d)G1 \tag{2.1}$$

$$\frac{dS}{dt} = k_{G1-S} \times G1 - (k_{S-G2M} + d)S \tag{2.2}$$

$$\frac{dG2M}{dt} = k_{S-G2} \times S - (k_{M-G1} + d)G2M \tag{2.3}$$

$$\frac{dD}{dt} = d \times (G1 + S + G2M) \tag{2.4}$$

We used matrix exponentials using Jax package in Python (version 3.11) to solve for the ODE system with the following Jacobian matrix:

$$Jac = \begin{bmatrix} -(k_{G1-S} + d) & 0 & 0 & 2k_{M-G1} \\ k_{G1-S} & -(k_{S-G2M} + d) & 0 & 0 \\ 0 & k_{S-G2} & -(k_{M-G1} + d) & 0 \\ d & d & d & 0 \end{bmatrix} \tag{2.5}$$

For each of the phases we can solve:

$$Y(t) = e^{Jac \times t} Y(0)$$

The rate parameters are assumed to follow a Hill function across the range of concentrations, C, with the following equation:

$$Hill(C) = E_{min} + \frac{E_{max} - E_{min}}{1 + (\frac{EC_{50}}{C})^k}$$

where the parameters of the model, E_{min}, E_{max}, EC_{50}, and c_k are constant coefficients defined as the effect at untreated, maximum effect, the concentration at half maximal effect, and the Hill slope, respectively. We assumed each drug can have a different effect on each cell cycle phase, hence, each phase transition rates is defined by a separate set of Hill parameters. Employing a Bayesian statistical approach, we estimate the Hill parameters for each of the cell cycle transition rates by optimizing the distance between the cell counts at each cell cycle phase and the ODE model predictions. For optimization, the Numpyro package (Python 3.11) was used where we defined a prior distribution for each Hill parameter, and fed the data, the prior distributions, and the ODE model to the optimizer which is based on the No-U-Turn Sampler (NUTS). We enforced the transition rates at each concentration to be lower or equal to the transition rates at untreated to avoid unexpected behavior as a result of drug treatment, which implies that the drug can only have an arresting effect on transitions between the cell cycle phases. This was imposed by defining a Beta distribution between

[0, 1] to serve as a fractional coefficient; with that, the prior distribution of the maximum effect (E_{max}) was constructed as a product of this coefficient and the corresponding effect at untreated (E_{min}). On the other hand, we assumed the relative cell death rate at untreated condition to be zero, such that the maximum cell death effect would represent a shift from untreated. The prior distributions used for E_{min} , EC_{50}, and c_k were LogNormal distributions to ensure positive values. The parameters used for prior distributions are summarized in Table 3. The accepted confidence interval for the parameters was determined based on the within-replicate variations per cell line. To ensure consistency, the median of the standard deviation of replicates for all conditions of a cell line was utilized. We used 3500 warm up and another 3500 samples for the fitting, with target acceptance probability 0.95%.

Parameter estimation

To estimate the posterior distribution of the transition and death rate parameters in our cell cycle model, we employed the No-U-Turn (NUTS) algorithm implemented in Numpyro [30]. The accepted confidence interval for the parameters was determined based on the within-replicate variations per cell line. To ensure consistency, the median of the standard deviation of replicates for all conditions of a cell line was utilized. The Hill parameters, representing each rate in the ODE model, were assigned a LogNormal prior distribution with initial mean and variance values presented in **Table 4**. The LogNormal distribution takes two parameters, which are the mean and variance. The Uniform distribution takes two parameters, which are the minimum and the maximum of the range of random variables. We discarded the first 3000 samples as burn-in and used an additional 3000 samples to ensure convergence.

Simultaneous fitting of all drug conditions

Since we did not have access to the dead cell numbers, we removed the last equation in the ODE model corresponding to cell death counts, and estimated the cell death rate from fitting other cell cycle phase counts present in the model. The minimum relative cell death was still

	G1 $\longrightarrow$ S	**S $\longrightarrow$ G2M**	**G2M $\longrightarrow$ G1**	**Cell death**
EC_{50}	LogNormal(0, 1)	LogNormal(0, 1)	LogNormal(0, 1)	LogNormal(0, 1)
c_k	LogNormal(0, 0.4)	LogNormal(0, 0.4)	LogNormal(0, 0.4)	LogNormal(0, 0.4)
E_{min}	LogNormal(-1, 2)	LogNormal(-1, 2)	LogNormal(-1, 2)	0
E_{max}	Uniform(0, 1) E_{min}	Uniform(0, 1.2) E_{min}	Uniform(0, 1) E_{min}	LogNormal(-1, 2)

Table 4: Parameters for the prior LogNormal distribution used in MCMC with Hill assumption.

considered zero at untreated, and each cell cycle transition rate was assumed decreasing with concentration. Additionally, we shared the parameters of the untreated condition across all drug treatments for each cell line. Meaning, we define a set of prior distributions for EC_{50}, C_k, and $_{Emax}$ per drug, but only one E_{min} that is shared among all 12 conditions. This way, at each run, we estimate 148 parameters corresponding to the 12 conditions we are fitting at the same time, as opposed to 16 parameters for the time that we fit each drug condition separately.

Drug combination

1. Directly fitting combination conditions to the model

Among the 12 conditions that we fitted to the model with the assumptions explained in the "GNE data" above, are the combinations. We compared the estimated parameters from fitting the combinations directly to the ODE model with single treatments; for instance, we compared paclitaxel alone, paclitaxel + 300 nM CDK4/6i (palbociclib), and paclitaxel + 300 nM CDK2i.

2. Predicting drug combination rates by Bliss Independence framework and single drug parameters

Using the estimated rates from single drug treatments among the 12 conditions we fitted at once, and employing the Bliss independence framework, we predicted the effects of drug combination. For each cell cycle transition rate, we first normalized it to the estimated shared untreated rate, and then for each pair of drugs, performed the Bliss independence operation,

and scaled the result back by multiplying to the untreated rate. The outcome is the expected transition rate for the combination of the drugs. Since the cell death rate is not inhibitory, and also we assumed no relative cell death at untreated, the expected combination effect from cell death rate is simply the sum of the two rates. Assume p_a to be one of the transition rates estimated for drug A, and p_b to be the same transition rate estimated for drug B, at a certain concentration. $\hat{p}_a$ and $\hat{p}_b$ are normalized to their corresponding untreated, p_0.

$$\hat{p}_a = \frac{p_a}{p_0},$$
$$\hat{p}_b = \frac{p_b}{p_0}$$

$$Bliss = (1 - (1 - \hat{p}_a) + (1 - \hat{p}_b) - \hat{p}_a \times \hat{p}_b) \times p_0 = \hat{p}_a \times \hat{p}_b \times p_0$$

3. Predicting drug combination by the classic adaptation of the Bliss Independence framework.

To compare our drug combination results with a baseline model, we used the Bliss independence applied directly on total cell counts, similar to what is usually used in the literature. To this end, we used the 4 replicates of single drug treatment data, and calculated the total cell counts for each condition. Then, the cell counts across the concentration for each condition were normalized to their corresponding untreated condition. The normalized values were then used to calculate the Bliss combination, and then scaled back to their original unit by multiplying by the untreated count. In the following, $\hat{N}_a$ and $\hat{N}_b$ are the cell counts that are normalized to the untreated cell counts, N_0. This operation was performed for each of the replicates and this way we calculated a confidence of interval shown in Figure 7.

$$\hat{N}_a = \frac{N_a}{N_0},$$
$$\hat{N}_b = \frac{N_b}{N_0}$$

$$Bliss = (1 - (1 - \hat{N}_a) + (1 - \hat{N}_b) - \hat{N}_a \times \hat{N}_b) \times N_0 = \hat{N}_a \times \hat{N}_b \times N_0$$

CP decomposition

Canonical polyadic decomposition (CP decomposition), also known as PARAFAC, was employed to analyze the multidimensional datasets. We utilized the CP decomposition developed in [20]. Briefly, the algorithm involves decomposing the data into a series of rank-one tensors, such that:

$$X \approx \sum_{r=1}^{R} \mathbf{c}_r \circ \mathbf{d}_r \circ \mathbf{p}_r, \tag{2.6}$$

where X is the original three-dimensional tensor, R is the number of components (positive integer), $r = 1, 2, ..., R$ (see **Figure 2B**), and $\circ$ represents the outer product between the vectors. The combination of the rank-one components form the factor matrices such that, for example, $C = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_R]$ that we plotted as heatmaps for easier interpretation.
To find the optimum decomposition, alternating least squares (ALS) was used while initializing the decomposition by singular value decomposition of the flattened data along each dimension.

$$\mathbf{X}_{(1)} \approx \mathbf{C}(\mathbf{P} \odot \mathbf{D})^T \tag{2.7}$$

$$\mathbf{X}_{(2)} \approx \mathbf{D}(\mathbf{P} \odot \mathbf{C})^T \tag{2.8}$$

$$\mathbf{X}_{(3)} \approx \mathbf{P}(\mathbf{D} \odot \mathbf{C})^T \tag{2.9}$$

in which $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$, and $\mathbf{X}_{(3)}$ are the flattened tensor along each axis, and $\odot$ is the Khatri-Rao product. The missing values within the data were imputed by a one-component PCA. In each ALS iteration, linear least square solving was performed on each dimension separately, such that we fix two of the dimensions and solve for the third; for example, if $\mathbf{D}$ and $\mathbf{P}$ are fixed:

$$\min_{\mathbf{C}} \|\mathbf{X}_{(1)} - \mathbf{C}(\mathbf{P} \odot \mathbf{D})^T\| \tag{2.10}$$

This is done similarly for the other two dimensions. 2000 iterations are performed.

Optimum number of components

To identify the optimal number of components for CP decomposition, we employed the core consistency diagnostic and the R2X metric. We utilized the TensorLy-viz library [38] to calculate the core consistency for a range of component numbers and selected the number of components at which the core consistency begins to decrease. Additionally, we used R2X as a measure of the model's explained variance. R2X is defined as the ratio of the sum of squares of the residual tensor to the sum of squares of the original tensor. The two matrices ensure the number of components selected does not overfit the data and also explain reasonable amount of variance within the data.

Genotype - phenotype associations

The input and output data

We used three types of input data to investigate the genotype - drug response associations: (1) The input data used in **Figure 3** from CP decomposition of cell cycle phase-specific measurements was a tensor of 54 (cell lines) $\times$ 324 (drug conditions) $\times$ 8 (phases), which was then factorized with 6 components. The cell line factor matrix was used in the logistic regression model. (2) The input data used in **Figure 3D** for PCA analysis, was the total counts of cells, which was a matrix of form 54 $\times$ 324. The matrix was normalized with respect to each cell line. The result of PCA analysis with 6 components was a matrix of 54 $\times$ 6 which was used in the logistic regression. (3) The input data used in **Figure 5** was a tensor of estimated rates with 54 $\times$ 324 $\times$ 4 (cell cycle rates), which was factorized using CP decomposition into 3 components. The cell line factor matrix 54 $\times$ 3 was used for logistic

regression.

At each case, the output data was a binary vector for each of the 7 cases. TNBC: triple negative breast cancer, HER2+: human epidermal growth factor receptor positive, HR+: hormone receptor positive, PTEN: phosphotase and TENsin homolog deleted on chromosome 10, PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha mutation, BRCA: breast cancer gene, and CDKN2A: cyclin dependent kinase inhibitor 2A (mutation in the p16 gene). We used Cellosaurus [39] portal to determine whether a cell lines has a specific mutation or not, and Dai et al [40] to determine breast cancer receptor status.

Logistic regression

For all three aforementioned cases, we performed data normalization by min max scaling using sklearn from Python 3.11. Since the prevalence of cell lines with positive receptor status or each mutation was less common, the data was imbalanced. We used SMOTE (synthetic minority oversampling technique) to balance the dataset by oversampling the rare class. In each case we used a 5-fold stratified cross validation and reported the mean of balanced accuracy for the test sets.

Data Availability Statement

The data is available upon request. For cell cycle gating, a dynamical gating approach implemented in https://github.com/datarail/DrugResponse was used.

Code Availability Statement

https://github.com/Genentech/cell_cycle_rate_model

Supplementary Material

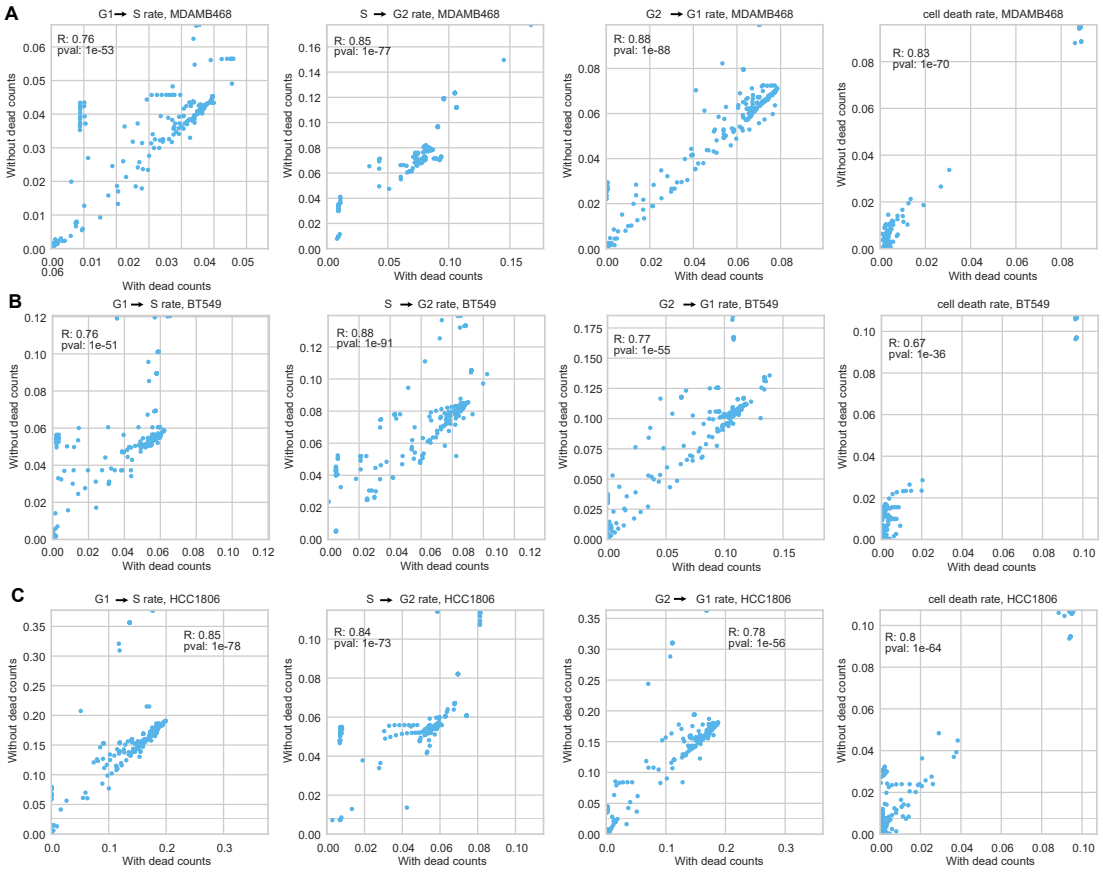


Figure S1: Spearman correlation between estimated parameters with and without using the number of dead cells from the experiment. The Spearman correlation of the G1 to S transition rate, S to G2 transition rate, G2M to G1 transition rate, and the death rate between the two aforementioned situations for MDAMB468 (**A.**), BT-549 (**B.**), and HCC1806 (**C.**), respectively. The three cell lines were selected at random from the panel of cell lines to serve as a showcase.

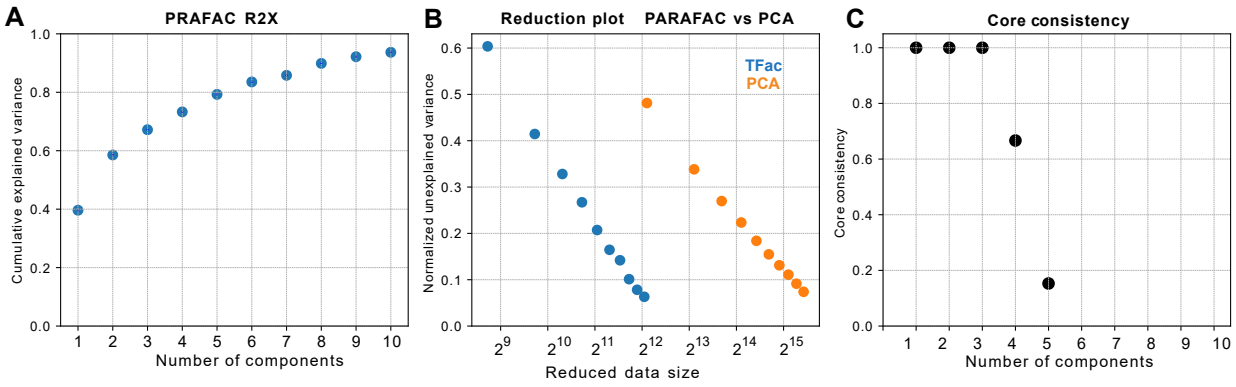


Figure S2: Decomposition of the HMS tensor of rates after fitting. A. R2X measure for up to 10 components showing cumulative explained variance. **B.** The reductions plot showing the power of data reduction of CP decomposition versus PCA. **C.** Core consistency metric to prevent overfitting.

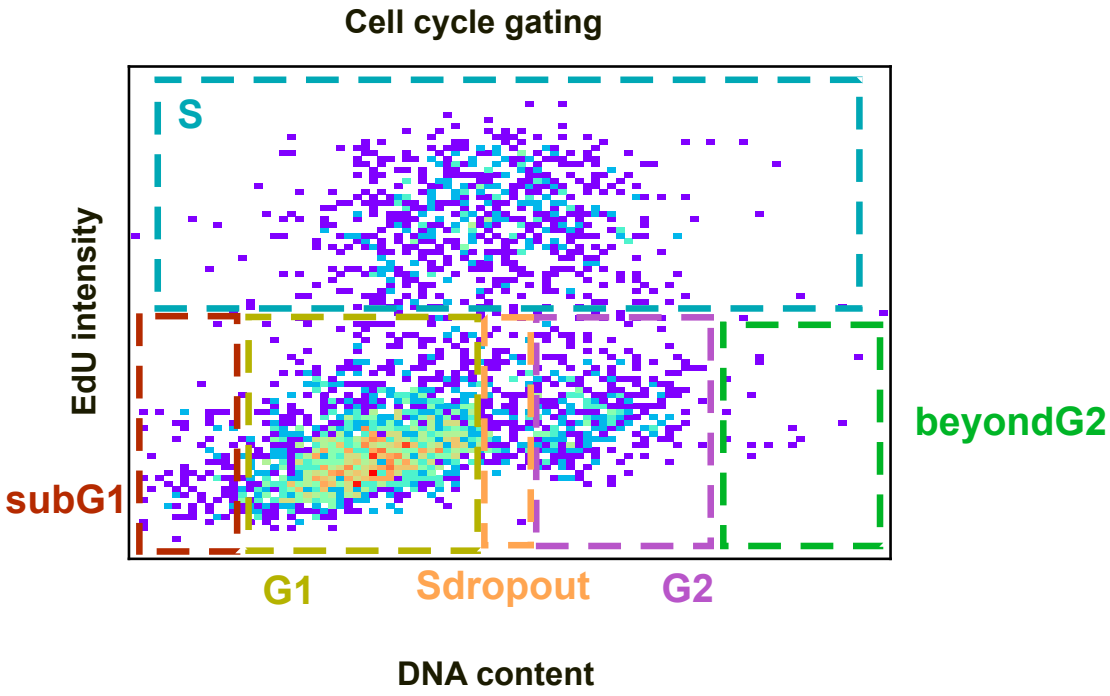


Figure S3: Cell cycle gating of the GNE dataset. Cell cycle gating with DNA concent and EdU intensity into 6 subphases, sub-G1 (low EdU, very low DNA), G1 (low EdU, low DNA), S (high EdU), G2 (low EdU, high DNA), S dropout (low EdU, medium DNA), beyond G2 (low EdU, very high DNA).

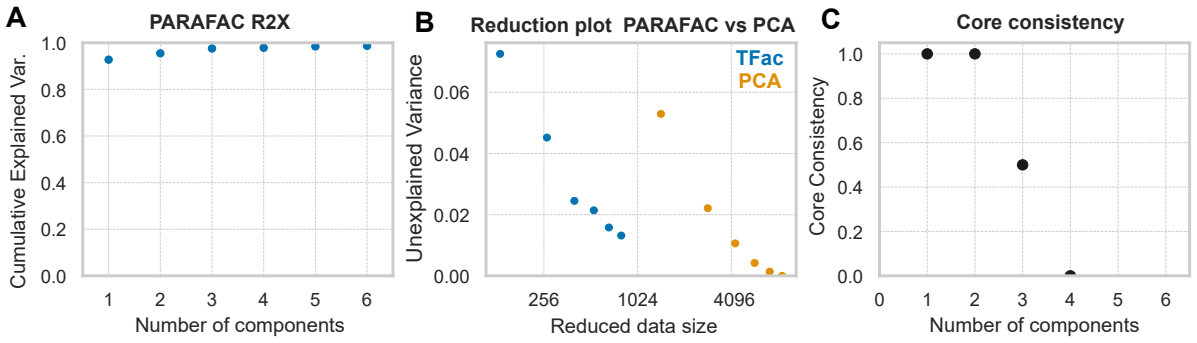


Figure S4: Decomposition of the GNE dataset. A. R2X measure for up to 6 components showing cumulative explained variance. **B.** The reductions plot showing the power of data reduction of CP decomposition versus PCA. **C.** Core consistency metric to prevent overfitting.

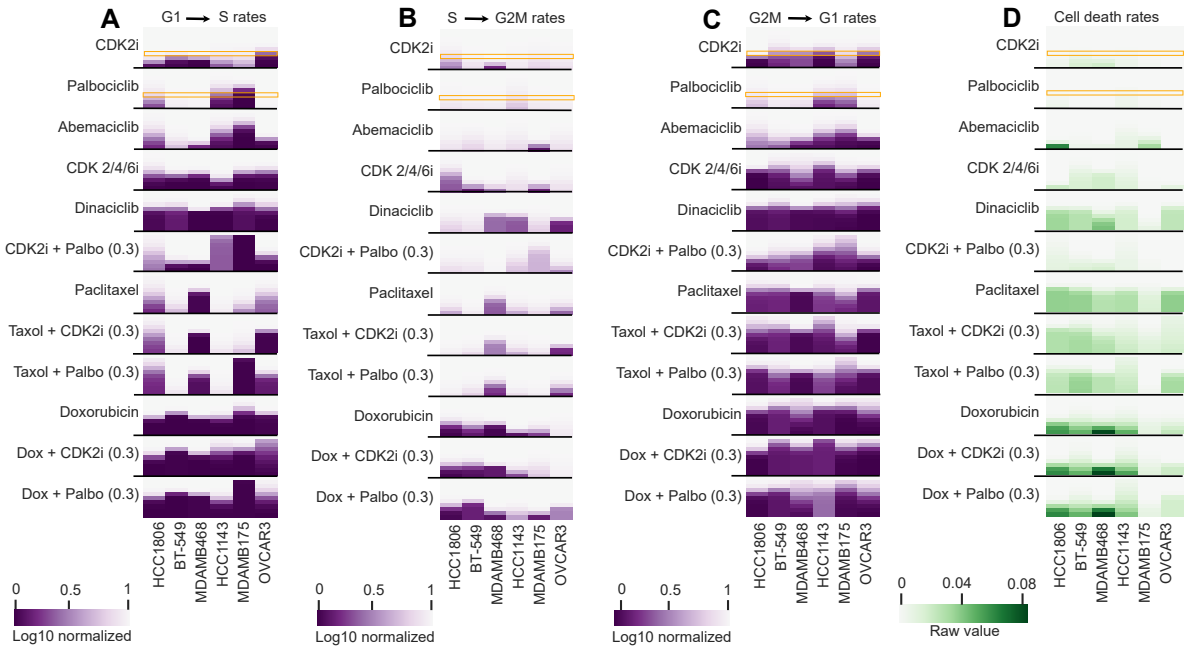


Figure S5: Comparison of estimated rates in fitting the individual and combinations to the ODE model. A-C. The estimated progression rates from G1 to S (**A**), S to G2M (**B**), and G2M to G1 (**C**) normalized to their corresponding untreated rate across all the cell lines and drug conditions. **D.** The estimated cell death rate across cell lines and conditions. Each condition is ordered with respect to concentration from untreated to the highest. 0.3 μM of palbociclib and CDK2i have been highlighted with an orange square.

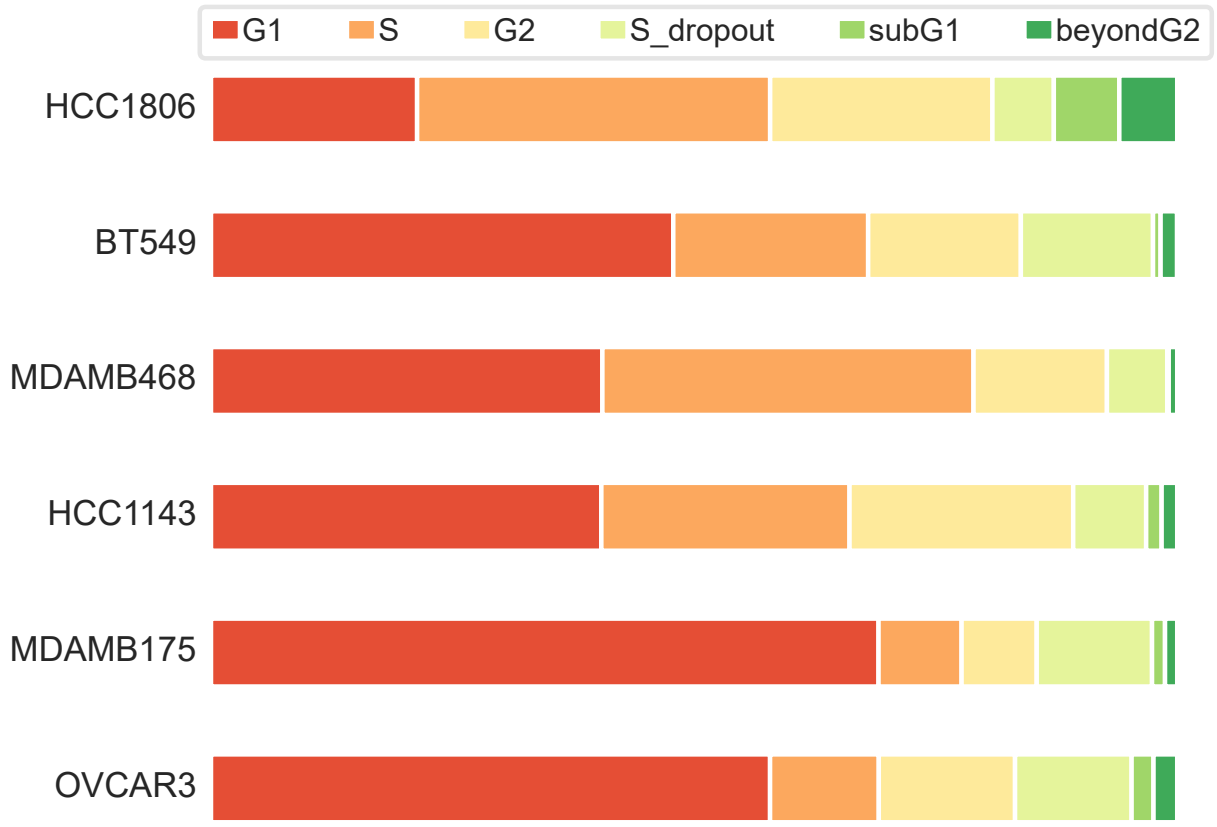


Figure S6: Comparison between the distribution of cell cycle phases at the time of drug administration for each cell line in the GNE dataset. The fraction of cells at each cell cycle phase G1 (red), S (orange), G2 (yellow), S dropout (chartruse green), sub-G1 (light green), and beyond G2 (dark green).

	mean	**SD**	**median**	n_{eff}	r_{hat}
EC_{50} G1	0.04	0.00	0.04	1592.57	1.0
EC_{50} S	0.53	0.05	0.52	2443.82	1.0
EC_{50} G2M	0.05	0.00	0.05	1897.38	1.0
EC_{50} D	9.53	6.59	8.04	1409.51	1.0
C_{Inf} coef G1	0.00	0.00	0.00	3366.56	1.0
C_{Inf} coef S	0.77	0.02	0.77	1561.36	1.0
C_{Inf} coef G2M	0.20	0.01	0.2	2020.60	1.0
C_{Inf} coef D	0.0	0.04	0.0	1769.45	1.0
C_k G1	2.49	0.1	2.48	1770.88	1.0
C_k S	7.81	1.32	7.64	2374.30	1.0
C_k G2M	2.25	0.32	2.23	2030.71	1.0
C_k D	2.07	1.14	1.79	1952.19	1.0
C_0 G1	0.01	0.00	0.01	1827.18	1.0
C_0 S	0.06	0.00	0.06	1576.12	1.0
C_0 G2M	0.04	0.00	0.04	1700.31	1.0

Table S1: Parameters estimates of MDAMB175VII treated with abemaciclib in the HMS dataset.

	mean	**SD**	**median**	n_{eff}	r_{hat}
EC_{50} G1	0.01	0.0	0.01	1975.56	1.0
EC_{50} S	5.01	0.65	4.98	1639.94	1.0
EC_{50} G2M	0.01	0.0	0.11	1924.98	1.0
EC_{50} D	14.55	5.60	14.37	950.36	1.0
C_{Inf} coef G1	0.02	0.00	0.02	1480.81	1.0
C_{Inf} coef S	1.19	0.01	1.20	1431.72	1.0
C_{Inf} coef G2M	0.12	0.01	0.12	1118.39	1.0
C_{Inf} coef D	0.01	0.03	0.01	1645.67	1.0
C_k G1	4.01	0.84	3.89	1413.21	1.0
C_k S	4.45	1.19	4.32	1377.96	1.0
C_k G2M	1.20	0.20	1.18	2007.56	1.0
C_k D	3.62	0.75	3.51	1872.48	1.0
C_0 G1	0.01	0.00	0.01	1827.18	1.0
C_0 S	0.06	0.00	0.06	1576.12	1.0
C_0 G2M	0.04	0.00	0.04	1700.31	1.0

Table S2: Parameters estimates of MDAMB175VII treated with palbociclib in the HMS dataset.

		Cell lines					
		HCC1806	BT-549	MDAMB468	HCC1143	MDAMB175	OVCAR3
CDK2i	EC_{50} G1	1.76	0.89	1.0	3.51	2.46	0.21
	EC_{50} S	0.87	0.91	0.06	0.001	0.04	11.63
	EC_{50} G2M	1.06	0.32	0.31	0.17	0.28	0.23
	EC_{50} D	2.76	3.19	1.78	1.20	2.29	3.6
palbociclib	EC_{50} G1	2.98	1.34	3.09	0.2	0.06	3.33
	EC_{50} S	2.91	3.68	0.07	0.88	0.15	0.74
	EC_{50} G2M	1.06	0.32	0.31	0.17	0.28	0.23
	EC_{50} D	1.52	0.02	4.83	0.06	2.1	4.06

Table S3: EC50s estimated for cell cycle rates of cell lines across drug treatments from the GNE dataset.

Bibliography

[1] Paul Dent, Yong Tang, Adly Yacoub, Yun Dai, Paul B Fisher, and Steven Grant. Chk1 inhibitors in combination chemotherapy: thinking beyond the cell cycle. *Molecular interventions*, 11(2):133, 2011.

[2] Shom Goel, Molly J DeCristo, April C Watt, Haley BrinJones, Jaclyn Sceneay, Ben B Li, Naveed Khan, Jessalyn M Ubellacker, Shaozhen Xie, Otto Metzger-Filho, et al. Cdk4/6 inhibition triggers anti-tumour immunity. *Nature*, 548(7668):471–475, 2017.

[3] Nischal Koirala, Nandini Dey, Jennifer Aske, and Pradip De. Targeting cell cycle progression in her2+ breast cancer: An emerging treatment opportunity. *International Journal of Molecular Sciences*, 23(12):6547, 2022.

[4] Andrea Rocca, Alberto Farolfi, Sara Bravaccini, Alessio Schirone, and Dino Amadori. Palbociclib (pd 0332991): targeting the cell cycle machinery in breast cancer. *Expert opinion on pharmacotherapy*, 15(3):407–420, 2014.

[5] Lie Yuan, Yongqing Cai, Liang Zhang, Sijia Liu, Pan Li, and Xiaoli Li. Promoting apoptosis, a promising way to treat breast cancer with natural products: A comprehensive review. *Frontiers in Pharmacology*, 12:3878, 2022.

[6] Nadine M Tung, Mark E Robson, Steffen Ventz, Cesar Augusto Santa-Maria, Paul Kelly Marcom, Rita Nanda, Payal D Shah, Tarah Jean Ballinger, Eddy Shih-Hsin Yang, Michelle E Melisko, et al. Tbcrc 048: A phase ii study of olaparib monotherapy in

metastatic breast cancer patients with germline or somatic mutations in dna damage response (ddr) pathway genes (olaparib expanded)., 2020.

[7] Sonia Pernas, Sara M Tolaney, Eric P Winer, and Shom Goel. Cdk4/6 inhibition in breast cancer: current practice and future directions. *Therapeutic advances in medical oncology*, 10:1758835918786451, 2018.

[8] Caroline F Thorn, Connie Oshiro, Sharon Marsh, Tina Hernandez-Boussard, Howard McLeod, Teri E Klein, and Russ B Altman. Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenetics and genomics*, 21(7):440, 2011.

[9] Prson Gautam, Leena Karhinen, Agnieszka Szwajda, Sawan Kumar Jha, Bhagwan Yadav, Tero Aittokallio, and Krister Wennerberg. Identification of selective cytotoxic and synthetic lethal drug responses in triple negative breast cancer cells. *Molecular cancer*, 15(1):1–16, 2016.

[10] Jeffrey W Tyner, Cristina E Tognon, Daniel Bottomly, Beth Wilmot, Stephen E Kurtz, Samantha L Savage, Nicola Long, Anna Reister Schultz, Elie Traer, Melissa Abel, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*, 562(7728):526–531, 2018.

[11] Helen K Matthews, Cosetta Bertoli, and Robertus AM de Bruin. Cell cycle control in cancer. *Nature Reviews Molecular Cell Biology*, 23(1):74–88, 2022.

[12] Xueqian Gong, Lacey M Litchfield, Yue Webster, Li-Chun Chio, Swee Seong Wong, Trent R Stewart, Michele Dowless, Jack Dempsey, Yi Zeng, Raquel Torres, et al. Genomic aberrations that activate d-type cyclins are associated with enhanced sensitivity to the cdk4 and cdk6 inhibitor abemaciclib. *Cancer cell*, 32(6):761–776, 2017.

[13] Richard S Finn, Judy Dering, Dylan Conklin, Ondrej Kalous, David J Cohen, Amrita J Desai, Charles Ginther, Mohammad Atefi, Isan Chen, Camilla Fowst, et al. Pd 0332991, a selective cyclin d kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal

estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Research*, 11(5):1–13, 2009.

[14] Nicholas C Turner, Yuan Liu, Zhou Zhu, Sherene Loi, Marco Colleoni, Sibylle Loibl, Angela DeMichele, Nadia Harbeck, Fabrice André, Mohamed Amine Bayar, et al. Cyclin e1 expression and palbociclib efficacy in previously treated hormone receptor–positive metastatic breast cancer. *Journal of Clinical Oncology*, 37(14):1169, 2019.

[15] Mario Niepel, Marc Hafner, Mirra Chung, and Peter K Sorger. Measuring cancer drug sensitivity and resistance in cultured cells. *Current protocols in chemical biology*, 9(2):55–74, 2017.

[16] Sean M Gross, Farnaz Mohammadi, Crystal Sanchez-Aguila, Paulina J Zhan, Aaron S Meyer, and Laura M Heiser. Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *bioRxiv*, pages 2020–07, 2020.

[17] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

[18] Caitlin E Mills, Kartik Subramanian, Marc Hafner, Mario Niepel, Luca Gerosa, Mirra Chung, Chiara Victor, Benjamin Gaudio, Clarence Yapp, Ajit J Nirmal, et al. Multiplexed and reproducible high content screening of live and fixed cells using dye drop. *Nature Communications*, 13(1):6918, 2022.

[19] Jia-Ren Lin, Mohammad Fallahi-Sichani, Jia-Yun Chen, and Peter K Sorger. Cyclic immunofluorescence (cycif), a highly multiplexed method for single-cell imaging. *Current protocols in chemical biology*, 8(4):251–264, 2016.

[20] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[21] Lei Zhong, Yueshan Li, Liang Xiong, Wenjing Wang, Ming Wu, Ting Yuan, Wei Yang, Chenyu Tian, Zhuang Miao, Tianqi Wang, et al. Small molecules in targeted cancer therapy: Advances, challenges, and future perspectives. *Signal transduction and targeted therapy*, 6(1):201, 2021.

[22] Angela DeMichele, Amy S Clark, Kay See Tan, Daniel F Heitjan, Kristi Gramlich, Maryann Gallagher, Priti Lal, Michael Feldman, Paul Zhang, Christopher Colameco, et al. Cdk 4/6 inhibitor palbociclib (pd0332991) in rb+ advanced breast cancer: phase ii activity, safety, and predictive biomarker assessment. *Clinical Cancer Research*, 21(5):995–1001, 2015.

[23] Maura N Dickler, Sara M Tolaney, Hope S Rugo, Javier Cortes, Veronique Dieras, Debra A Patt, Hans Wildiers, Martin Frenzel, Andrew Koustenis, and Jose Baselga. Monarch1: Results from a phase ii study of abemaciclib, a cdk4 and cdk6 inhibitor, as monotherapy, in patients with hr+/her2-breast cancer, after chemotherapy for advanced disease., 2016.

[24] Anand Shah, Erik Bloomquist, Shenghui Tang, Wentao Fu, Youwei Bi, Qi Liu, Jingyu Yu, Ping Zhao, Todd R Palmby, Kirsten B Goldberg, et al. Fda approval: ribociclib for the treatment of postmenopausal women with hormone receptor–positive, her2-negative advanced or metastatic breast cancer. *Clinical Cancer Research*, 24(13):2999–3004, 2018.

[25] Marc Creixell, Hyuna Kim, Farnaz Mohammadi, Shelly R Peyton, and Aaron S Meyer. Systems approaches to uncovering the contribution of environment-mediated drug resistance. *Current Opinion in Solid State and Materials Science*, 26(5):101005, 2022.

[26] Pepijn M Schoonen, Francien Talens, Colin Stok, Ewa Gogola, Anne Margriet Heijink, Peter Bouwman, Floris Foijer, Madalena Tarsounas, Sohvi Blatter, Jos Jonkers, et al. Progression through mitosis promotes parp inhibitor-induced cytotoxicity in homologous recombination-deficient cancer cells. *Nature communications*, 8(1):15981, 2017.

[27] Marc Hafner, Caitlin E Mills, Kartik Subramanian, Chen Chen, Mirra Chung, Sarah A Boswell, Robert A Everley, Changchang Liu, Charlotte S Walmsley, Dejan Juric, et al. Multiomics profiling establishes the polypharmacology of fda-approved cdk4/6 inhibitors and the potential for differential clinical activity. *Cell chemical biology*, 26(8):1067–1080, 2019.

[28] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[29] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[30] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[31] Kevin D Freeman-Cook, Robert L Hoffman, Douglas C Behenna, Britton Boras, Jordan Carelli, Wade Diehl, Rose Ann Ferre, You-Ai He, Andrea Hui, Buwen Huang, et al. Discovery of pf-06873600, a cdk2/4/6 inhibitor for the treatment of cancer. *Journal of Medicinal Chemistry*, 64(13):9056–9077, 2021.

[32] Kevin D Freeman-Cook, Robert L Hoffman, Douglas C Behenna, Britton Boras, Jordan Carelli, Wade Diehl, Rose Ann Ferre, You-Ai He, Andrea Hui, Buwen Huang, et al. Discovery of pf-06873600, a cdk2/4/6 inhibitor for the treatment of cancer. *Journal of Medicinal Chemistry*, 64(13):9056–9077, 2021.

[33] Mallavarapu R Srividya, Balaram Thota, Bangalore C Shailaja, Arimappamagan Arivazhagan, Kandavel Thennarasu, Bangalore A Chandramouli, Alangar S Hegde, and Vani Santosh. Homozygous 10q23/pten deletion and its impact on outcome in glioblas-

toma: a prospective translational study on a uniformly treated cohort of adult patients. *Neuropathology*, 31(4):376–383, 2011.

[34] Daniel L Gustafson and Rodney L Page. Cancer chemotherapy. *Withrow and MacEwen's small animal clinical oncology*, pages 157–179, 2013.

[35] Christopher C Mills, EA Kolb, and Valerie B Sampson. Development of chemotherapy with cell-cycle inhibitors for adult and pediatric cancer therapycombination therapies for cancer. *Cancer research*, 78(2):320–325, 2018.

[36] Anne Fassl and Piotr Sicinski. Chemotherapy and cdk4/6 inhibition in cancer treatment: Timing is everything. *Cancer Cell*, 37(3):265–267, 2020.

[37] Beatriz Salvador-Barbero, Mónica Álvarez-Fernández, Elisabet Zapatero-Solana, Aicha El Bakkali, María del Camino Menéndez, Pedro P López-Casas, Tomas Di Domenico, Tao Xie, Todd VanArsdale, David J Shields, et al. Cdk4/6 inhibitors impair recovery from cytotoxic chemotherapy in pancreatic adenocarcinoma. *Cancer cell*, 37(3):340–353, 2020.

[38] Marie Roald and Yngve Mardal Moe. Tlviz: Visualising and analysing tensor decomposition models with python. *Journal of Open Source Software*, 7(79):4754, 2022.

[39] Amos Bairoch. The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT*, 29(2):25, 2018.

[40] Xiaofeng Dai, Hongye Cheng, Zhonghu Bai, and Jia Li. Breast cancer cell line classification and its relevance with breast tumor subtyping. *Journal of Cancer*, 8(16):3131, 2017.

Chapter 3

A Lineage Tree-Based Hidden Markov Model Quantifies Cellular Heterogeneity and Plasticity

Farnaz Mohammadi, Shakthi Visagan, Sean Gross, Luka Karginov, J.C. Lagarde, Laura M. Heiser, Aaron S. Meyer

communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-04208-9 OPEN

A lineage tree-based hidden Markov model quantifies cellular heterogeneity and plasticity

Farnaz Mohammadi[1], Shakthi Visagan[1], Sean M. Gross[2], Luka Karginov[3], J. C. Lagarde[1], Laura M. Heiser[2] & Aaron S. Meyer[1,4,5,6]✉

Individual cells can assume a variety of molecular and phenotypic states and recent studies indicate that cells can rapidly adapt in response to therapeutic stress. Such phenotypic plasticity may confer resistance, but also presents opportunities to identify molecular programs that could be targeted for therapeutic benefit. Approaches to quantify tumor-drug responses typically focus on snapshot, population-level measurements. While informative, these methods lack lineage and temporal information, which are particularly critical for understanding dynamic processes such as cell state switching. As new technologies have become available to measure lineage relationships, modeling approaches will be needed to identify the forms of cell-to-cell heterogeneity present in these data. Here we apply a lineage tree-based adaptation of a hidden Markov model that employs single cell lineages as input to learn the characteristic patterns of phenotypic heterogeneity and state transitions. In benchmarking studies, we demonstrated that the model successfully classifies cells within experimentally-tractable dataset sizes. As an application, we analyzed experimental measurements in cancer and non-cancer cell populations under various treatments. We find evidence of multiple phenotypically distinct states, with considerable heterogeneity and unique drug responses. In total, this framework allows for the flexible modeling of single cell heterogeneity across lineages to quantify, understand, and control cell state switching.

[1] Department of Bioengineering, University of California, Los Angeles, CA, USA. [2] Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA. [3] Department of Bioengineering, University of Illinois, Urbana Champaign, IL, USA. [4] Department of Bioinformatics, University of California, Los Angeles, CA, USA. [5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, USA. [6] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, CA, USA. ✉email: ameyer@ucla.edu

Chemotherapy and targeted therapies selectively eliminate fast-proliferating or oncogene-addicted cells and are among the primary treatments for cancer. However, long-term therapeutic efficacy is inevitably limited by widespread intratumoral heterogeneity[1,2]. Cell-to-cell variability in drug response can originate from cell-intrinsic factors—such as genomic alterations, epigenetic mechanisms like changes in chromatin state[3], and variable protein levels[4,5]—or cell-extrinsic factors such as spatial variability in the surrounding vasculature and environmental stressors[6–8]. Moreover, cell plasticity, where cells adopt new characteristics such as those of other cell types, is observed in cancer cells, and can affect their sensitivity to therapy[9].

Large-scale profiling studies can find molecular features that associate with drug response using population-level samples[10,11]. These associations, while valuable, can miss the contribution of cell-to-cell heterogeneity, and especially stochastic changes in individual cell states that compound to effects on overall tumor drug response[3,12,13]. The most common methods for quantifying drug response are metrics of tumor cell population expansion or contraction[14–17]. Recent research has made efforts to track phenotypic measurements of fitness at the single-cell level[18,19], however, even single-cell measurements are typically performed with snapshots that subsequently miss the role of individual cells in the overall population response[20]. Though population heterogeneity is usually defined through molecular measurements, studies that have explicitly linked molecular and phenotypic variation have been able to identify mechanisms that underly cell-to-cell variation that would otherwise remain hidden[21], and studies starting with phenotypic analysis have generally found that phenotypic variability arises from a small number of molecular factors leading to the phenotypic variation[4,22,23].

Measurements accompanied by lineage relationships are uniquely valuable for studying inherited phenotypes within families of individuals. This value is evident in linkage studies wherein relatives are used to identify or refine the genetic determinants of disease[24–26]. Notably, linkage studies can identify genetic determinants with greater power than even large association studies because relatives essentially serve as internal controls[27]. Linkage studies also start with the phenotype of individuals, rather than grouping based on molecular differences, ensuring discoveries are phenotypically consequential. While the inherited factors are different between cells (e.g., proteins, RNA) and people (DNA), such approaches are likely to be similarly useful with populations of cells. Recently, constructing phylogenic trees of cancer cells using lineage tracing and single-cell sequencing has helped to characterize the directionality of metastatic seeding, though these methods are limited to tracking slow processes such as mutational differences[28]. Lineage-resolved data has also demonstrated value in uncovering cell-to-cell heterogeneity due to transient differences outside of cancer[22,23]. Therefore, tools to analyze and explore these data will be critical to uncovering new forms and sources of cell-to-cell variation.

Hidden Markov models (HMMs) provide an efficient strategy to infer discrete states from measurements when a series of co-dependent observations are made. An example of this is their widespread use in time series analysis, where each measurement is dependent on those that came before[29,30]. Recognizing this co-dependence allows HMMs to make accurate inferences even in the presence of extremely noisy measurements since each neighboring measurement provides accumulating evidence[31]. These models derive their relative simplicity by assuming a Markov process, meaning that the current behavior of a system can be assumed to be independent of its earlier history should its current state be known. This assumption naturally applies in many contexts. In the case of cells, this assumption aptly captures cell inheritance because daughter cells inherit both molecular signals and their environment from their predecessor. Indeed, several recent examples of cell-to-cell inheritance mechanisms can be represented as a Markov process through linear chains or cycles of states[12,22,23]. HMMs have been adapted to lineage trees (tHMMs) so that each measurement across the tree can similarly provide accumulating evidence for a prediction. Just like with time-series data, these models can provide very accurate predictions despite noisy measurements and limited information by recognizing the co-dependence between related measurements[32,33]. tHMMs have been used in a multitude of applications, from image classification to comparative genomics[34,35]. These models have been fit to lineages collected from stem cells and bacteria colonies, but have always required custom implementations[36,37]. Improvements in cell tracking and high-throughput imaging promise to make tHMM models valuable techniques for studying the plasticity of heterogeneous cell populations. However, widespread use of these models still depends on more easily usable implementations, examples of successful tHMM-based discoveries, and standards for experimental application.

Here, we develop an extensible implementation of tHMMs with a defined interface for integrating diverse types of measurements on cell lineage trees. This model allows us to quantify the dynamics and phenotypic features of drug response heterogeneity. We leveraged information about the relationships between cells to analyze the cell cycle responses of populations of breast cancer cells to a panel of therapies, and how normal breast cells respond to growth factor treatment. Single-cell measurements of the cell cycle revealed extensive variation not captured by population-level measurement. Using the tHMM model, we inferred the number of phenotypically distinct subpopulations, the characteristics of those subpopulations, the transition probabilities from one state to another, and each cell's expected state. We also confirmed that the tHMM model could use patterns of inheritance to predict cell behavior. This work, therefore, provides a flexible phenotype-driven route to discovering cell-to-cell variation in drug response, demonstrates an overall strategy for quantifying the dynamics of cell heterogeneity, and implements a very general software tool for the widespread use of tHMM models.

Results

Lineage information provides unique information about the source and structure of cell-to-cell heterogeneity. Single cells grow and then divide into two daughter cells, eventually forming a binary genealogical tree, also known as a lineage tree. We collected single-cell measurements in the form of lineage trees to track these relationships. The life cycle of each cell before division includes G1, S, G2, and M phases that must pass one after another. To illustrate the unique value of lineage measurements in analyzing intra-tumoral and drug response heterogeneity, we collected cell fates (whether cells ultimately divide or die) alongside either cell lifetimes (MCF10A) or individual cycle phase durations (AU565). Two random subsets of the tracked lineages of the breast cancer cell line AU565 are plotted in Fig. 1a. The single cell lineages reveal striking variation in cell cycle phase durations and cell division dynamics despite coming from the same sample. Population-level measurements would be unable to identify this difference as the starting and ending cell numbers are the same. Measurements that record or reflect the history of cells (e.g., CFSE staining, Luria-Delbruck experiment) can help to identify these variations within cell populations but must make assumptions about the dynamics of heterogeneity[13,23]. Lineage measurements, by contrast, provide sufficiently rich temporal

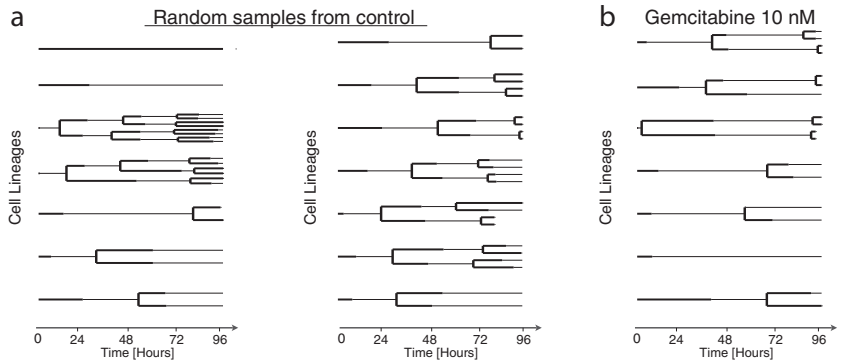


Fig. 1 Total cell number is insufficient to distinguish the structure of heterogeneous populations. a Randomly sampled lineages of untreated AU565 cells from the same replicate and experiment. **b** Randomly sampled lineages of AU565 cells treated with 5 nM gemcitabine from a single replicate and experiment. Each line indicates the lifetime of one cell. A line branching into two lines indicates cell division. The G1 and S/G2 phase durations are indicated by solid thick and thin lines, respectively.

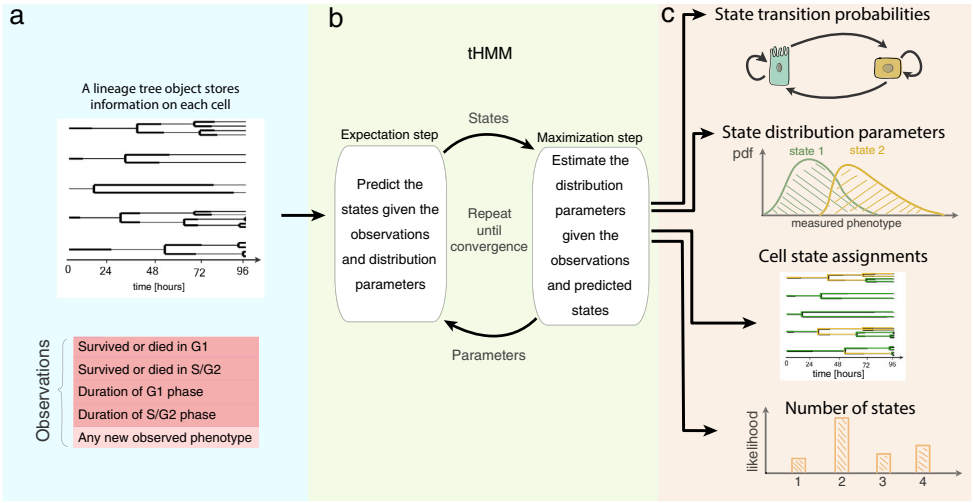


Fig. 2 The tHMM model. a Input data takes the form of single-cell measurements across time, where the lineage relationship between cells is known. **b** The fitting process includes expectation and maximization steps, where model parameters are iteratively updated until convergence. **c** Output predictions of the model after fitting including the tree of inferred cell states, probabilities of transition between each state, starting abundance of each cell state, and distributions that describe the behavior of cells within each state. The model likelihood can be used to estimate the number of distinguishable cell states.

information to quantify the specific structure of the phenotypic heterogeneity.

As further exploration of the cell tracking data, we randomly sampled lineages from gemcitabine-treated AU565 cells (Fig. 1b). Gemcitabine is a chemotherapy agent that disrupts DNA replication and results in the extension of and apoptosis in S phase[38]. We found that the S/G2 phase lengths in treated cells were noticeably extended compared to untreated cells, slowing population growth. There was generally striking variation between lineages of a single condition, including anywhere from zero to three cell divisions, but tightly shared behavior among cells and their relatives in each lineage. These observations demonstrate some of the unique advantages of collecting lineage-based measurements.

A lineage tree-based hidden Markov model infers the state of cells given measurements on lineage trees. Given the unique insights that single-cell measurements on lineage trees can provide, we implemented a strategy for classifying cells based on their phenotype and lineage relationships. We used a tree-based hidden Markov model (tHMM) to fit a set of measurements made across a lineage tree (Fig. 2a). Like a typical hidden Markov model, a tHMM can infer the hidden discrete "states" of cells given a series of measurements where a state is defined by specific phenotype distributions. The inference of these states takes place using an iterative strategy wherein the states of each cell are predicted by the phenotype of both the cell and its relatives in a lineage (expectation step), and then each distribution of phenotypes is fit to match the cells within that state (maximization step) (Fig. 2b). This expectation-maximization (EM) process repeats until convergence.

After fitting, the model can provide a variety of information (Fig. 2c). First, it infers the starting and transition probabilities of each state. Second, the distribution of cells' phenotypes in each state are estimated and can be compared to distinguish how cells of each state behave. For instance, if we use the growth rates of cells as their

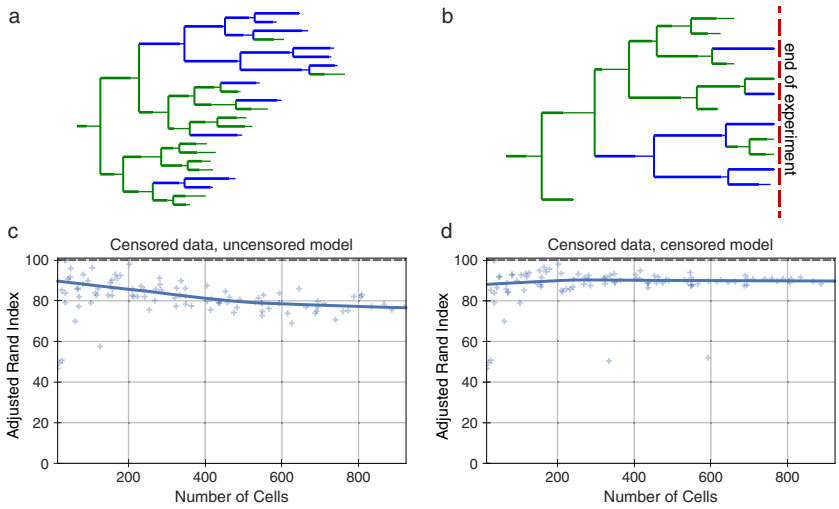


Fig. 3 Experiments of finite time necessitate data censorship corrections. a An example synthetic, uncensored two-state lineage. **b** An example synthetic, censored two-state lineage. Cells in state 0 and 1 are shown in green and blue, respectively. **c** State assignment accuracy with censored lineages using an uncorrected model. **d** State assignment accuracy with censored lineages using the corrected model. Each scatter point represents the state assignment accuracy of the model when fit to a lineage with the indicated number of cells. The solid lines show the Lowess trendline of the individual run accuracies. 100 trials are plotted. The accuracy of state assignment is measured by the adjusted rand index, a similarity measure that can serve as an accuracy for unlabeled clustering problems.

phenotype, we may observe a subpopulation of cells with shorter times to division, and another with longer times. Moreover, the state of each individual cell can be predicted from the fit data or new measurements. Finally, the model provides a likelihood of each cell's observations and therefore the data overall. This last quantity can be used, for example, to estimate the number of distinguishable cell states. When implementing these processes, we ensured that a cell's measurements were defined through a modular interface, allowing many other forms of data to be easily integrated, such as cell morphology or molecular measurements.

Experiments of finite time necessitate corrections for experimental censorship. Modeling the duration of each cell's lifetime is complicated by the influence of experimental parameters. Specifically, cells measured at the beginning or end of an experiment persist beyond the experiment's duration and so, while we observe these cells, we do not know their exact lifetimes. Data censorship occurs when a measurement is systematically affected by an undesired influence. For instance, in our case, phase durations are censored because the experiment started after cells had already begun their initial cell cycle phase or the experiment ended before they had completed their last phase. Previously, this has been addressed by removing incompletely observed cells[22]. However, doing so results in a systematic bias, where longer-lived cells are preferentially eliminated. On the other hand, ignoring the truncation of these values also creates bias by creating an upper bound on the cells' lifetimes (Fig. 3b, c).

To correct for this effect in our model, we marked cells that encountered the start or end bounds of the experiment. When estimating the properties of these cells' lifetime we instead used a censored estimator or the survival function of the distribution[39]. Because the labels are interchangeable in our classification, we used the adjusted rand index[40], a similarity measure that can serve as an accuracy measure for clustering results. Using synthetic data, we verified that this correction resulted in accurate phenotype estimations (Fig. 3d, Supplementary Figs. 3, 10). Thus, accounting for cells that outlive the bounds of the experiment through a censored estimator removes the contribution of this experimental confounder.

Synthetic lineage benchmarks show a tHMM can accurately infer population behavior. To evaluate how accurately a tHMM model could infer the behavior of multi-state cell populations, we used synthetic populations of cells in a wide variety of configurations, such as various populations sizes, numbers of states, and abundance of the states. In each case, we determined that the tHMM model could accurately infer the hidden states and parameters of a population given at least 100 cells. This synthetic data included uncensored (Supplementary Figs. 1, 2, 8, 9; Supplementary Tables 1, 2) or censored (Supplementary Figs. 4, 10, 3, 15; Supplementary Tables 1–3) situations. Synthetic data were created by lengthening the simulated experiment time, in effect creating deeper lineages, or by increasing the number of initial cells to have a greater number of lineages, increasing the experiment's breadth. In addition to varying the number of cells in a population, we benchmarked populations with varied cell state percentages (Supplementary Figs. 4, 5) and varied the degree of phenotypic differences between states (Supplementary Figs. 6, 7; Fig. 5). This benchmarking consistently showed that the tHMM model would provide accurate results across a range of circumstances, and generally provided accurate results with datasets consisting of at least 10 lineages, 100 cells overall, and 10 cells from each state.

More specifically, one of the benchmarking studies we performed was with data matching our measurements of AU565, where G1 and S/G2 phase durations were represented by a gamma distribution, and their corresponding cell fate represented by a Bernoulli distribution (Fig. 4). The choice of the gamma distribution for cell cycle phase was inspired by a previous study[41] and verified by evaluating a variety of distributions; the gamma distribution fit the cell lifetime data best. Although the tHMM model was fit with no information about the true underlying parameters of the simulated cells, it distinguished the pre-assigned two underlying cell states'

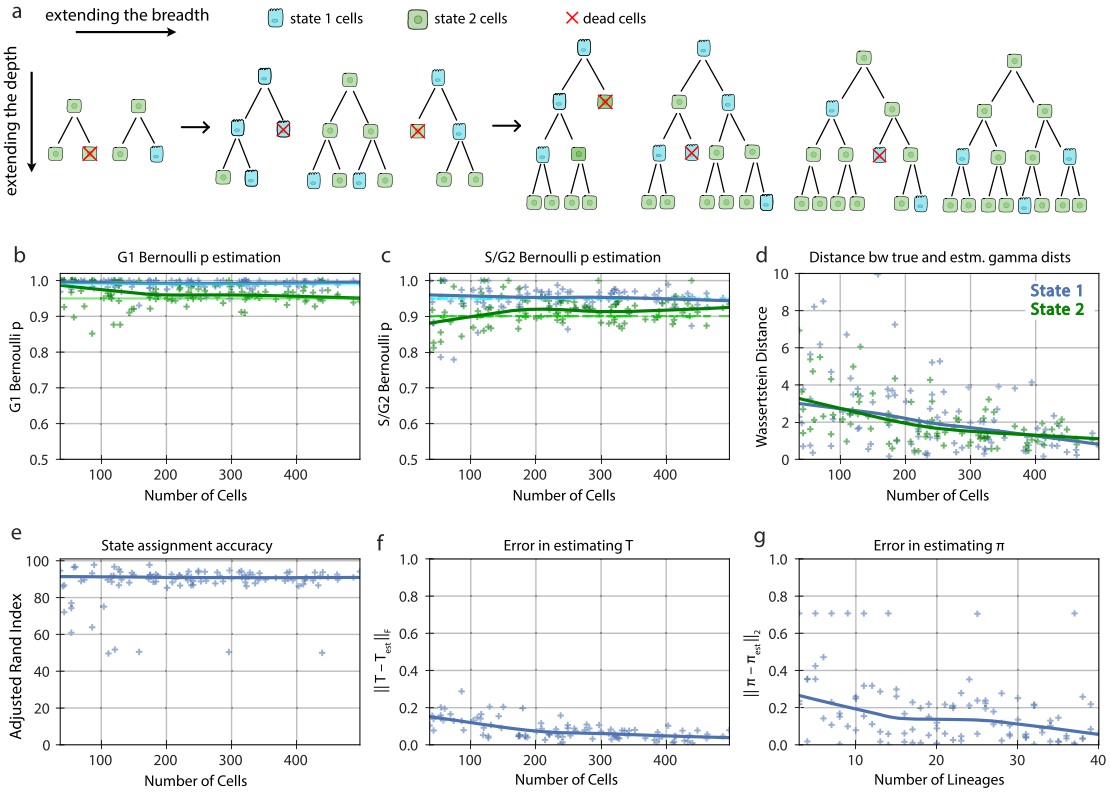


Fig. 4 Model performance on censored lineages of two states with increasing breadth and depth. a Synthetic two-state populations of increasing breadth (increasing number of initial cells and therefore lineages) and of increasing depth (increasing experiment time and therefore more cells in each lineage) are analyzed. The states are shown as green and blue colors. Red indicates cell death. **b**, **c** The accuracy of estimating the Bernoulli parameters for G1 and S/G2 phase, respectively. Each point in the scatter plots represents the inferred value for a model evaluation trial with the number of cells shown in the x-axis. The dark solid lines are the Lowess trendline across the individual trials. The light green and light blue lines show the true value of the parameters. **d** The distance between the true and estimated gamma distributions associated with phase lengths for the two states. **e** The state assignment accuracy. **f** The errors in the estimated and transition rate matrices. **g** The initial probability vector. Note that the Wasserstein distance between the true and estimated distributions for each state is much lower than the distance between two distributions that are quite similar (Fig. 5b). 100 simulation trials are plotted.

phenotypes (Fig. 4b–d) and member cells with >95% accuracy (Fig. 4e). The Wasserstein distance metric was used to quantify the difference between the true and estimated cell cycle phase duration distributions to show the accuracy of parameter estimation (Fig. 4d). On the population level, the difference between the true and estimated transition probabilities, as calculated by the sum of squared difference, was less than 0.1 for 100 cells or more. Starting probabilities were compared to their corresponding true values using the Euclidean distance and showed less than a 0.2 error for populations with 10 lineages or more (Fig. 4f, g). Thus, we are confident that with similar experimental data, we should derive accurate results.

Lineage information improves cell state identification with heritable phenotypes. Cells of even very distinct molecular states can have partly overlapping phenotypes due to non-heritable variation. Therefore, we sought to evaluate how different two states need to be for us to accurately identify them as distinct (Fig. 5a). We varied the G1 phase duration of two states from identical to very distinct (Fig. 5b) and quantified the state assignment accuracy of our model (Fig. 5c). While the phenotypic observation of a given state had to be different for our model to accurately assign cells, even moderately overlapping phenotypes (Wasserstein distance of ~20) could be distinguished by using the lineage relationships of cells. As a baseline comparison, we analytically identified the optimal classifier in the absence of lineage information (see Methods). The tHMM consistently outperformed this approach (Fig. 5c). The model performance in censored and uncensored populations was similar (Supplementary Figs. 6, 7). This shows that lineage relationships can be used to identify cell states with partially overlapping phenotypes more accurately.

Likelihood-based model selection can effectively identify the number of distinct states. One does not usually know the number of distinct cell states within a population. Further, the number of distinct states may vary depending upon the environmental context of the cells, particularly for phenotypic measurements[42,43]. To test whether we could infer the number of phenotypically distinct states, we performed model selection using the Bayesian information criterion (BIC) while varying the number of states in synthetic data (Fig. 6). We normalized the BIC values such that zero corresponds to the state with the highest likelihood. The synthetic populations included approximately 250 to 650 cells with known cell phase fate and phase lengths (Supplementary Table 3). The

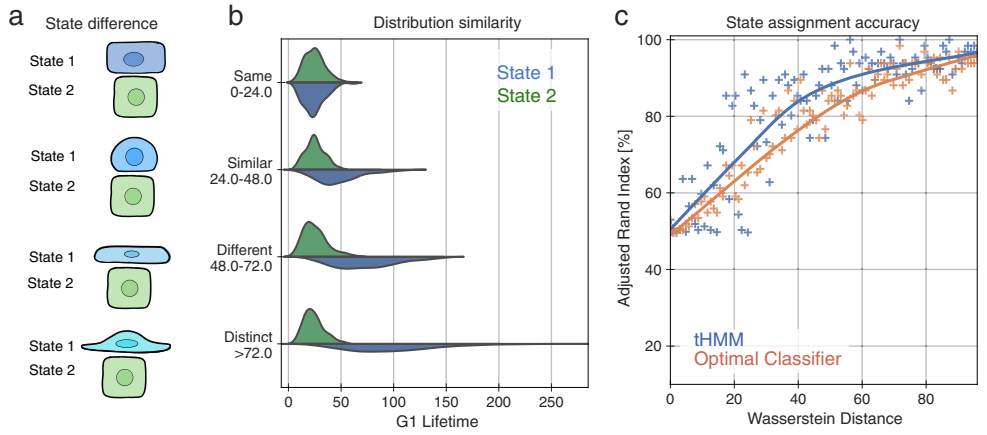


Fig. 5 Model performance versus the difference between states. a Cartoon of how two states can vary in their phenotypic similarity, in a synthetic population of two states. On the top, cells might be virtually indistinguishable (here based on shape). On the bottom, they might be so different that looking at one cell is sufficient to identify its state. **b** The distribution of G1 duration is varied in state 1 (blue) while the other state is kept constant. **c** State assignment accuracy versus the Wasserstein distance between state phenotypes. Each point represents the accuracy of state assignment for a lineage created by a set of parameters that yield the shown Wasserstein distance between the two-state distributions. 100 simulation trials are plotted. Either the tHMM model (blue) or an optimal classifier without lineage information (orange) was used. The solid lines show a Lowess trendline of the model accuracy.

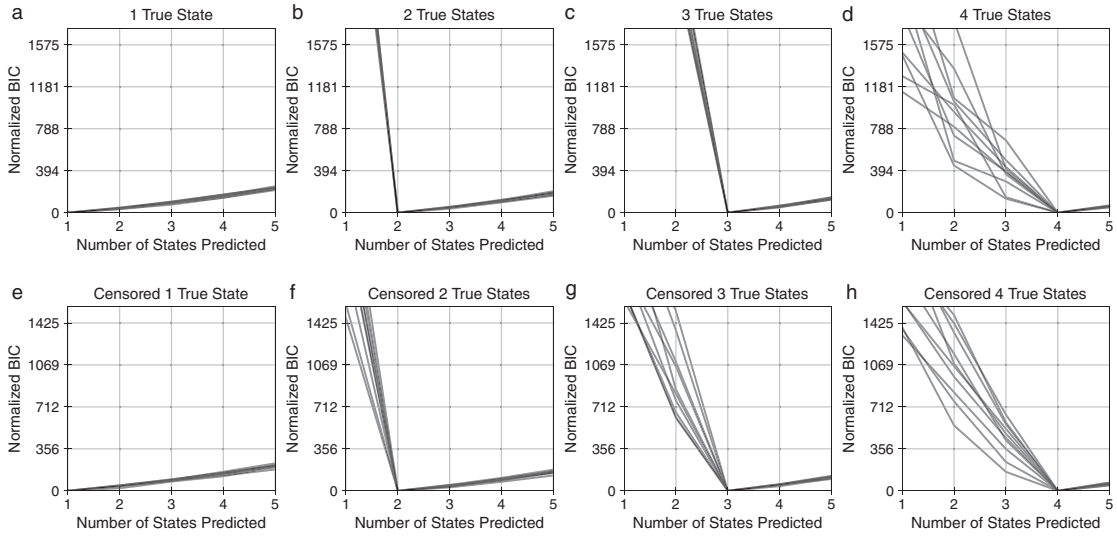


Fig. 6 Model selection effectively identifies the number of distinct states in synthetic data. a–d Model BIC for synthetic uncensored lineages with 1–4 true states. **e–h** Model BIC for synthetic censored lineages with 1–4 true states. BIC values are normalized such that the optimum is equal to 0. The minimum BIC value corresponds to the predicted number of states in each repetition. 10 trials plotted.

inferred number of cell states was consistently correct for both uncensored (Fig. 6a–d) and censored lineages (Fig. 6e–h). This indicated that model selection can help to identify the appropriate number of cell states for a set of measurements.

tHMM infers several distinct subpopulations in experimental drug response data. As an application of our model, we used phenotypic measurements from two cell lines. With the first, AU565, we measured of the G1 and S/G2 phase durations and terminal cell fates of cells in a control condition and when treated with 3 concentrations of gemcitabine or lapatinib. For the second, MCF10A, we measured the overall cell lifetimes and terminal fates of cells treated with PBS or single concentrations of the growth factors EGF, HGF, or OSM. Cells were imaged every 30 minutes and then tracked over time to assemble lineage relationships. The lapatinib and gemcitabine-treated AU565 populations (including control) contained a total of 5290 and 4537 cells, respectively. The MCF10A population contained 1306 cells. Lineages included 1–5 generations of cells. The model was fit to each experiment's data across all conditions, enforcing that the initial and transition probabilities are shared across concentrations but allowing the phenotype distributions to vary. We enforced a unidirectional phenotypic shift with drug concentration in AU565 cells, reflecting the expectation of a dose-response effect on cell phenotype within each state. The cell fate

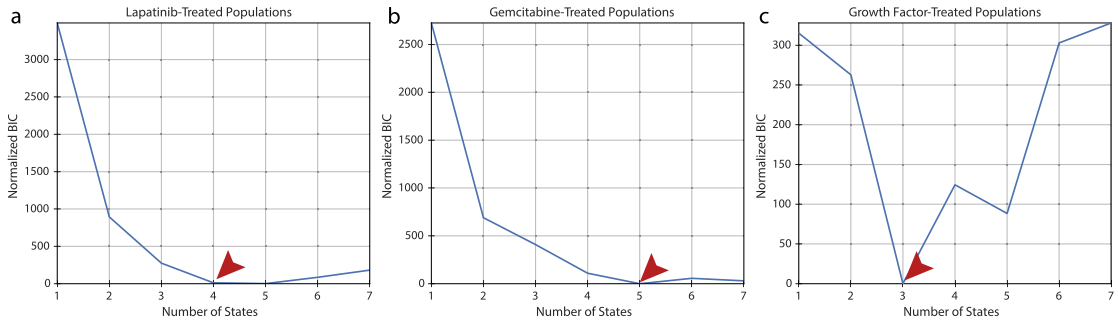


Fig. 7 BIC-based model selection infers the number of phenotypically distinct states. Normalized BIC values for (**a**) AU565 cells in control and treated with 5 nM, 25 nM, and 250 nM of lapatinib; (**b**) AU565 cells in control and treated with 5 nM, 10 nM, and 30 nM of gemcitabine; and (**c**) MCF10A cells treated with PBS, 10 ng/ml EGF, 40 ng/ml HGF, and 10 ng/ml OSM. The BIC values for all conditions were normalized such that the minimum value was zero. The arrows in (**a**–**c**) point to the optimal number of states.

parameters were estimated without constraints. We assumed the number of states is shared across drug concentrations in AU565 cells and across growth factor treatments in MCF10A cells. To determine the number of cell states, we compared models of 1–7 states using the BIC, where the lowest BIC value across numbers of states indicates the most optimal model correcting for complexity (Fig. 7a–c). The data for each compound indicated the presence of multiple inherited states.

To verify the model's predictive ability, we additionally implemented a cross-validation scheme for the lineage data. Briefly, roughly 20% of the cells were chosen at random and then masked from the fitting process. The model parameters were estimated using only the unmasked cells, though all cells received state assignments through use of their relatives. At the end, the log-likelihood of the masked cells' observations were evaluated using the fit model. We tested this cross-validation approach by creating synthetic cell populations of 2–5 true states with conditions matching the experimental data. For each scenario, we were able to identify the correct number of states based on which gave the highest log-likelihood (Supplementary Fig. 16a–c, f, g, Supplementary Table 3). Cross-validating the experimental data again confirmed the 4 and 5 phenotypic states within the lapatinib and gemcitabine data, respectively (Supplementary Fig. 16d, e). It also directly demonstrated that the inclusion of multiple states enables the tHMM model to predict unseen data, and that this prediction is dependent on inheritance; a no-inheritance model, in which all transitions were equally likely, performed relatively poorly (Supplementary Fig. 16d, e).

Lapatinib response is defined by both stable and interconverting states. We fit the lapatinib-treated data to the model with 4 states based on our BIC-based model selection, confirmed by cross-validation (Fig. 7a, Supplementary Fig. 16f). Fitting revealed states of widely varying persistence over generations, from less than a 0.01 probability of remaining in state 2 to a 0.94 probability of remaining in states 1 and 3 (Fig. 8a). Interestingly, states 2 and 4 formed a cycle wherein the most probable transition was between the two (Fig. 8a, Supplementary Fig. 11).

Examining the phenotypes of each state revealed distinct drug responses. Lapatinib is an EGFR/HER2 inhibitor that induces cell cycle arrest in G1 phase[44]. Every state displayed a dose-dependent increase in G1 phase lifetime with lapatinib treatment, and G1 effects were more pronounced as compared to those involving S/G2 (Fig. 8b–i, Supplementary Fig. 11). While the probability of survival at the end of the cell cycle phase decreased at higher concentrations, very few cell death events were observed (Fig. 8h, i, Supplementary Fig. 11). Consequently, the chances of cell death likely have high uncertainty at higher concentrations of lapatinib. States 2 and 4 were highly arrested in both G1 and S/G2 phase; in contrast, states 1 and 3 experienced little arrest in G1 and no arrest in S/G2 phase (Fig. 8f, g). Thus, cell states seemed to be primarily distinguished based on the degree of lapatinib response. The cycle between states 2 and 4 seems to reflect the observation that cells more highly arrested in G1 than G2/S give rise to cells that spend longer in G2/S than G1, and vice versa (Fig. 8, Supplementary Fig. 11).

Gemcitabine-treated populations are clustered into phase-specific responses. Gemcitabine is a chemotherapy agent that induces cell cycle arrest and apoptosis in S phase by disrupting DNA repair. The AU565 cells were treated with 5, 10, and 30 nM of gemcitabine; model selection, confirmed by cross-validation, inferred 5 states in the population (Fig. 7b, Supplementary Fig. 16a/g). Examining the 5-state fit revealed relatively stable states 1, 3, and 4 (Fig. 9a–e, Supplementary Fig. 12). States 2 and 5 formed a cycle with high rates of interconversion.

Gemcitabine modulated both G1 and S/G2 cell cycle phases and these effects were variable across the five identified states (Fig. 9f–i). State 5 showed S/G2-specific arrest and always resulted in cell death at the highest concentration (Fig. 9g, i, Supplementary Fig. 12). Meanwhile, cells in state 4 grew almost normally, with some cell death in G1 at the highest concentration (Fig. 9f–i). At the highest concentration, state 3 represents the cells arrested at S/G2 that have not divided even once, and state 5 is the representative of almost all cells undergoing cell death at S/G2 (Fig. 8i, Supplementary Fig. 12).

Lastly, we wished to explore whether the phenotypic heterogeneity we observed was limited to cancer cells or cytotoxic drug treatment. To determine this, we tracked non-tumorigenic MCF10A breast cells. These cells are normally grown in the presence of epidermal growth factor (EGF); we compared this condition to growth factor withdrawal (PBS) or rescue with hepatocyte growth factor (HGF) or oncostatin M (OSM)[45]. Each growth factor consistently promoted proliferation on a population level compared to the PBS control, though with considerable inter- and intra-lineage variation (Supplementary Fig. 13). BIC-based model selection inferred the presence of 3 distinct states (Fig. 7c). Inspecting the model revealed generally more dynamic transitions between states as compared to the AU565 experiments (Supplementary Fig. 14). Due to the lack of growth factors, most cells arrested in the PBS condition; few observations of either division or death events is the reason for the division probability being 0.5 (Supplementary Fig. 14g). State 1 was distinct in being relatively less responsive to HGF and OSM treatments (Supplementary Fig. 14a/f), while state 1 displayed

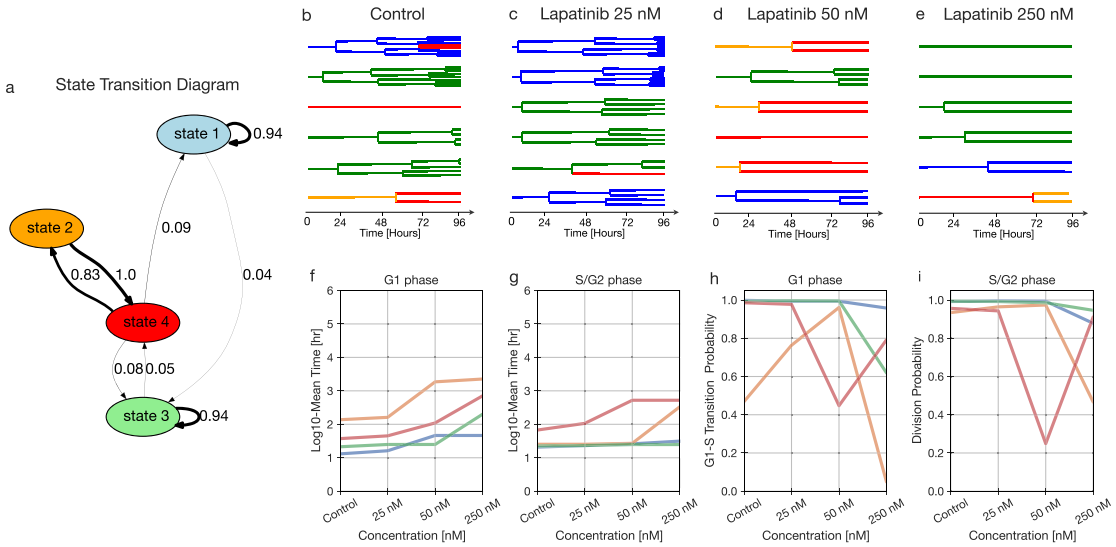


Fig. 8 Lapatinib response is defined by phenotypically-distinct stable and interconverting states. a State transition graph showing the probability of state transitions among the predicted states. Transitions with less than a 0.03 probability have been removed. **b**–**e** A sample of lineage trees after fitting the model and state assignment (control, 25 nM, 50 nM, and 250 nM). **f**–**g** The $\log_{10}$ of fit mean time of G1 and S/G2 phase durations for different concentrations. **h**, **i** The Bernoulli parameter, indicating the probability of G1-to-S phase transition versus cell death (**h**), and the probability of division versus cell death (**i**) for each concentration.

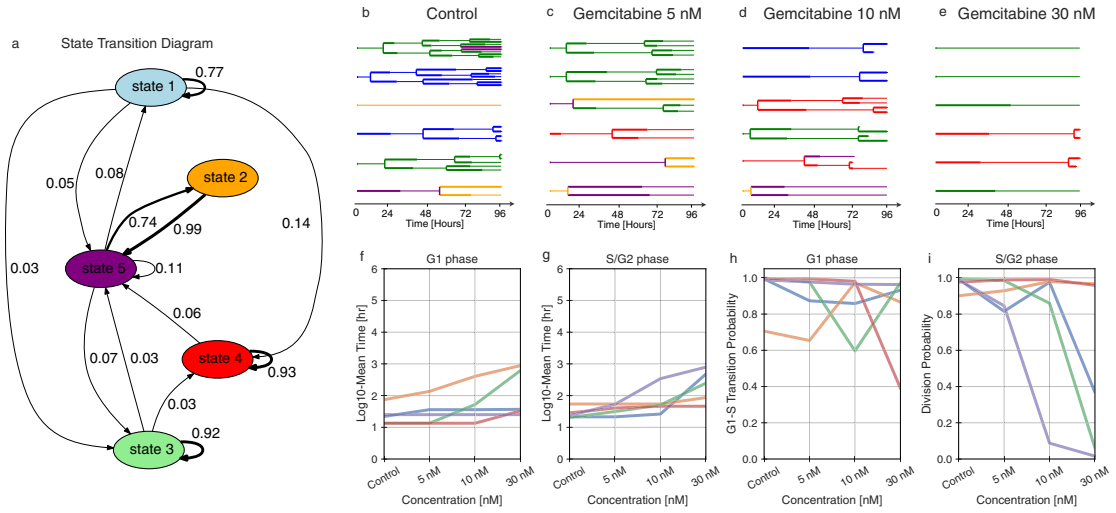


Fig. 9 State-specific inferences of the gemcitabine-treated data. a State transition graph showing the probability of state transitions among the predicted states. The transitions with less than a 0.03 probability have been removed. **b**–**e** A sample of lineage trees after fitting the model and state assignment (control, 5 nM, 10 nM, and 30 nM). **f**, **g** The $\log_{10}$ fit mean time of G1 and S/G2 phase durations for different concentrations. **h**, **i** The Bernoulli parameter, indicating the probability of G1-to-S phase transition versus cell death (**h**), and the probability of division versus cell death (**i**) for each concentration.

higher rates of cell death overall (Supplementary Fig. 14a/g). In total, the tHMM model was effective in identifying subsets of cells with divergent phenotypic responses to drug treatment alongside the relationships between cells in the population.

Discussion

Heterogeneity and plasticity in cancer cells enables them to adapt in response to therapy. Even in the absence of genetic mutations, other heritable variation serves as a substrate for selection[46,47]. In this paper, we introduced a tree-based hidden Markov model that clusters single cells from heterogeneous populations based solely on their phenotypic traits and relationships. Model benchmarking showed that it can provide accurate results using feasible experimental designs. Of particular importance, the tHMM could recognize subpopulations even at lower frequencies (Supplementary Figs. 6, 7). Comparing the model to more standard clustering, the tHMM showed that lineage information helps to

identify cell states more accurately (Fig. 5). Using cross-validation, we were able to show that accounting for cell inheritance allowed the model to accurately predict unseen observations (Supplementary Fig. 16). Several critical advancements in the current work are a modular interface for using tHMM models with various phenotypes (Fig. 2), proper censorship handling (Fig. 3), strategies for model evaluation (Fig. 7, Supplementary Fig. 16), and demonstrating that such a model can be applied to study cancer heterogeneity at baseline and in response to perturbation.

We used single-cell lineage tracking data of AU565 cancer cells treated with lapatinib and gemcitabine as a demonstration of the model. G1 and S/G2 cell cycle phase durations and cell fate measurements were used as relevant cell phenotypes to quantify the anti-cancer effects of these drugs. We were able to identify 4 and 5 distinct subpopulations within the lapatinib and gemcitabine-treated data, respectively (Fig. 7). The phenotypic features of each state were quantified in parallel (Figs. 8, 9). Lapatinib is known to inhibit cell proliferation by inhibiting Akt/mTOR pathway activity, which is a key regulator of G1 phase progression[48]. Similarly, our analysis in the lapatinib-treated population indicated that cells, regardless of their state, experienced a prolonged G1 phase, but individual states varied in their susceptibility. In gemcitabine-treated cells, we observed that most states were highly heritable, with more varied phenotypic effects. This included cells that became arrested in S/G2 and underwent apoptosis (state 5), cells that were selectively arrested in G1 (state 1), and cells that hardly responded to drug treatment at all (state 4; Fig. 9). While gemcitabine canonically works by inducing cell arrest in G2/S, previous work has characterized its effects on G1 phase by separating the effects on both cell cycle phases[49]. They similarly identified that G1 arrest was associated with cell death, which is also evident in cells of state 3 where G1 arrest is seen alongside cell death in both phases (Fig. 9f–i). Our results would further suggest that those cells with G1 effects are molecularly and heritably distinct from those that are arrested in S/G2. MCF10A cells with growth factor-induced proliferation showed a very distinct pattern of variation, suggesting that the phenotypic cell states identified by the model reflect a confluence of cell features and treatment conditions (Supplementary Figs. 13, 14).

We present several lines of evidence supporting the accuracy of the model and the existence of heritable cell states. First, across a diverse array of benchmarking experiments, we show that the model can derive accurate conclusions from synthetic data with properties like those we observed in the experimental measurements (e.g., Fig. 4). Through an informatic model selection scheme, we find statistical evidence for the existence of multiple states (Figs. 6, 7). Examining these cell states, we find patterns consistent with the biological mechanisms of the compounds we used to alter cell proliferation (Figs. 8, 9). Reassuringly, we were able to confirm that the abundance of cell states was consistent across experimental replicates, ruling out the possibility that state differences arose from day-to-day variation between experiments. Finally, we showed that the model could more effectively predict the behavior of unseen cells with the inclusion of multiple cell states, and that this prediction is dependent on allowing inheritance between cell generations (Supplementary Fig. 16). While we have considered the use of experimental control conditions, it is important to keep in mind that the variation observed here arises both through external perturbation and natural variation within the population. Consequently, we have not been able to identify a context in which one might expect to not observe multiple states, supporting the general usefulness of our approach. While experiments in which distinct cell lines are mixed can help to validate methods in which cell relationships are inferred, such as pseudotime methods[50], the cell relationships are not modeled here because they are explicitly known through the measured lineage relationships. Ultimately, experiments uncovering molecular markers and mechanisms of these cell states will provide the best independent validation for their biological significance.

Modeling advancements will further improve on our approach. Cells may express a continuum of, rather than discrete, phenotypic states[22]. If this is the case, a continuous latent variable model would lead to a refined view of the population-level heterogeneity. A discrete model like the one used here should, however, still provide an accurate estimate by breaking up the continuous state-space into discrete steps. Continuous latent variable models also have additional challenges in implementation and interpretation[51,52]. Careful handling of each states' phenotypic distributions might also improve the model's accuracy and power to identify distinct states. For example, the eventual fate of cells and their cell cycle durations are likely correlated which could be handled through a multivariate distribution accounting for this covariance[53]. This becomes even more important with the incorporation of other phenotypic information such as migration, cell shape, or other features, all of which are likely to be correlated to some extent.

Experimental advancements will improve the utility and accuracy of single-cell analysis using lineage information. Currently our experimental data is limited to 96 hours, covering up to five generations of cells. However, traits such as resistance may develop over more generations and longer timescales[13,54,55]. Longer data collection becomes challenging due to factors such as phototoxicity and cell stress[56]. Improved imaging modalities and experimental platforms might allow for longer tracking experiments, with reduced phototoxicity, in more physiologically representative environments such as engineered 3D extracellular matrix[57,58]. Currently, the model is agnostic as to whether the heterogeneity it identifies is pre-existing or induced by drug treatment. Collecting data in which cells are tracked before and after drug treatment, and after a wash-out, would help to link pre- and post-exposure cell phenotypes[59].

While we have identified states that represent phenotypically distinct subpopulations of cells, we currently cannot comment on the molecular factors leading to these phenotypes. Molecular barcoding has been a popular approach for identifying subpopulations of cells with genetic predispositions toward unique phenotypes, but we do not expect it would identify the same subpopulations as we do here[55]. Unlike in barcoding experiments, we do not see a bottleneck in the clonality of cells that survive treatment, and rapid interconversion between states should corrupt the relationship between ancestor phenotype and descendent molecular state[3,12]. However, we expect that single-cell molecular analyses, such as single-cell tracking tied with transcriptional profiling of the same cells at the end of the experiment, should allow us to align molecular and phenotypic states in the same populations of cells[60]. Such experiments would also provide a common baseline by which to link lineage-based phenotypic analysis and various snapshot measurements of the same cell population. In this way it should be able to pinpoint the underlying molecular mechanisms driving distinct phenotypic responses.

In total, the pipeline developed here provides a unique approach for understanding the structure of dynamic, heterogeneous tumor populations. By capturing the dynamics of state transitions, it links single-cell phenotypes to overall population behavior. Incorporating molecular measurements, and a broader set of drug interventions, will then also help to identify means of modulating state and overall population behavior. Ultimately, we expect this integrative view will help to identify treatments alone and in combination that allow for population-level control by affecting the growth of and transitions between individual cell subpopulations.

Methods

Experimental cell lineage data. Stable cell line creation, drug treatments, and tracking of AU565 and MCF10A cells were performed as described in Gross et al.[61] and Gross et al.[45], respectively. Briefly, AU565 cells were co-transfected with a transposase plasmid (Addgene #34879) and a donor plasmid that drove expression of a nuclear-localized mCherry, puromycin resistance, and a fragment of HDHB fused to the clover fluorescent protein, which was used to track progression through the cell cycle[62]. Cells stably expressing the nuclear and cell cycle reporter were selected for 7 days with 0.75 μg/ml puromycin. The phase of the cells is determined based on whether the amount of fluorescence is greater within nucleus or the cytoplasm[62]. As a result, the reporter signal is invariant to changes in exposure and background. To track drug responses AU565 reporter cells were plated into 24-well plates with fluorobrite media containing 10% FBS, glutamine, and penicillin-streptomycin. 24 hours later fresh media containing escalating doses of lapatinib and gemcitabine was added. MCF10A cells were cultured in growth media (DMEM/F12, 5% horse serum, 20 ng/ml cholera toxin, 10 μg/ml insulin, and 1% Pen/Strep), grown to 50–80% confluency, and detached with 0.05% trypsin-EDTA. 7 hours after seeding 75000 cells, they were washed with PBS and the experiment media (DMEM/F12, 5% horse serum, 0.5 μg/ml hydrocortisone, 100 ng/ml cholera toxin, and 1% Pen/Strep) was added to the 8 well-plates which was followed by 18 hours of incubation. Afterward, cells were treated with growth factors 10 ng/ml EGF, 40 ng/ml HGF, and 10 ng/ml OSM in fresh experiment media. After drug addition, plates were placed in the IncuCyte S3 and four image locations per treatment were imaged every 30 minutes. AU565 were imaged for 96 hours and MCF10A cells for 48 hours. After half the experiment times, fresh media and drugs/growth factors were added. Cell lineages from the IncuCyte images were manually tracked in FIJI[63] to record cell division, death, and the transition from G1 to S/G2 phase (in AU565). AU565 cells are non-motile and fewer than 4% of cells were within one cell length of the image boundary, ensuring minimal sampling bias from the microscopy field of view. Three biological replicates were collected and combined in the final data set. To verify that results did not reflect batch effects, we checked that state assignments were not enriched or depleted within a replicate.

Lineage tree-based hidden Markov model. The core assumption of a Markov chain is that the next state and current observations are only dependent on the current state. Proof of the expressions below involving cell state assignment (expectation step), including the upward recursion, downward recursion, and Viterbi algorithms, can be found in Durand[33]. All other model elements, including the emissions distribution fitting, model evaluation strategies, and censorship corrections were developed in this study.

Basic model structure. The initial probabilities of a cell being in state k are represented by the vector π that sums to 1:

$$\pi_k = P(z_1 = k), \quad k \in \{1, \ldots, K\} \tag{1}$$

where z indicates the state and K is the total number of states. The probability of state i transitioning to state j is represented by the $K \times K$ matrix, T, in which each row sums to 1:

$$T_{i,j} = T(z_i \to z_j) = P(z_j | z_i), \quad i,j \in \{1, \ldots, K\} \tag{2}$$

The emission likelihood matrix, EL, is based on the cell observations. It is defined as the probability of an observation conditioned on the cell being in a specific state:

$$EL(n,k) = P(x_n = x | z_n = k) \tag{3}$$

where x_n is the observation for cell number n, with a total of N cells in a lineage. Separate observations were assumed to be independent; for instance, cell fate is assumed to be independent from the duration of each cell phase. This facilitates calculating the likelihood of observations, because we can multiply the likelihood of all observations together for the overall likelihood.

Assigning cell states (expectation step)

Upward recursion: An upward-downward algorithm for calculating the probabilities in hidden Markov chains was proposed by Erphaim and Merhav[64] which suffered from underflow. This problem was originally solved by Levinson[65], where they adopted a heuristic-based scaling, and then was improved by Devijver[66] where they introduced smooth probabilities. Durand[33], however, revised this approach for hidden Markov trees to avoid underflow when calculating $P(Z|X)$ probability matrices. To explain we need the following definitions:

$p(n)$ is the parent cell of cell n, and $c(n)$ is the children of cell n.

$\bar{X}$ is the observation of the whole tree and $\bar{X}_a$ is a subtree of $\bar{X}$ which is rooted at cell a.

$\bar{Z}$ is the complete hidden state tree.

$\bar{X}_{a/b}$ is the subtree rooted at a except for the subtree rooted at cell b, if $\bar{X}_b$ is a subtree of $\bar{X}_a$.

For the state prediction we start by calculating the marginal state distribution (MSD) matrix. MSD is an $N \times K$ matrix that for each cell is marginalizing the transition probability over all possible current states by traversing from root to leaf cells:

$$MSD(n,k) = P(z_n = k) = \sum_i P(z_n = k | z_{n-1} = i) \times P(z_{n-1} = i) \tag{4}$$

During upward recursion, the flow of upward probabilities is calculated from leaf cells to the root cells generation by generation. First, for leaf cells, the probabilities (β) are calculated by:

$$\beta_n(k) = P(z_n = k | X_n = x_n) = \frac{EL(n,k) \times MSD(n,k)}{NF_l(n)} \tag{5}$$

in which X_n is the leaf cell's observation, and NF (Normalizing Factor) is an $N \times 1$ matrix that is the marginal observation distribution. Since $\sum_k \beta_n(k) = 1$, we find the NF for leaf cells using:

$$NF_l(n) = \sum_k EL(n,k) \times MSD(n,k) = P(X_n = x_n) \tag{6}$$

For non-leaf cells the values are given by:

$$\beta_n(k) = P(z_n = k | \bar{X}_n = \bar{x}_n) = \frac{EL(n,k) \times MSD(n,k) \times \prod_{v \in c(n)} \beta_{n,v}(k)}{NF_{nl}(n)} \tag{7}$$

where we calculate the non-leaf NF using:

$$NF_{nl}(n) = \sum_k \left[EL(n,k) \times MSD(n,k) \prod_{v \in c(n)} \beta_{n,v}(k) \right] \tag{8}$$

and linking β between parent-daughter cells is given by:

$$\beta_{p(n),n}(k) = P(\bar{X}_n = \bar{x}_n | z_{p(n)} = k) = \sum_j \frac{\beta_n(j) \times T_{k,j}}{MSD(n,j)} \tag{9}$$

By recursing from leaf to root cells, the β and NF matrices are calculated as upward recursion. The NF matrix gives a convenient expression for the observation log-likelihoods. For each root cell we have:

$$P(\bar{X} = \bar{x}) = \prod_n \frac{P(\bar{X}_n = \bar{x}_n)}{\prod_{v \in c(n)} P(\bar{X}_v = \bar{x}_v)} = \sum_n NF(n) \quad n \in \{1, \ldots, N\} \tag{10}$$

The overall model log-likelihood is given by the sum over root cells:

$$\log P(\bar{X} = \bar{x}) = \sum_n \log NF(n) \tag{11}$$

Downward recursion: For computing downward recursion, we need the following definition for each root cells:

$$\gamma_1(k) = P(z_1 = k | \bar{X}_1 = \bar{x}_1) = \beta_1(k) \tag{12}$$

The other cells follow in an $N \times K$ matrix by writing the conditional probabilities as the summation over the joint probabilities of parent-daughter cell:

$$\gamma_n(k) = P(z_n = k | \bar{X}_1 = \bar{x}_1) = \frac{\beta_n(k)}{MSD(n,k)} \sum_i \frac{T_{i,k} \gamma_{p(n)}(i)}{\beta_{p(n),n}(i)} \tag{13}$$

Viterbi algorithm: Given a sequence of observations in a hidden Markov chain, the Viterbi algorithm is commonly used to find the most likely sequence of states. Equivalently, here it returns the most likely sequence of states of the cells in a lineage tree using upward and downward recursion[33].

The algorithm follows an upward recursion from leaf to root cells. We define δ, an $N \times K$ matrix:

$$\delta_n(k) = \max_{\bar{z}_{c(n)}} \left\{ P(\bar{X}_n = \bar{x}_n, \bar{Z}_{c(n)} = \bar{z}_{c(n)} | z_n = k) \right\} \tag{14}$$

and the links between parent-daughter cells as:

$$\delta_{p(n),n}(k) = \max_{\bar{z}_n} \left\{ P(\bar{X}_n = \bar{x}_n, \bar{Z}_n = \bar{z}_n | z_{p(n)} = k) \right\} = \max_k \left\{ \delta_n(k) T_{k,k} \right\} \tag{15}$$

We initialize from the leaf cells as:

$$\delta_n(k) = P(X_n = x_n | z_n = k) = EL(n,k) \tag{16}$$

and for non-leaf cells use:

$$\delta_n(k) = \left[\prod_{v \in c(n)} \delta_{n,v}(k) \right] \times EL(n,k) \tag{17}$$

The probability of the optimal state tree corresponding to the observations tree, assuming root cell is noted as cell 1, is then given by:

$$Z^* = \max_k \{ \delta_1(k) \pi_k \} \tag{18}$$

which arises from maximization over the conditional emission likelihood (EL) probabilities by factoring out the root cells as the outer maximizing step over all possible states.

Fitting the cell phenotypes (maximization step). In the maximization step, we find the maximum likelihood of the hidden Markov model distribution parameters. We estimate the initial probabilities, the transition probability matrix, and the parameters of the observation distributions. The maximum likelihood estimation of the initial probabilities can be found from each state's representation among the root cells:

$$\pi_k^* = \gamma_1(k) \tag{19}$$

Similarly, the transition probability matrix is estimated by calculating the prevalence of each transition across the lineage trees:

$$T_{i,j}^* = \frac{\sum_{n=1}^{N-1} \xi_n(i,j)}{\sum_{n=1}^{N-1} \gamma_n(i)} \tag{20}$$

where

$$\xi_n(i,j) = \left(\frac{\gamma_{p(n)}(i)}{\frac{\beta_n(i)}{MSD(n,i)} T(i,j)} \right)^T \times \frac{\beta_n(j)}{MSD(n,j)} \tag{21}$$

Estimating emissions distribution parameters. In the current study, we used two emissions distributions; first, a Bernoulli distribution for the probability of each cell fate, either at the end of each cell cycle phase or at the end of cell's lifetime; second, a gamma distribution for the durations of each cell cycle phase or overall cell lifetime. To estimate the distribution parameters after finding the cell state assignments, we calculated their maximum likelihood estimation weighted by their proportional assignment to that state. The initial and transition probabilities were shared across drug concentrations.

For estimating the Bernoulli distribution parameter for cell fate, we simply found the state assignment-weighted sample mean of the observations. To estimate the gamma distribution parameters, we fit all concentrations of each drug simultaneously and assumed that increasing drug concentration had a unidirectional effect on the observed phenotype within each state. This was implemented, using sequential least-squares programming (SLSQP)[67], through a linear constraint on the scaling parameter of the gamma distributions between concentrations so that higher concentrations had equal or greater average durations. The gamma distribution likelihood fitting is a convex optimization problem, indicating that local optimization can arrive at the globally optimal solution. Linear constraints do not change this property, and we confirmed fitting with different starting points arrived at the same solution. We used censored estimators to handle the effect of time censorship (explained below) in the duration distribution fitting. This was done by fitting uncensored and censored observations to the complete and survival distributions, respectively, and using the accumulated log-likelihood to estimate the distribution parameters.

Baum-Welch. Since both the hidden states and model parameters are unknown, we applied expectation-maximization (EM), known as the Baum-Welch algorithm in the case of HMMs, to find both the model parameters and cell states.

The expectation-maximization algorithm consists of two steps: expectation and maximization. During expectation, the probabilities of all cells being in specific states are calculated, such that for every cell and every state we have $P(z_n = k|X_n)$ and $P(z_n = k, z_{n+1} = l|X_n)$. The expectation step is calculated by the upward and downward recursion algorithms described above. In the maximization step, described above, the distribution parameters of each state, the initial (π) probabilities, and the transition probability (T) matrices are estimated, given the state assignments of each cell.

The expectation-maximization algorithm is initialized by randomly assigning the cells to states using a Dirichlet distribution. During fitting we iteratively switch between the expectation and maximization steps and then calculate the likelihood. If the likelihood improves less than a set threshold, we take that to indicate convergence.

Model evaluation. To find the most likely number of states corresponding to the observations, the Bayesian Information Criterion (BIC) was used[68]. The BIC requires the number of degrees of freedom, which we calculate using the number of independent parameters. Our model estimates a k element initial probability vector, a $k \times k$ transition matrix, and a $k \times m$ matrix of state-wise parameters where k is the number of states and m is the number of parameters associated with observation distributions. For the phase-specific observation distributions we have a total of 6 parameters, including 2 Bernoulli parameters and 2 pairs of shape and scale parameters for the gamma distribution. Since the row-sums for transition and initial probability matrices must be 1, these values are not entirely independent. From distribution analysis of the phase lengths, we realized the shape parameter of the gamma distribution remains constant over different conditions, while the scale parameter changes. Therefore, the shape parameter was shared between the populations treated with 4 different concentrations of the same compound. Each condition, therefore, introduced 2 free parameters (1 Bernoulli parameter and 1 scale parameter). For the MCF10A experiments, terminal fates and cell cycle durations were also assumed to be Bernoulli- and gamma-distributed, respectively. The shape of cell lifetime was similarly shared among the four conditions (PBS, EGF, HGF, and OSM).

The Wasserstein or Kantorovich–Rubinstein metric is a measure of distance between two distributions. This metric was used to determine the difference between state emissions[69]. An analytical solution, the absolute value of the difference in distribution means, was used for the gamma distribution.

Model benchmarking. We used emission distributions to represent the phenotypic characteristics of the cells within the lineages. To create our synthetic data, we considered two possible options as our set of observations throughout an experiment. In one case, we modeled the overall cell fate and cell lifetime; in the second, we modeled the phase-specific fate and duration. In both, we used a Bernoulli distribution for the fate outcomes and a gamma distribution for durations. The state assignment accuracy was calculated using the Rand Index[40]. The difference between true and estimated probability matrices was assessed using the Frobenius norm, or the sum of each element squared.

Synthetic lineage data generation. We generated synthetic lineage trees with K discrete states and N total number of cells for benchmarking. Lineages were composed of two primary data structures: the state and emissions trees. The state tree was randomly seeded with a root cell determined by the starting probabilities, then expanded by randomly sampling transitions based on the transition probability matrix. The lineages were extended by either increasing the number of initial cells, resulting in a greater number of lineages (breadth), or by lengthening the experiment time resulting in each lineage containing more cells (depth). After creating the tree of states with the desired number of cells, the emission tree is built upon it. Emissions were randomly sampled from the distributions for each cell's state. Finally, the effects of the emissions were applied to the tree when necessary. If any cells died, their progeny were marked as unobserved by making their emissions equal to NaN (Not a Number). If applicable, the effects of finite-duration experiments were also applied. Cells existing outside of the experiment duration were marked as unobserved, and those crossing the bounds of an experiment were marked as censored with duration clipped by the experiment.

Time censorship. Our phenotypic measurements include the cell fate (progression or cell death) and duration. These measurements are made for each cell cycle phase (G1 or S/G2) in the case of AU565 cells and for the entire lifetime for MCF10A cells. These measurements can contain incomplete information due to the bounds of an experiment. For instance, it is unknown when initial cells present at the start of the experiment began their cell cycle. The same is true of the cells present at the end of the experiment because we do not observe their end. Hence, a cell's lifetime and/or fate may be partially observed. To ensure our synthetic data is a close reflection of experimental data, we incorporated this effect in our synthetic data. Cells with lifetimes that extend beyond the end of the experiment were marked as censored for the lifetime estimation.

Cell overall lifetime observations. The parameters are reflective of the cell phenotypes we observed with 5 nM lapatinib treatment. Supplementary Figs. 1–5 are based on these parameters. Each figure includes 100 trials.

Transition probability matrix: $T = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$

The initial probability vector is then calculated as the stationary distribution of states from transition probability matrix, satisfying $\pi = \pi * T$.

In this case, we have: $\pi = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

The same T and π were used for phase-specific emissions.

In Supplementary Table 1, "Bern p" refers to the Bernoulli parameter, the cell division probability at the end of its lifetime which is equal to 1—the probability of cell death at the end of its lifetime. "Shape" and "Scale" refer to the gamma distribution parameters. The cells' lifetimes were fit to gamma distributions.

Cell cycle phase-specific observations. The synthetic data used in Figs. 3, 4, Supplementary Figs. 8–10 were created based on the following parameters. These parameters are based on estimations from AU565 cells treated with 5 nM lapatinib. Each figure includes 100 trials.

In Supplementary Table 2, "G1 bern" and "S/G2 bern" are the cell division probabilities at the end of G1 and S/G2 phase, respectively. The "G1 shape" and "G1 scale" are the gamma distribution parameters of G1 phase lengths. "S/G2 shape" and "S/G2 scale" are the gamma distribution parameters of S/G2 phase lengths.

To benchmark the model with 5 states, we simulated 25–500 lineages, each with up to 31 cells, to create a population with 5 states. Like with the experimental data, we assumed the experiment ends after 96 hours and censored the cells' observations accordingly. The model parameters, including the transition probabilities and initial probabilities are listed below. The analysis results are

shown in Supplementary Fig. 14.

$$T = \begin{bmatrix} 0.6 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.05 & 0.8 & 0.05 & 0.05 & 0.05 \\ 0.01 & 0.1 & 0.7 & 0.09 & 0.1 \\ 0.1 & 0.1 & 0.05 & 0.7 & 0.05 \\ 0.1 & 0.1 & 0.05 & 0.05 & 0.7 \end{bmatrix} \quad (22)$$

$$\pi = \begin{bmatrix} 0.13 \\ 0.33 \\ 0.16 \\ 0.18 \\ 0.18 \end{bmatrix} \quad (23)$$

Figure 6 uses the first 4 states of Supplementary Table 3 as the parameter set for the emissions matrix to simulate varying state numbers in the BIC calculation.

Varying emission differences. To create synthetic data with subpopulations of varying dissimilarity (Fig. 5), we use the phase-specific parameters, with the values for the G1 phase gamma scale parameter for state 1 varying over [4, 20]. This results in an increase in the Wasserstein distance between the two cell states, allowing us to measure state assignment accuracy for different dissimilarity amounts between the two states. Likewise, for Supplementary Figs. 6, 7, we simulated the overall cell lifetime and varied the gamma distribution scale parameter from 1 to 8 for state 1.

Optimal baseline classifier. To compare the tHMM with a classifier that ignores heritability, we manually calculated the optimal classification boundary between the gamma distributions for state 1 and state 2. The best choice of classification boundary between two gamma distributions is the point at which the likelihood of the random variable, x, is equal between the two distributions:

$$p(x|G(k_1, \theta_1)) = p(x|G(k_2, \theta_2)) \quad (24)$$

where k_1, θ_1, k_2, and θ_2 are the shape and scale parameters of the gamma distribution corresponding to state 1 and 2, respectively. The shape parameter was shared between the two distributions. Consequently, this can be simplified to:

$$x = \frac{k \ln \frac{\theta_2}{\theta_1}}{\frac{1}{\theta_1} - \frac{1}{\theta_2}} \quad (25)$$

We assigned the classification labels to the observations using this classification boundary, which formed the baseline accuracy shown in Fig. 5c. As states 1 and 2 are identical at the very first point, we used the distribution mean ($k \times \theta$) as the threshold.

Cross-validation. To split the lineage data into train and test sets, we randomly selected 20% of cells from each condition and masked their observations such that they would not contribute to the fitting process. This was performed by setting the log-likelihood of the masked cells' observations to be uniformly zero for all the states. During the Baum-Welch fitting, the algorithm estimates the parameters using only the training cells. However, during the expectation step, the state of masked cells is still inferred via information about their relatives. After the fitting converges, we calculate the log-likelihood of the test cells' observations given their state assignments. This is accumulated into an overall likelihood of the held-out observations given the tHMM state assignments and fit.

To test this cross-validation scheme's ability to determine the optimum number of states for a cell population, we created synthetic populations with 2–5 true states. States 1–n were used, where n is the number of true states, to generate data that is like the experimental data. The state observation distributions shown in Supplementary Table 2. The transition probabilities were generated by adding 0.1 elementwise to the identity matrix and then normalizing it. The initial probabilities for all states were equal. Fitting was performed with models including 1–7 states. The optimum number of states was taken to be the smallest number of states at which the log-likelihood plateaus.

Lowess trendline. Locally Weighted Scatterplot Smoothing (Lowess) was used to provide the trendlines in the figures with repeated model runs.

Statistics and reproducibility. The experiments were repeated in three independent biological replicates and yielded similar results.

Data availability

The experimental lineage data for AU565 and MCF10A cell lines can be found at https://github.com/meyer-lab/tHMM and https://doi.org/10.5281/zenodo.7195355 The synthetic data from which we plotted Figs. 3c, d, 4b–g, 5c–7 uses the code in the file named after the corresponding figure number. Data used in Figs. 7a, b, 8, 9 uses the AU565 cell line experimental lineage data, and Fig. 7c and 10 use the MCF10A lineage data. The cell lines used in this study (AU565, MCF10A) can be made available upon request.

Code availability

All analysis were implemented in Python v3.9 and can be found at https://github.com/meyer-lab/tHMM. The repository can also be found at Zenodo[70].

Received: 9 August 2021; Accepted: 1 November 2022;
Published online: 17 November 2022

References

1. Di Maio, M. et al. Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. *Lancet Oncol.* **6**, 669–677 (2005).
2. De Roock, W. et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol* **11**, 753–762 (2010).
3. Sharma, S. V. et al. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* **141**, 69–80 (2010).
4. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).
5. Sigal, A. et al. Variability and memory of protein levels in human cells. *Nature* **444**, 643–646 (2006).
6. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17**, 284–299 (2016).
7. Falkenberg, K. J. & Johnstone, R. W. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat. Rev. Drug Discov.* **13**, 673–691 (2014).
8. Inde, Z. & Dixon, S. J. The impact of non-genetic heterogeneity on cancer cell death. *Crit. Rev. Biochem. Mol. Biol.* **53**, 99–114 (2018).
9. Pisco, A. O. & Huang, S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer* **112**, 1725–1732 (2015).
10. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
11. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nat.* **483**, 603–607 (2012).
12. Gupta, P. B. et al. Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* **146**, 633–644 (2011).
13. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
14. Gett, A. V., Sallusto, F., Lanzavecchia, A. & Geginat, J. T cell fitness determined by signal strength. *Nat. Immunol.* **4**, 355–360 (2003).
15. Arai, T. et al. Tumor Doubling Time and Prognosis in Lung Cancer Patients: Evaluation from Chest Films and Clinical Follow-up Study. *Jpn. J. Clin. Oncol.* **24**, 199–204 (1994).
16. Bourhis, J. et al. Potential doubling time and clinical outcome in head and neck squamous cell carcinoma treated with 70 GY in 7 weeks. *Int. J. Radiat. Oncol.* **35**, 471–476 (1996).
17. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
18. Huang, D. et al. High-Speed Live-Cell Interferometry: A New Method for Quantifying Tumor Drug Resistance and Heterogeneity. *Anal. Chem.* **90**, 3299–3306 (2018).
19. Tyson, D. R., Garbett, S. P., Frick, P. L. & Quaranta, V. Fractional proliferation: A method to deconvolve cell population dynamics from single-cell data. *Nat. Methods* https://doi.org/10.1038/nmeth.2138 (2012).
20. O'Connor, J. P. B. et al. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin. Cancer Res* **21**, 249–257 (2015).
21. Chen, K. et al. Phenotypically supervised single-cell sequencing parses within-cell-type heterogeneity. *iScience* **24**, 101991 (2020).
22. Kuchen, E. E., Becker, N. B., Claudino, N. & Höfer, T. Hidden long-range memories of growth and cycle speed correlate cell cycles in lineage trees. *Elife* **9**, e51002 (2020).
23. Mitchell, S., Roy, K., Zangle, T. A. & Hoffmann, A. Nongenetic origins of cell-to-cell variability in B lymphocyte proliferation. *Proc. Natl. Acad. Sci. USA.* **115**, E2888–E2897 (2018).

24. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
25. Brumpton, B. et al. Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. https://doi.org/10.1101/602516 (2019).
26. Concannon, P. et al. Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes* **58**, 1018–1022 (2009).
27. Concannon, P., Rich, S. S. & Nepom, G. T. Genetics of Type 1A Diabetes. *N. Engl. J. Med.* **360**, 1646–1654 (2009).
28. Quinn, J. J. et al. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).
29. Choo, K. H., Tong, J. C. & Zhang, L. Recent applications of Hidden Markov Models in computational biology. *Genomics. Proteomics Bioinformatics* **2**, 84–96 (2004).
30. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
31. Yanagawa, M. et al. Single-molecule diffusion-based estimation of ligand effects on G protein–coupled receptors. *Sci. Signal.* **11**, eaao1917 (2018).
32. Crouse, M. S., Nowak, R. D. & Baraniuk, R. G. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process* **46**, 886–902 (1998).
33. Durand, J. B., Gonçalvès, P. & Guédon, Y. Computational methods for hidden Markov tree models-An application to wavelet trees. *IEEE Trans. Signal Process* **52**, 2551–2560 (2004).
34. Choi, H. & Baraniuk, R. G. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans. Image Process* **10**, 1309–1321 (2001).
35. Bykova, N. A., Favorov, A. V. & Mironov, A. A. Hidden Markov models for evolution and comparative genomics analysis. *PLoS One* **8**, e65012–e65012 (2013).
36. Olariu, V. et al. Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics* **25**, 2824–2830 (2009).
37. Nakashima, S., Sughiyama, Y. & Kobayashi, T. J. Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics* **36**, 2829–2838 (2020).
38. Lund, B., Kristjansen, P. E. G. & Hansen, H. H. Clinical and preclinical activity of 2′,2′- difluorodeoxycytidine (gemcitabine). *Cancer Treat. Rev.* https://doi.org/10.1016/0305-7372(93)90026-N (1993).
39. Bolstad, B. M. Comparing some iterative methods of parameter estimation for censored gamma data. (1998).
40. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
41. Chao, H. X. et al. Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Mol. Syst. Biol.* **15**, e8604–e8604 (2019).
42. Lee, J. A. et al. Microbial phenotypic heterogeneity in response to a metabolic toxin: Continuous, dynamically shifting distribution of formaldehyde tolerance in Methylobacterium extorquens populations. *PLoS Genet* **15**, e1008458–e1008458 (2019).
43. van Boxtel, C., van Heerden, J. H., Nordholt, N., Schmidt, P. & Bruggeman, F. J. Taking chances and making mistakes: non-genetic phenotypic heterogeneity and its consequences for surviving in dynamic environments. *J. R. Soc. Interface* **14**, 20170141 (2017).
44. Tang, L., Wang, Y., Strom, A., Gustafsson, J. A. & Guan, X. Lapatinib induces p27Kip1-dependent G1 arrest through both transcriptional and post-translational mechanisms. *Cell Cycle* https://doi.org/10.4161/cc.25728 (2013).
45. Gross, S. M. et al. A LINCS microenvironment perturbation resource for integrative assessment of ligand-mediated molecular and phenotypic responses. https://doi.org/10.1101/2021.08.06.455429 (2021).
46. Brock, A., Chang, H. & Huang, S. Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet.* **10**, 336–342 (2009).
47. Fan, Y. & Meyer, T. Molecular control of cell density-mediated exit to quiescence. *Cell Rep.* **36**, 109436 (2021).
48. Zhu, X. et al. Autophagy stimulates apoptosis in HER2-overexpressing breast cancers treated by lapatinib. *J. Cell. Biochem.* **114**, 2643–2653 (2013).
49. Johnson, T. I. et al. Quantifying cell cycle-dependent drug sensitivities in cancer using a high throughput synchronisation and screening approach. *EBioMedicine* **68**, 103396 (2021).
50. Campbell, K. R. & Yau, C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat. Commun.* **9**, 1–12 (2018). *2018 91*.
51. Lartillot, N. A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data. *Bioinformatics* **30**, 488–496 (2013).
52. Ding, J. et al. Reconstructing differentiation networks and their regulation from time series single-cell expression data. *Genome Res.* **28**, 383–395 (2018).
53. Tentner, A. R. et al. Combined experimental and computational analysis of DNA damage signaling reveals context-dependent roles for Erk in apoptosis and G1/S arrest after genotoxic stress. *Mol. Syst. Biol.* **8**, 568 (2012).
54. Hata, A. N. et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* **22**, 262–269 (2016).
55. Bhang, H. C. et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* **21**, 440–448 (2015).
56. Burke, R. T. & Orth, J. D. Through the Looking Glass: Time-lapse Microscopy and Longitudinal Tracking of Single Cells to Study Anti-cancer Therapeutics. *J. Vis. Exp.* **14**, 53994 (2016).
57. Han, K. et al. CRISPR screens in cancer spheroids identify 3D growth-specific vulnerabilities. *Nature* **580**, 136–141 (2020).
58. Schwartz, A. D. et al. A biomaterial screening approach reveals microenvironmental mechanisms of drug resistance. *Integr. Biol. (Camb).* **9**, 912–924 (2017).
59. Emert, B. L. et al. Variability within rare cell states enables multiple paths toward drug resistance. *Nat. Biotechnol.* **39**, 865–876 (2021).
60. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2017).
61. Gross, S. M. et al. Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *bioRxiv* 2020.07.24.219907 (2021).
62. SL, S. et al. The proliferation-quiescence decision is controlled by a bifurcation in CDK2 activity at mitotic exit. *Cell* **155**, 369 (2013).
63. Meijering, E., Dzyubachyk, O. & Smal, I. Methods for cell and particle tracking. *Methods in Enzymology* https://doi.org/10.1016/B978-0-12-391857-4.00009-4 (2012).
64. Ephraim, Y. & Merhav, N. Hidden Markov processes. *IEEE Trans. Inf. Theory* **48**, 1518–1569 (2002).
65. Levinson, S. E., Rabiner, L. R. & Sondhi, M. M. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell Syst. Tech. J* **62**, 1035–1074 (1983).
66. Devijver, P. A. Baum's forward-backward algorithm revisited. *Pattern Recognit. Lett.* **3**, 369–373 (1985).
67. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020). *2020 173*.
68. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2011).
69. Vallender, S. S. Calculation of the Wasserstein Distance Between Probability Distributions on the Line. *Theory Probab. & Its Appl* **18**, 784–786 (1974).
70. Mohammadi, F., Visagan, S., Lagarde, J. & Meyer, A. S. Meyer-lab/tHMM. https://doi.org/10.5281/ZENODO.7195355 (2022).

Acknowledgements

This work was supported by the Jayne Koskinas Ted Giovanis Foundation for Health and Policy, NIH U01-CA215709 (A.S.M.), NIH U54-CA209988 (L.M.H.), NIH U54-HG008100 (L.M.H.). The authors thank Scott Taylor for his critical feedback that helped to improve the manuscript. The authors thank Ali Farhat, Adam Weiner, and Nikan Namiri for early exploratory work.

Author contributions

A.S.M. and L.M.H. conceived of the study; A.S.M. conceived of the model; A.S.M, F.M., S.V. designed model; A.S.M., F.M., J.L., L.K., S.V. performed computational experiments; S.M.G. performed the experiments; F.M., J.L., L.K., S.M.G. conducted data analysis; A.S.M. and L.M.H. supervised the research; all authors wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-04208-9.

Correspondence and requests for materials should be addressed to Aaron S. Meyer.

Peer review information *Communications Biology* thanks Barbara Bravi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: George Inglis. This article has been peer reviewed as part of Springer Nature's Guided Open Access initiative.

A lineage tree-based hidden Markov model quantifies cellular heterogeneity and plasticity

Supplementary Information

Farnaz Mohammadi[1], Shakthi Visagan[1], Sean M. Gross[2], Luka Karginov[3], JC Lagarde[1], Laura M. Heiser[2], and Aaron S. Meyer[1,4,5,6]

[1]Department of Bioengineering, University of California, Los Angeles, USA.

[2]Department of Biomedical Engineering, Oregon Health and Science University, Portland, USA.

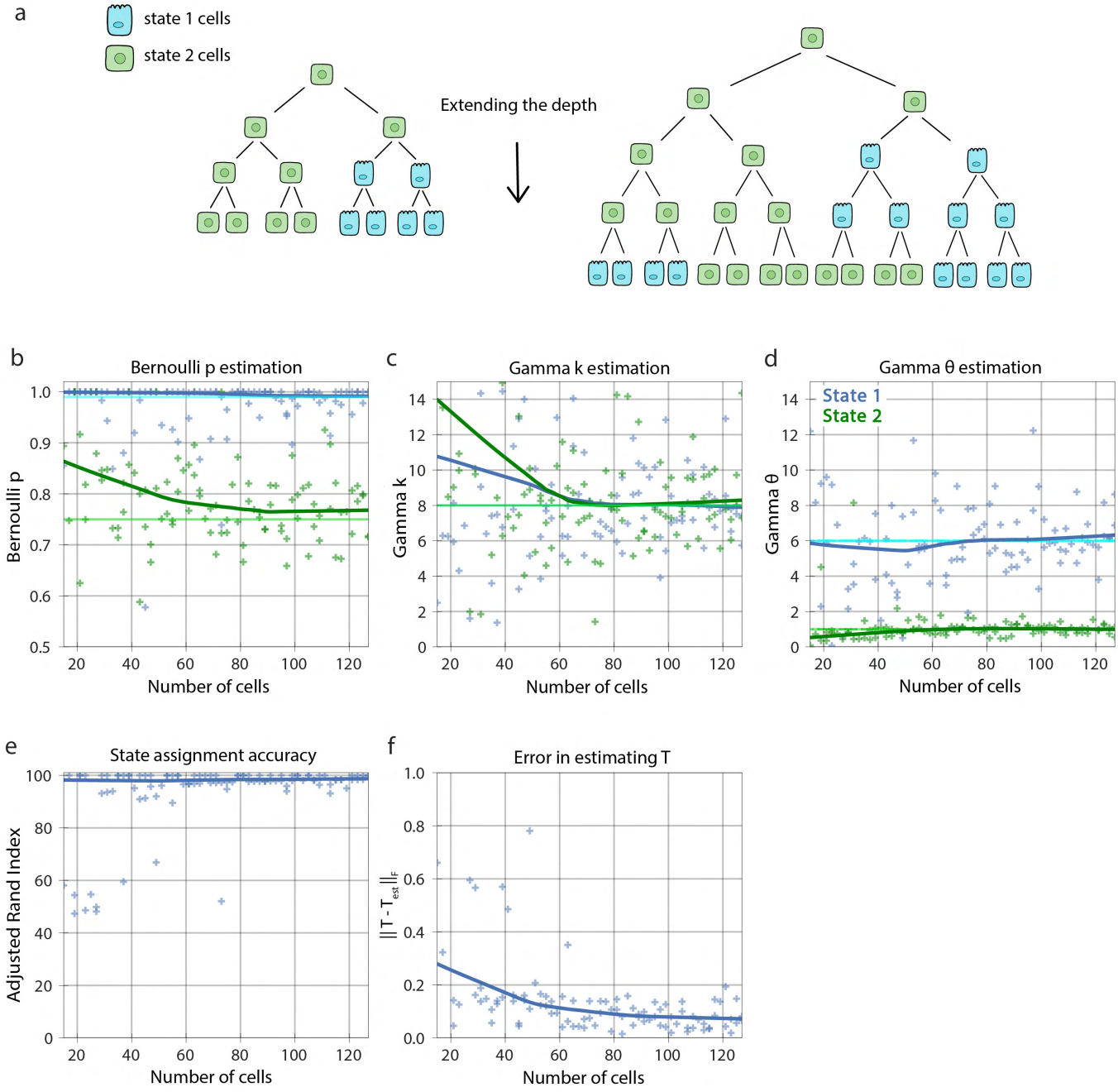
[3]Department of Bioengineering, University of Illinois, Urbana Champaign, USA.

[4]Department of Bioinformatics, University of California, Los Angeles, USA.

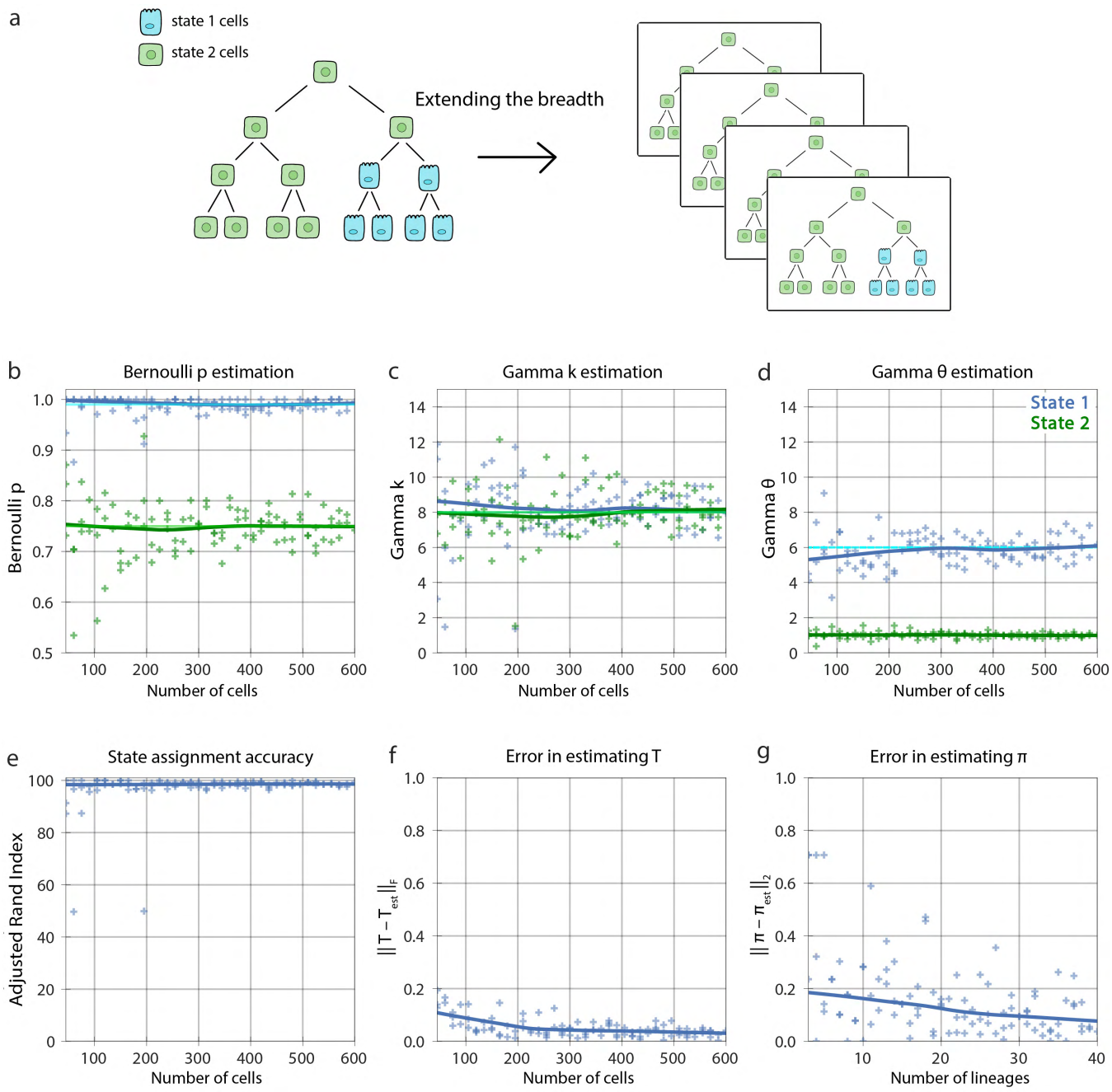
[5]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, USA.

[6]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, USA.

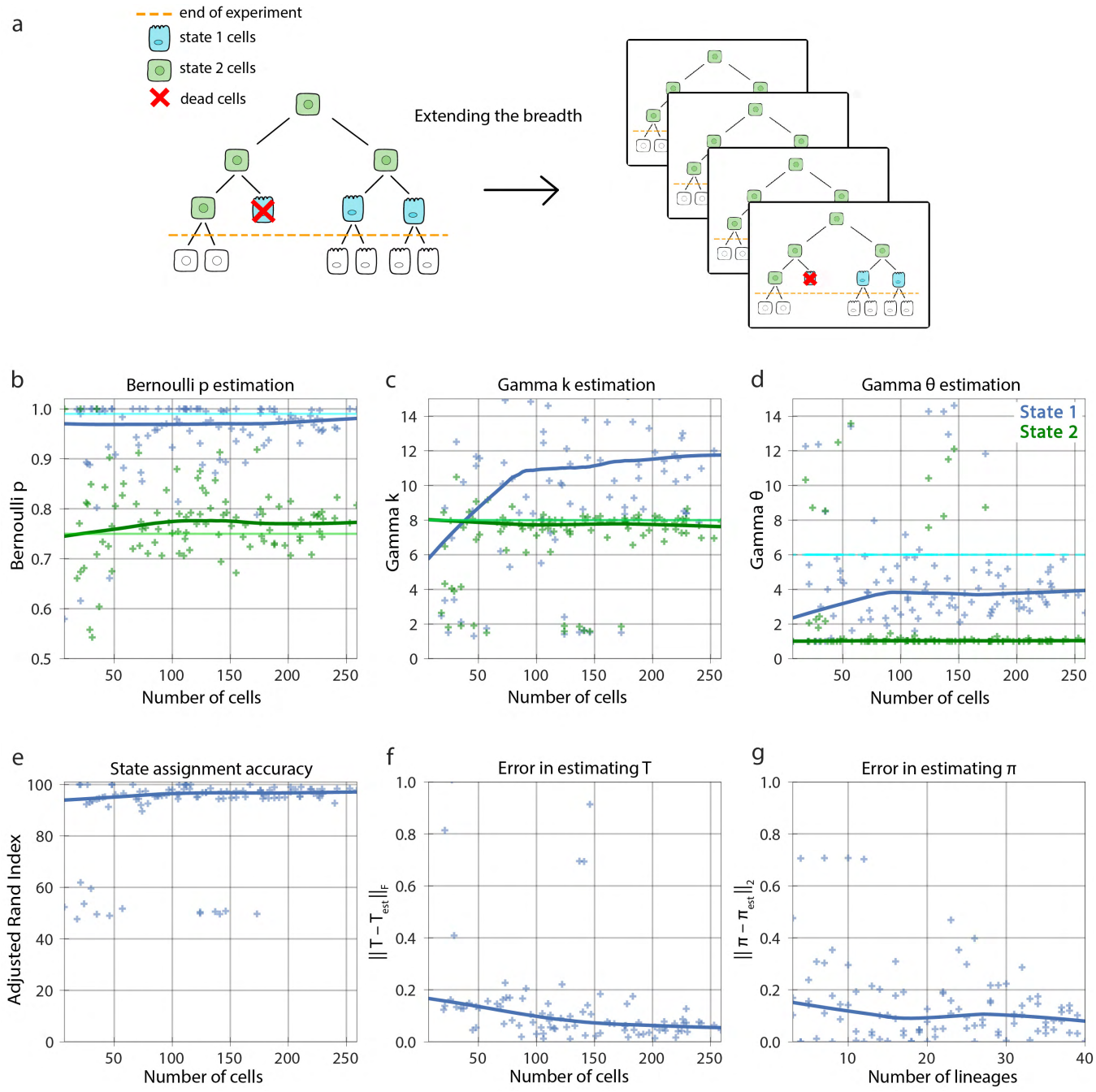
Supplementary Figures



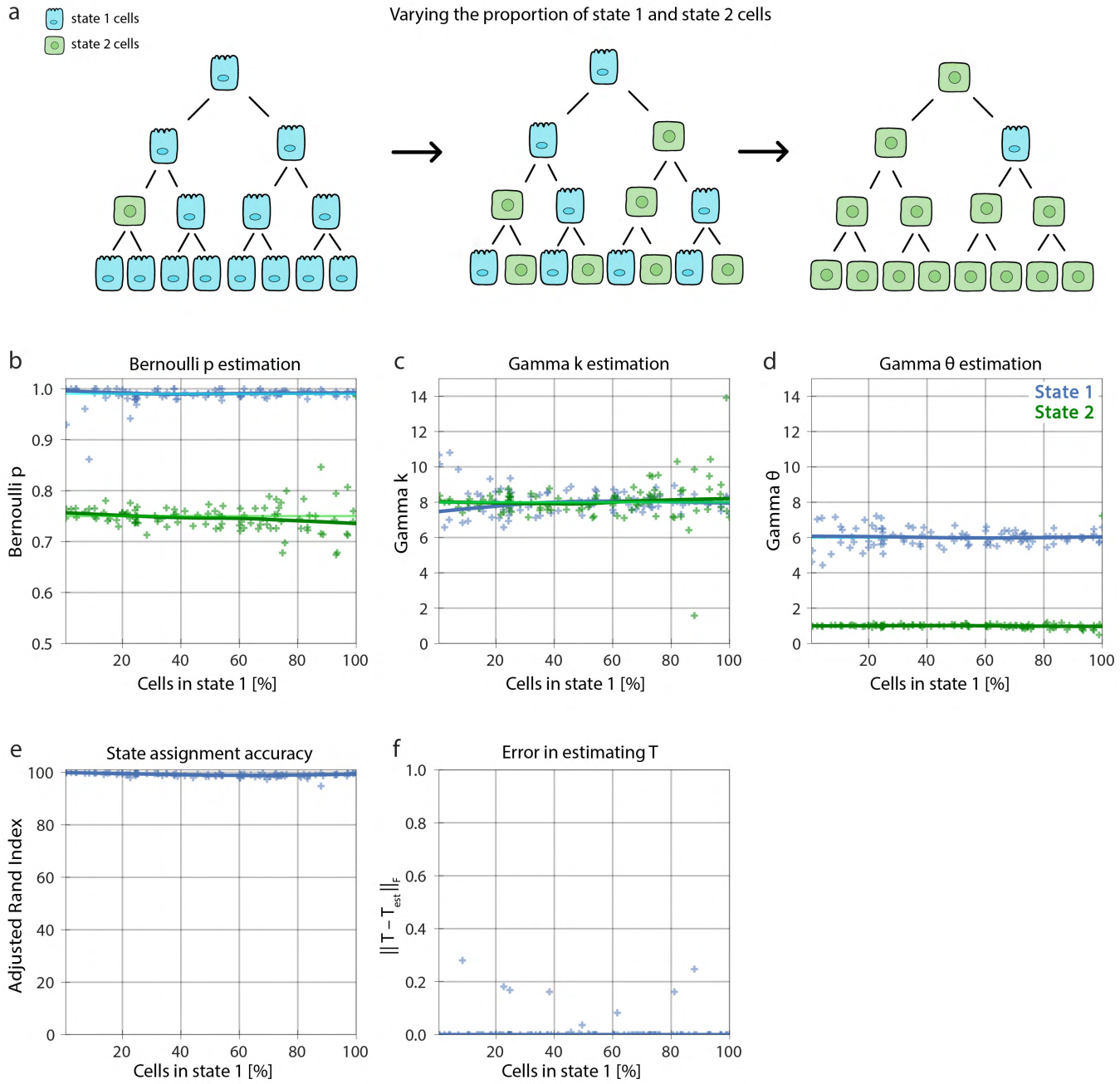
Supplementary Figure 1: Performance on synthetic uncensored single lineages of increasing size with two states. (a) Visual representation of increasing the lineage size with two states. **(b)** The Bernoulli parameter for states 1 and 2 as the number of cells increase. **(c)** The shape parameter, k, and **(d)** scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. **(e)** The state assignment accuracy as the number of cells increases. **(f)** The error in the estimate of the transition probability matrix, T as the number of cells increase. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



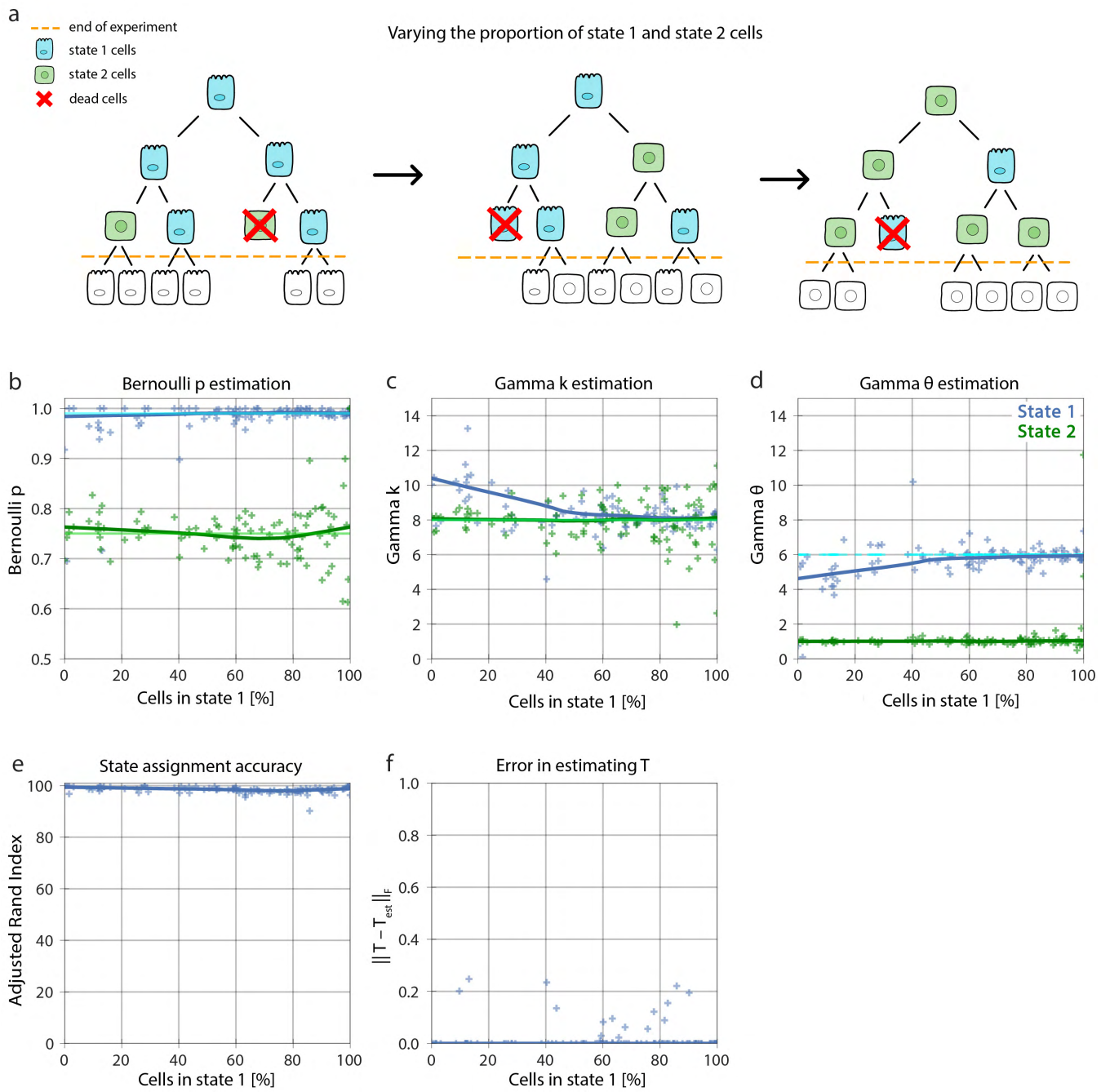
Supplementary Figure 2: Performance on synthetic uncensored lineages of increasing number with two states. (a) Visualization of increasing the number of fully observed lineages. **(b)** The Bernoulli parameter for states 1 and 2 as the number of cells increase. **(c)** The shape parameter, k, and **(d)** scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. **(e)** The state assignment accuracy as the number of cells increases. **(f)** The error in the estimate of the transition probability matrix, T, as the number of cells increase. **(g)** The errors in the estimate of the initial probability matrix, π, as the number of lineages increase. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



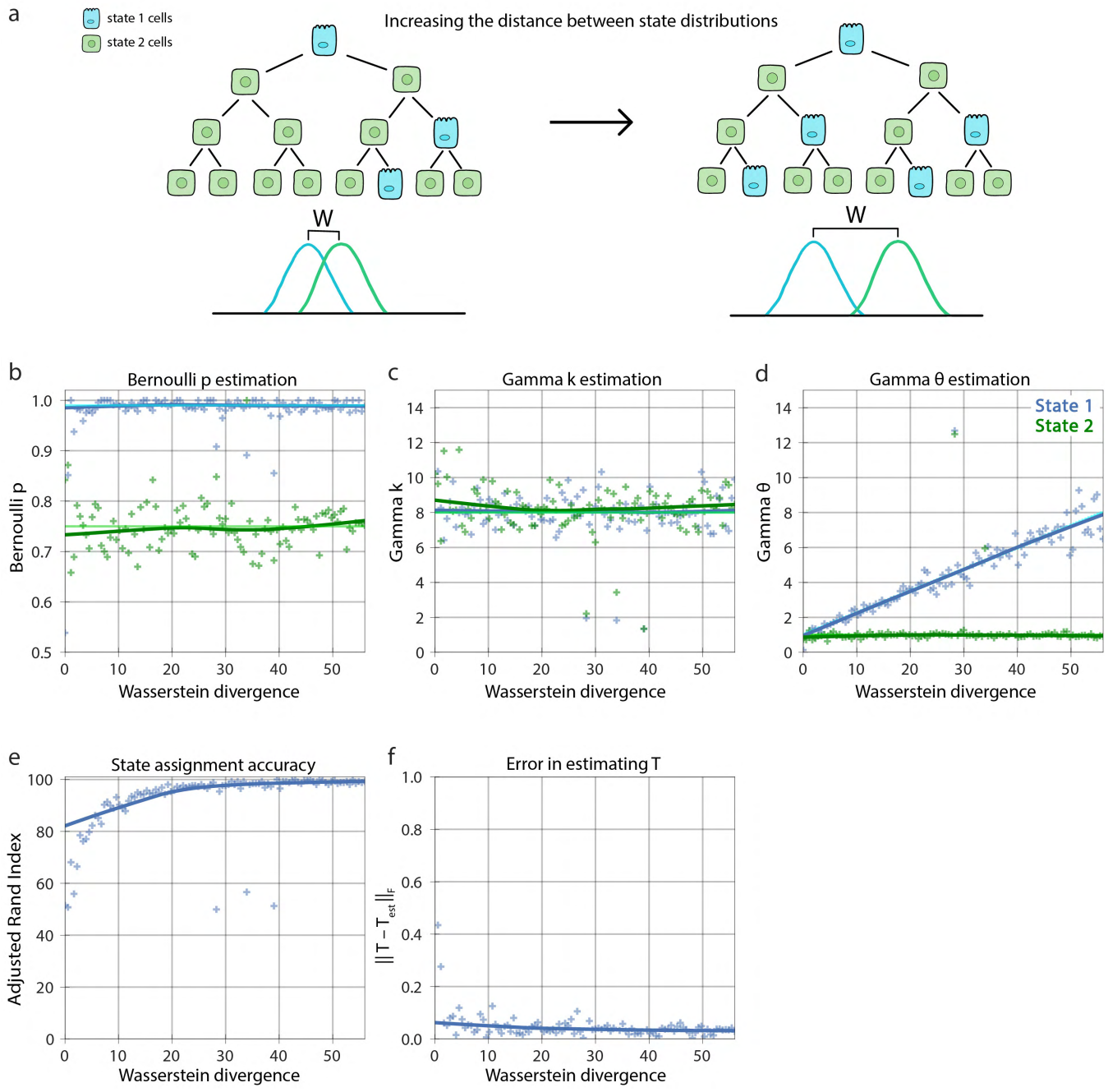
Supplementary Figure 3: Performance on synthetic censored lineages of increasing number with two states. (a) Visualization of the number of censored lineages increasing. **(b)** The Bernoulli parameter for states 1 and 2 as the number of cells increase. **(c)** The shape parameter, k, and **(d)** scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. **(e)** State assignment accuracy as the number of cells increase. **(f)** The error in the estimate of the transition probability matrix, T, as the number of cells increase. **(g)** The errors in the estimate of the initial probability matrix, π, as the number of lineages increase. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



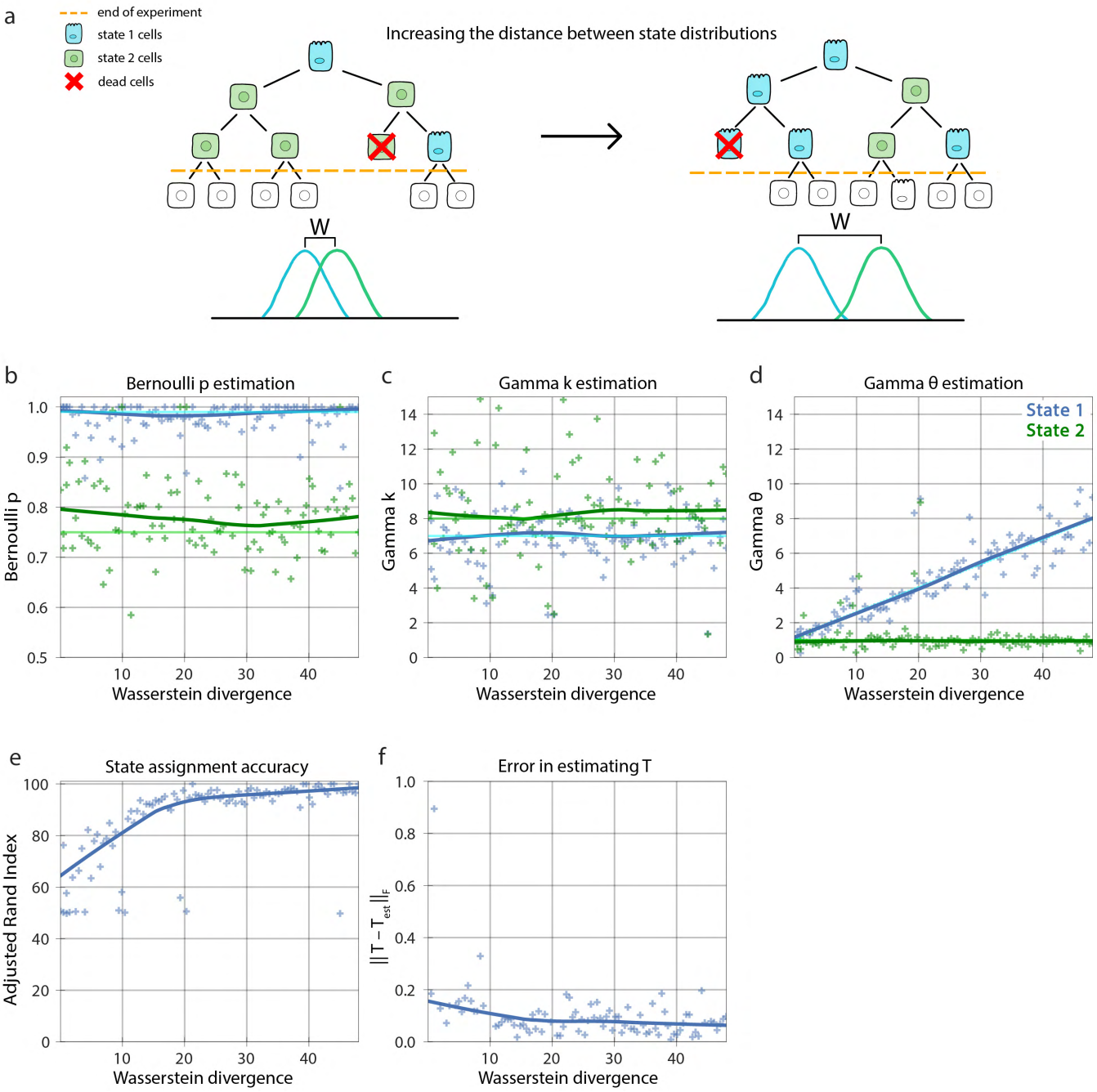
Supplementary Figure 4: Model performance relative to the presence of each state for an uncensored lineage in a synthetic two-state dataset. (a) Visualization of the distribution of cells in the lineage transitioning between state 1 and state 2. **(b)** The Bernoulli parameter for states 1 and 2 as the proportion of cells in state 1 increase. **(c)** The shape parameter, k, and **(d)** scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the proportion of cells in state 1 increase. **(e)** The state assignment accuracy as the proportion of cells in state 1 increase. **(f)** The errors in the estimate of the transition probability matrix, T, as proportion of cells in state 1 increase. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



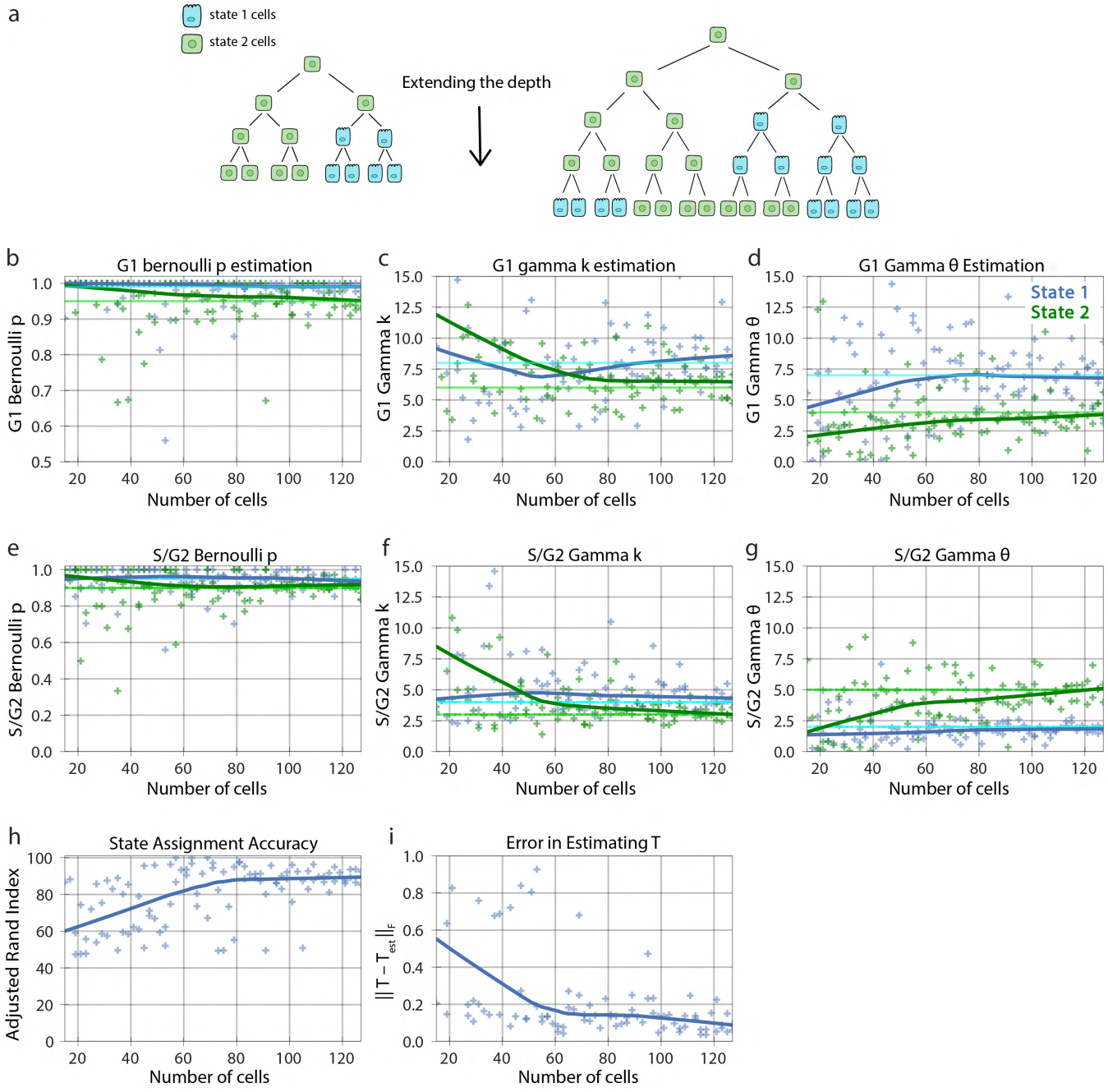
Supplementary Figure 5: Change in model performance when varying the presence of a state for a censored lineage in a synthetic two-state dataset. (a) Visualization of the proportion of cells in a censored lineage transitioning between state 1 and 2. **(b)** The Bernoulli parameter for states 1 and 2 as the proportion of cells in state 1 increase. **(c-d)** The shape parameter, k, and scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the proportion of cells in state 1 increase. **(e)** The state assignment accuracy as the proportion of cells in state 1 increase. **(f)** The error in the estimate of the transition probability matrix, T, as the proportion of cells in state 1 increase. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



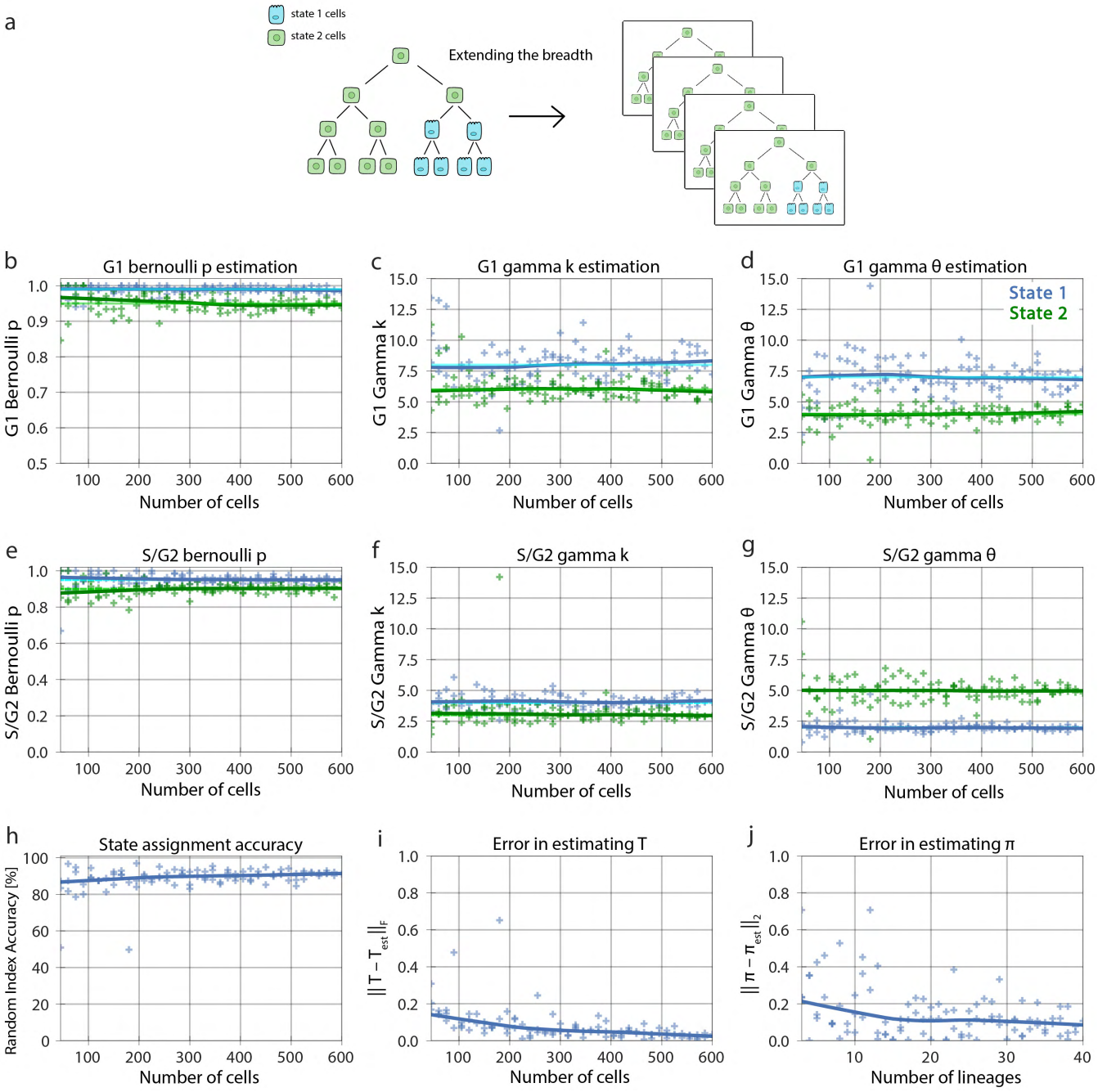
Supplementary Figure 6: Change in model performance when varying state distribution similarity for an uncensored population of lineages in a synthetic two-state dataset. (a) Visualization of the Wasserstein divergence increasing as the state distribution in the lineage varies. **(b)** The cell fate Bernoulli parameter compared to the true value. **(c)** The shape parameter, k, **(d)** the scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime compared to the true values. **(e)** The state assignment accuracy as the Wasserstein divergence increases. **(f)** The errors in the estimate of the transition probability matrix, T, as the Wasserstein divergence increases. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



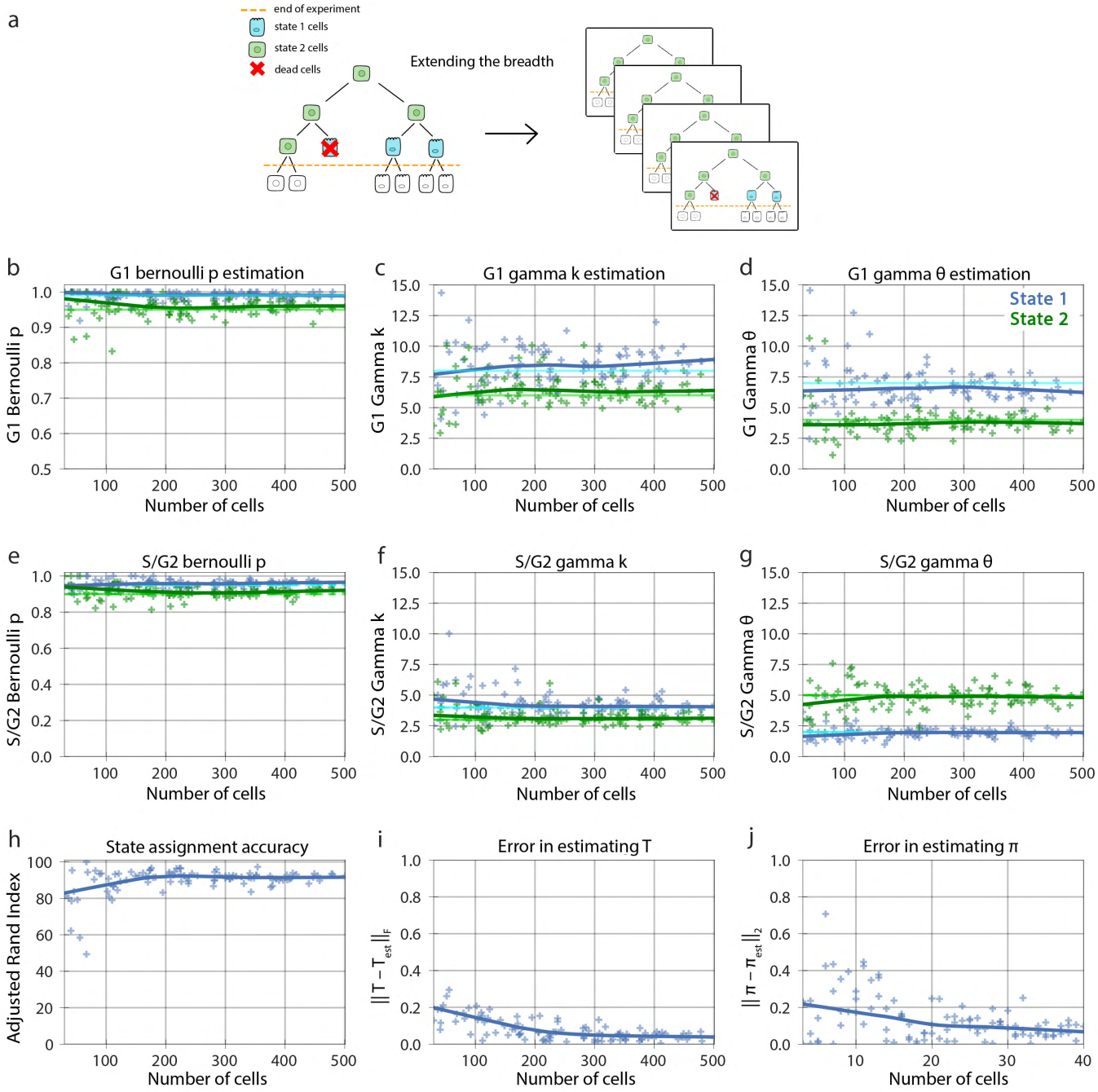
Supplementary Figure 7: Change in model performance when varying state distribution similarity for a censored population of lineages in a synthetic two-state dataset. (a) Visualization of the Wasserstein divergence increasing as the state distribution in the censored lineage varies. **(b)** The Bernoulli, **(c)** Gamma shape, and **(d)** Gamma scale parameters for states 1 and 2 as the Wasserstein divergence increases. **(e)** The state assignment accuracy as the Wasserstein divergence increases. **(f)** The error in the estimate of the transition probability matrix, T, as the Wasserstein divergence increases. In **(b-d)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



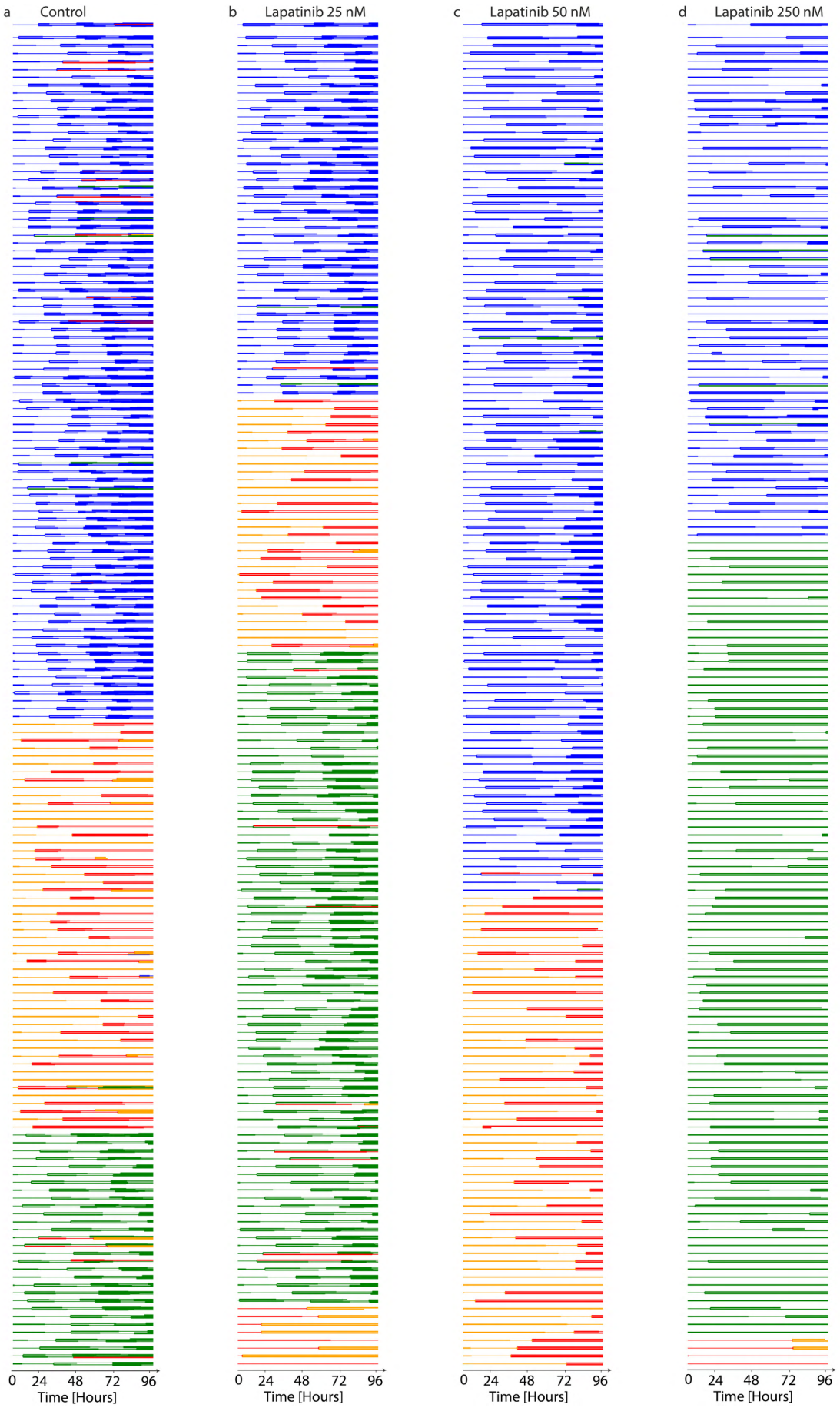
Supplementary Figure 8: Performance of increasing cell numbers in an uncensored single lineage in a synthetic two-state dataset. (a) Visualization of a single lineage with increasing cell number. **(b)** The G1 phase Bernoulli, **(c)** Gamma shape, and **(d)** Gamma scale parameters for states 1 and 2 as the number of cells increase. **(e)** The S/G2 phase Bernoulli, **(f)** Gamma shape, and **(g)** Gamma scale parameter for states 1 and 2 as the number of cells increase. **(h)** The state assignment accuracy as the number of cells increases. **(i)** The errors in the estimate of the transition probability matrix, T, as the number of cells increase. In **(b-g)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



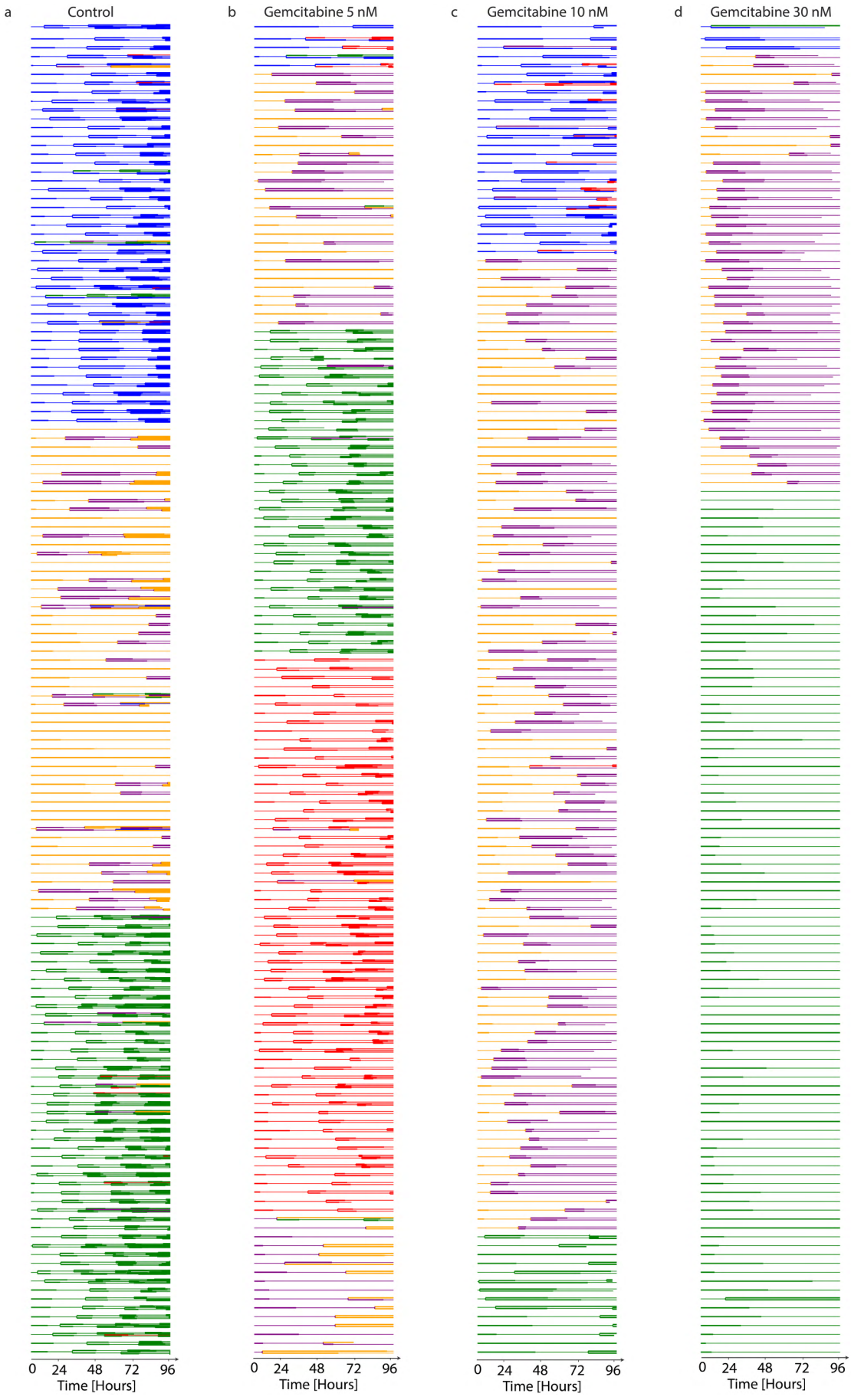
Supplementary Figure 9: Performance of increasing lineage numbers in an uncensored population in a synthetic two-state dataset. (a) Visualization of the number of uncensored lineages within a population increasing. **(b)** The G1 phase Bernoulli, **(c)** Gamma shape, and **(d)** Gamma scale parameters for states 1 and 2 as the number of cells increase. **(e)** The S/G2 phase Bernoulli, **(f)** Gamma shape, and **(g)** Gamma scale parameters for states 1 and 2 as the number of cells increase. **(h)** The state assignment accuracy as the number of cells increases. **(i)** The error in the estimate of the transition probability matrix, T, as the number of cells increase. **(j)** The errors in the estimate of the initial probability matrix, π, as the number of lineages increase. In **(b-g)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



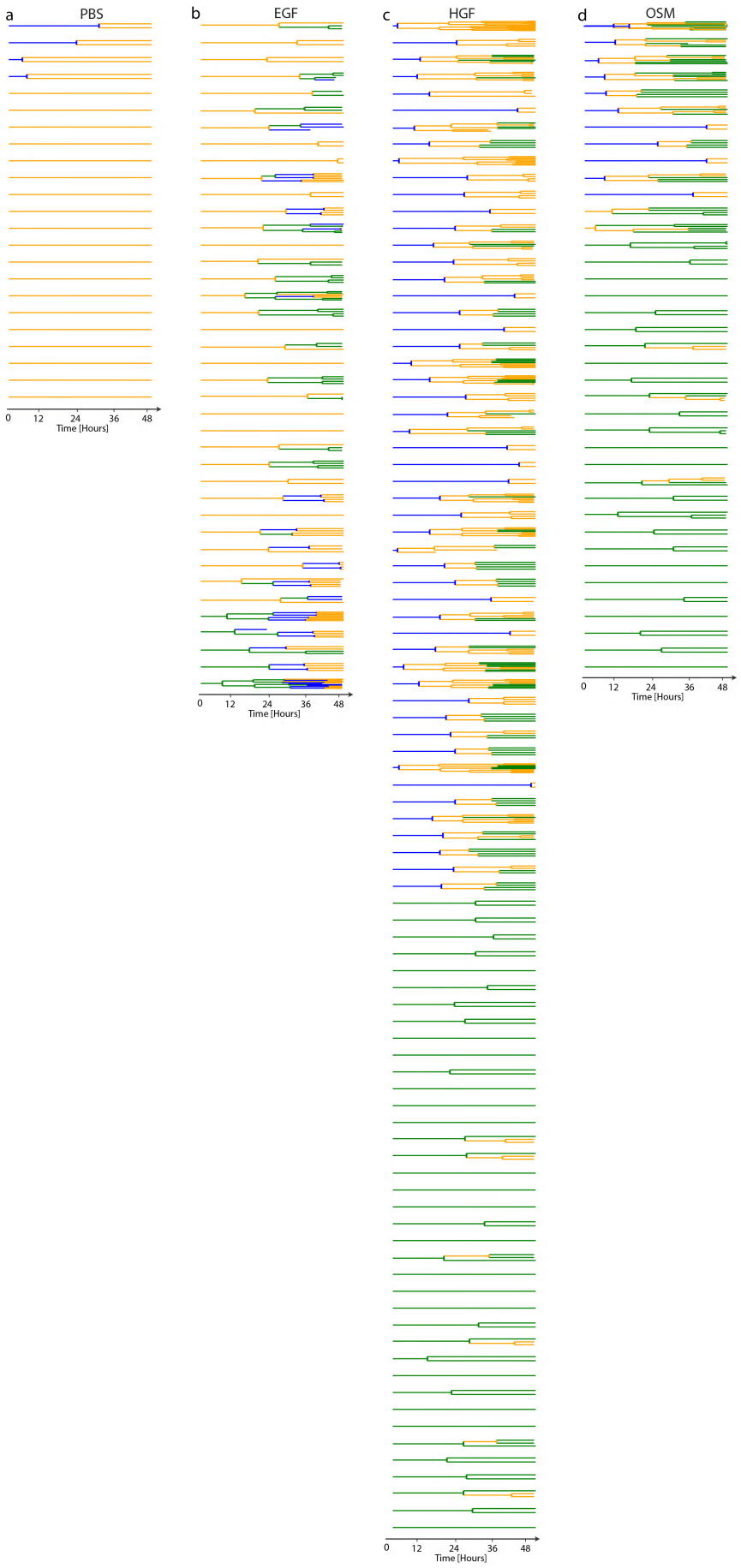
Supplementary Figure 10: Performance of increasing lineage numbers in a censored population in a synthetic two-state dataset. (a) Visualization of the number of censored lineages within a population increasing. **(b)** The G1 phase Bernoulli, **(c)** Gamma shape, and **(d)** Gamma scale parameters for states 1 and 2 as the number of cells increase. **(e)** The S/G2 phase Bernoulli, **(f)** Gamma shape, and **(g)** Gamma scale parameter for states 1 and 2 as the number of cells increase. **(h)** The state assignment accuracy as the number of cells increases. **(i)** The error in the estimate of the transition probability matrix, T, as the number of cells increase. **(j)** The error in the estimate of the initial probability matrix, π, as the number of lineages increase. In (b-g) the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.


a
Control
b
Lapatinib 25 nM
c
Lapatinib 50 nM
d
Lapatinib 250 nM
0
24
48
72
96
Time [Hours]

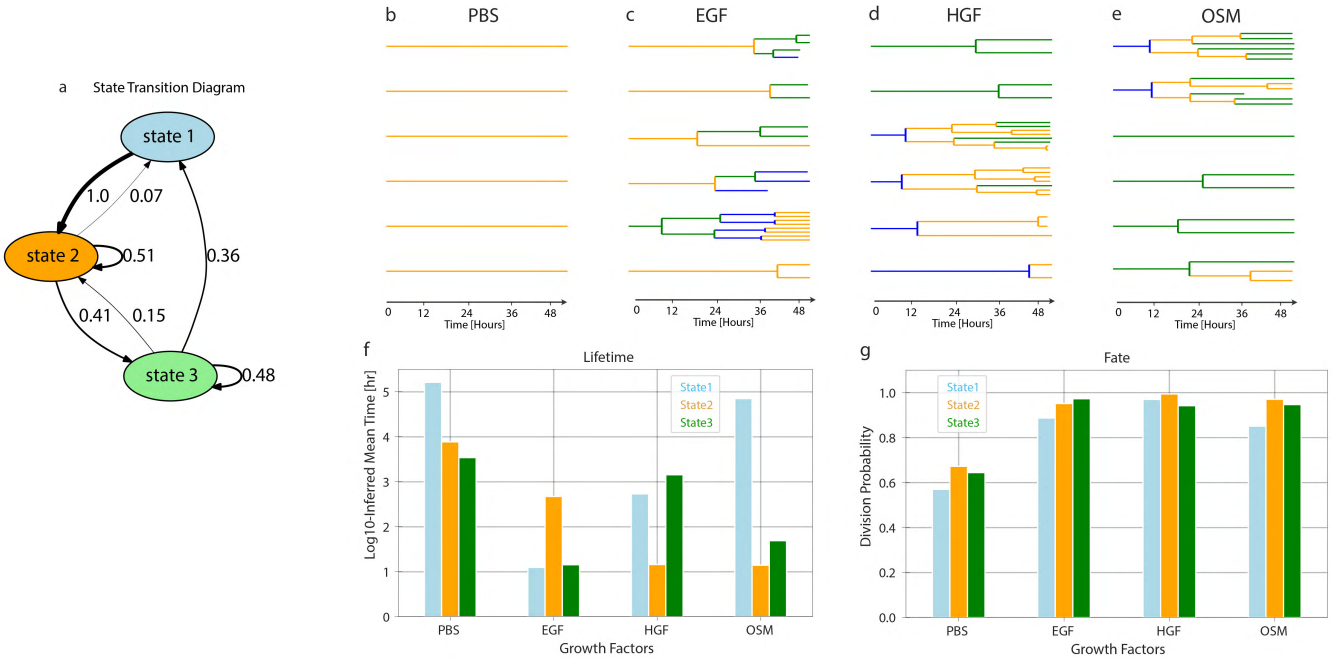
Supplementary Figure 11: The single cell data after fitting and state assignment for lapatinib-treated lineages. (a) Control, **(b)** 25 nM, **(c)** 50 nM, and **(d)** 250 nM lapatinib treatment. Each line represents a cell, and the length of the line represents the cell's lifetime. G1 and S/G2 phase durations are depicted by thick and thin lines, respectively. Termination and branching indicate cell death and division, respectively. Different colors show the state of each cell. Blue: state 1, orange: state 2, green: state 3, red: state 4, purple: state 5, and olive: state 6.


a
Control
b
Gemcitabine 5 nM
c
Gemcitabine 10 nM
d
Gemcitabine 30 nM
0 24 48 72 96
Time [Hours]

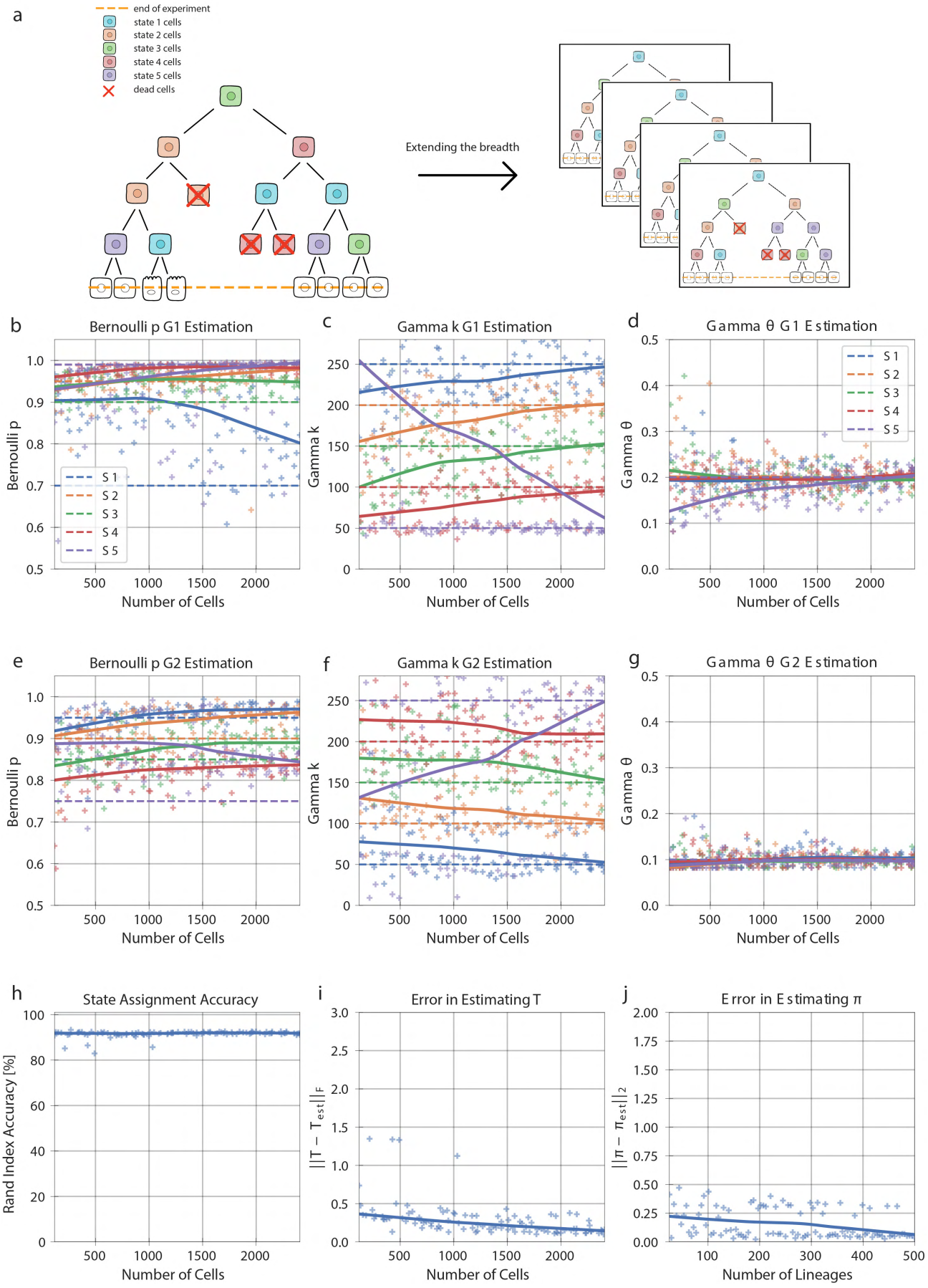
Supplementary Figure 12: The single cell data after fitting and state assignment for gemcitabine-treated lineages. (a) Control, **(b)** 5 nM, **(c)** 10 nM, and **(d)** 30 nM gemcitabine treatment. Each line is a cell, and the length of the line represents the cell's lifetime. G1 and S/G2 phase durations are depicted by thick and thin lines, respectively. Termination and branching indicate cell death and division, respectively. Different colors show the state of each cell. Blue: state 1, orange: state 2, green: state 3, red: state 4, and purple: state 5.


a
PBS
b
EGF
c
HGF
d
OSM
0
12
24
36
48
Time [Hours]

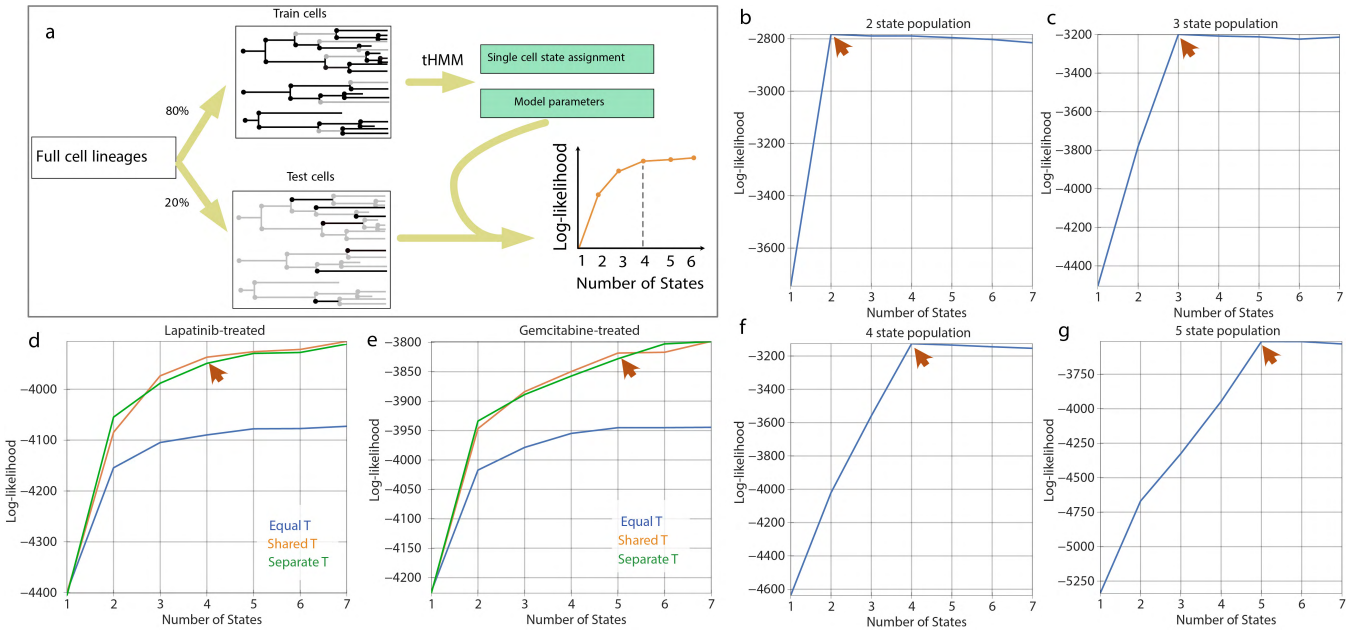
Supplementary Figure 13: The single cell data after fitting and state assignment for growth factor treated MCF10A lineages. (a) PBS, **(b)** EGF, **(c)** HGF, and **(d)** OSM treatments. Each line represents a cell, and the length of the line represents the cell's lifetime. Termination and branching indicate cell death and division, respectively. Different colors show the state of each cell. Blue: state 1, orange: state 2, and green: state 3.



Supplementary Figure 14: State-specific emissions of the growth factor treated MCF10A population. (a) State transition graph showing the probability of state transitions among the states. Transitions with less than 0.03 probability have been removed. **(b-e)** A sample of lineage trees after fitting the model and state assignment (PBS, EGF, HGF, and OSM). **(f-g)** The inferred log10 mean time of cell cycle durations for different conditions at different states. **(h-i)** The Bernoulli parameter, which is the probability of dividing versus dying for different conditions and states.


a
end of experiment
state 1 cells
state 2 cells
state 3 cells
state 4 cells
state 5 cells
dead cells
Extending the breadth
b
Bernoulli p G1 Estimation
Bernoulli p
Number of Cells
c
Gamma k G1 Estimation
Gamma k
Number of Cells
d
Gamma θ G1 Estimation
Gamma θ
Number of Cells
S 1
S 2
S 3
S 4
S 5
e
Bernoulli p G2 Estimation
Bernoulli p
Number of Cells
f
Gamma k G2 Estimation
Gamma k
Number of Cells
g
Gamma θ G2 Estimation
Gamma θ
Number of Cells
h
State Assignment Accuracy
Rand Index Accuracy [%]
Number of Cells
i
Error in Estimating T
$||T - T_{est}||_F$
Number of Cells
j
Error in Estimating π
$||\pi - \pi_{est}||_2$
Number of Lineages

Supplementary Figure S15: Performance of increasing lineage numbers in a censored population in a synthetic five-state dataset. (a) Visualization of the number of censored lineages within a population increasing. **(b)** The G1 phase Bernoulli, **(c)** Gamma shape, and **(d)** Gamma scale parameters for each state as the number of cells increase. **(e)** The S/G2 phase Bernoulli, **(f)** Gamma shape, and **(g)** Gamma scale parameter for each state as the number of cells increase. **(h)** The state assignment accuracy as the number of cells increases. **(i)** The error in the estimate of the transition probability matrix, T, as the number of cells increase. **(j)** The error in the estimate of the initial probability matrix, π, as the number of lineages increase. In **(b-g)** the light green and blue solid lines show the true value of the parameters, and the dark green and blue solid lines show the Lowess trend of estimations.



Supplementary Figure S16: Performance of increasing lineage numbers in a censored population in a synthetic five-state dataset. (a) Visualization of the cross-validation process. 80% of cells are randomly selected as the training set, and the remaining cells serve as the test set. The log likelihood of the observations of the test given the trained model will determine the optimum number of states. **(b–d)** The log likelihood from the cross-validation approach for lapatinib and gemcitabine-treated AU565 cells. “Equal T” refers to the scenario where all transitions to and from each state are equal, simulating the absence of inheritance, “Shared T” refers to the scenario where we estimate a transition matrix that is shared between all concentrations, and “Separate T” simulates the scenario where each concentration has a separate transition matrix. **(f–g)** The log likelihood plot for a 2, 3, 4, and 5 state synthetic model using the cross-validation scheme. The arrows in **(b-g)** point to the optimal number of states.

Supplementary Table 1. State distribution parameters for cell cycle phase non-specific observation for a two-state population.

State	Bern p	Shape	Scale
State 1	0.99	8	6
State 2	0.75	8	1

Supplementary Table 2. State distribution parameters for cell cycle phase-specific observation for a two-state population.

State	G1 bern	S/G2 bern	G1 shape	G1 scale	S/G2 shape	S/G2 scale
State 1	0.99	0.95	8	7	4	2
State 2	0.95	0.9	6	4	3	5

Supplementary Table 3. State distribution parameters for cell cycle phase non-specific observation for a five-state population.

State	G1 bern	S/G2 bern	G1 shape	G1 scale	S/G2 shape	S/G2 scale
State 1	0.7	0.99	250	0.2	50	0.1
State 2	0.95	0.9	200	0.2	100	0.1
State 3	0.9	0.85	150	0.2	150	0.1
State 4	0.99	0.75	100	0.2	200	0.1
State 5	0.99	0.75	50	0.2	250	0.1