Title: Integrating transcriptomics and bulk time course data into a mathematical framework to describe and predict therapeutic resistance in cancer

Authors: Kaitlyn Johnson[1], Grant R. Howard[1], Daylin Morgan[1], Eric A. Brenner[1,2], Andrea L. Gardner[1], Russell E. Durrett[1,2], William Mo[1], Aziz Al'Khafaji[1,2], Eduardo D. Sontag[4,5,6], Angela M. Jarrett[3,7], Thomas E. Yankeelov[1,3,7,8,9], Amy Brock[1, 2, 3]

Affiliations:

1) Department of Biomedical Engineering, The University of Texas at Austin
2) Institute for Cellular and Molecular Biology, The University of Texas at Austin
3) Livestrong Cancer Institutes, Dell Medical School, The University of Texas at Austin
4) Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115, USA
5) Department of Bioengineering, Northeastern University, Boston, MA, 02115, USA.
6) Laboratory of Systems Pharmacology, Program in Therapeutics Science, Harvard Medical School, Boston, MA, 02115, USA
7) Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin
8) Department of Diagnostic Medicine, The University of Texas at Austin
9) Department of Oncology, The University of Texas at Austin

Corresponding author: Amy Brock, amy.brock@utexas.edu

Key words: mathematical modeling, mathematical oncology, chemoresistance, intratumor heterogeneity, population dynamics

Abstract

A significant challenge in the field of biomedicine is the development of methods to integrate the multitude of dispersed data sets into comprehensive frameworks to be used to generate optimal clinical decisions. Recent technological advances in single cell analysis allow for high-dimensional molecular characterization of cells and populations, but to date, few mathematical models have attempted to integrate measurements from the single cell scale with other types of longitudinal data. Here, we present a framework that actionizes static outputs from a machine learning model and leverages these as measurements of state variables in a dynamic model of treatment response. We apply this framework to breast cancer cells to integrate single cell transcriptomic data with longitudinal bulk cell population (bulk time course) data. We demonstrate that the explicit

inclusion of the phenotypic composition estimate, derived from single cell RNA-sequencing data (scRNA-seq), improves accuracy in the prediction of new treatments with a concordance correlation coefficient (CCC) of 0.92 compared to a prediction accuracy of CCC = 0.64 when fitting on longitudinal bulk cell population data alone. To our knowledge, this is the first work that explicitly integrates single cell clonally-resolved transcriptome datasets with bulk time-course data to jointly calibrate a mathematical model of drug resistance dynamics. We anticipate this approach to be a first step that demonstrates the feasibility of incorporating multiple data types into mathematical models to develop optimized treatment regimens from data.

Introduction

The development of resistance to chemotherapy is a major cause of treatment failure in cancer. Intratumoral heterogeneity and phenotypic plasticity play significant roles in therapeutic resistance (1)(2) and individual cell measurements such as flow and mass cytometry (3) and scRNA-seq (4) have been used to capture and analyze this cell variability (5–8). Although these assays destructive nature can limit the time resolution of data acquisition, snapshot information alone has provided immense insight to the field: illuminating novel molecular insight about distinct subpopulations (9), developing detailed hypothesis about population structure (10), and even demonstrating the ability to predict clinical outcomes (1). However, outside of the field of differentiation (11), most information gleamed from “omics” data sets have not been directly linked to growth and treatment response dynamics of the bulk cell population—which are critical to understanding the dynamics of cancer progression.

Longitudinal bulk cell population data in cancer have been used to calibrate mathematical models of heterogeneous subpopulations (10,12,13) of cancer cells. These models describe cancer cells dynamically growing and responding to drug with differential growth rates and drug sensitivities. Knowledge of these model parameters have enabled the theoretical optimization of treatment protocols (14–16), and have been applied to prolong tumor control in both mice (10) and patients (12,17). Critical to the success of these modeling endeavors is the ability to identify and validate critical model parameters from available data (18). Identifiable and practical models are necessarily limited in their

capacity to explain biological complexity based on the availability and feasibility of longitudinal data, which is often limited to total tumor volume or total cell number in time. While complex relationships between distinct cell subpopulations is critical to some responses (9), the ability to track the relevant subpopulations longitudinally for model calibration and parameter estimation remains a challenge (19).

One way to resolve this challenge would be to work with both types of data and use them jointly to inform the calibration of a dynamic model. In this study, we sought to develop a flexible framework for integrating informatics outputs from high-throughput single-cell resolution data with bulk time-course data to demonstrate the feasibility of utilizing multimodal data sources in mathematical oncology. The integration of single cell data into a mathematical modeling framework has been successfully employed in the field of differentiation by quantifying the changing proportion of cells in distinct cell states over time (11). This approach is more complex in cancer, where the effects of exponential growth and death due to drug exposure results in changes in phenotypic composition that may be independent of directed transitions between cell states. To better understand these dynamics, we collect bulk time-course data throughout treatment with chemotherapy doxorubicin. We combine this with snapshots of lineage-traced scRNA-seq data and build a classifier to estimate phenotypic composition, via the proportion of sensitive and resistant cells, at distinct time points during treatment response. Despite differences in data acquisition, time resolution, and data uncertainty, we demonstrate that these two measurement sources can be used to estimate cell number in time and phenotypic composition in time, which can be compared to their corresponding model outputs. To account for different time resolutions in the measurement sources, we develop an integrated calibration scheme to incorporate both data types. We validate the model results by demonstrating that they can accurately predict the response dynamics to new treatment regimens. We propose this framework as a crucial next step towards combining tumor composition information with bulk time-course data to improve prediction and optimization of treatment outcomes.

Results

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

To describe and predict the dynamics of cancer cells in response to treatment, we chose to use a mathematical model that describes sensitive and resistant cell subpopulations growing, dying, and transitioning from the sensitive, *S*, to resistant, *R*, state as a direct result of treatment (15). This model was chosen because it represents a relatively simple phenomenological model of two subpopulations differing in their degree of drug sensitivity, that accounts for the ability of cells to transition directly from sensitive to resistant phenotypes following drug exposure, as has been observed in cancer cell systems (20).

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

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

In this model (Fig 1A), sensitive and resistant cells grow *via* a logistic growth hypothesis at their own intrinsic growth rates (r_S and r_R) and a joint carrying capacity (K), which will vary based on the experimental scenario: either taking the value of K_N for the carrying capacity of the cells in the bulk time course experiment or K_ϕ for the carrying capacity of the cells in the scRNA-seq experiment (Table 1, Supp Table S1). Sensitive and resistant cells are killed by the drug at a rate of d_S and d_R respectively, that is proportional to the number of cells in each subpopulation and the effective dose, $u(t)$, following the log-kill hypothesis. By definition, we set $d_S > d_R$ such that sensitive cells will be more susceptible to death due to treatment than resistant cells. Treatment drives cells from the sensitive subpopulation into the resistant subpopulation at a rate α, which is linearly proportional to the number of sensitive cells present and $u(t)$.

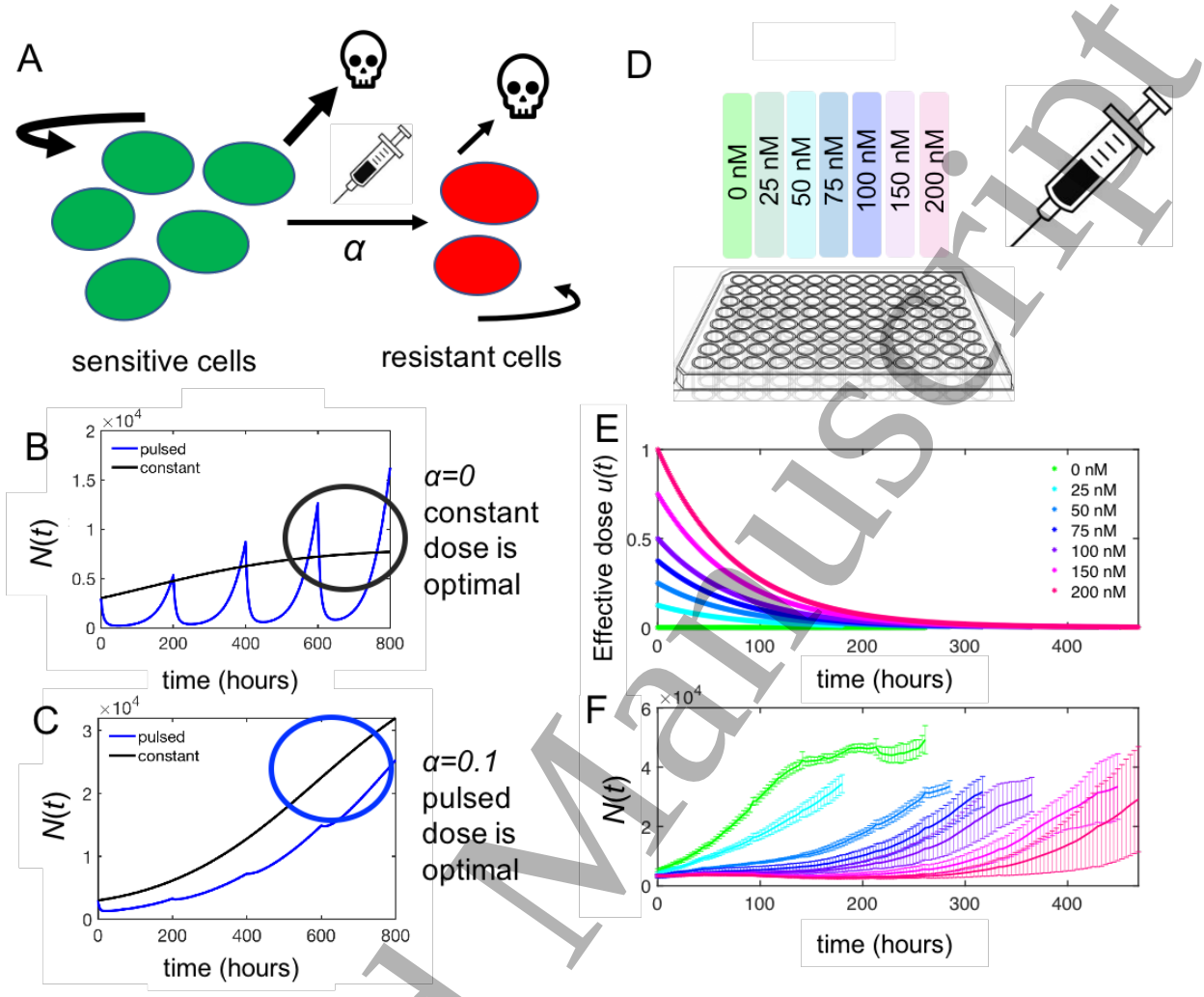


Fig 1. Mathematical Model of Treatment-induced Resistance and its Implications. A. Sketch of the model structure (Eq 1). The model describes sensitive and resistant subpopulations growing exponentially at independent growth rates. In response to treatment, sensitive and resistant cells are killed by the drug. The exposure to drug drives sensitive cells into the resistant phenotype. B. Example trajectory of model predicted total cell number in time for a constant dose (black) and a pulsed dose (blue) for the case where there is no drug-induced resistance (α = 0), indicating that the optimal treatment is the constant dose C. Example trajectory of model predicted total cell number in time for a constant dose (black) and a pulsed dose (blue) for the case where the drug does induce resistance (α > 0), indicating that in this case the optimal treatment is a pulsed treatment. D. Schematic of experimental set-up using time-resolved fluorescence microscopy to measure the number of MDA-MB-231 GFP labeled breast cancer cells in response to doxorubicin concentrations ranging from 0-200 nM treated for 24 hours and then allowed to recover in growth media. E. Estimated effective dose dynamics ($u(t)$) of the various pulse-treatments of doxorubicin. F. Measured number of cells in time, colored by drug concentration as in B, from six replicate wells. Error bars represent 95% confidence intervals around the mean cell number at each time point. Images were converted to cell number estimates every 4 hours. Time of monitoring ranged from 1 week (168 hours) for the untreated control to ~2.5 weeks (469 hours) for the 200 nM dose.

To incorporate time-dependent effects of a treatment on the cell population, we make a simple assumption about the pharmacokinetics of pulsed drug treatments, assuming exponential decay of the effective dose, $u(t)$, of the drug, as has been shown by others in greater detail (21,22).

$$u(t) = k_1 C_{drug} e^{-k_2 t}, \quad \text{(Eq. 2)}$$

where C_{drug} is the concentration of doxorubicin in nM, k_1 is a scaling factor used to non-dimensionalize the effective dose, and k_2 is an estimated rate of decay of the effect of doxorubicin pulse-treatment on breast cancer cells. The effective dose decays over a time scale consistent with experimental measurements of doxorubicin fluorescence dynamics *in vitro* (21,22) .

Previous work has demonstrated the theoretical implications of treatment-induced resistance (α in our model*)* on determining optimal treatment regimens (15). Simulations from our model (Eq.1) also revealed the importance of the degree of drug-induced resistance (α) in treatment optimization. We simulated a resistance-preserving therapy (i.e., $\alpha = 0$), and found that a constant dosing regimen optimizes tumor control (black line Fig 1B), leading to a lower maximum tumor cell number than the pulsed treatment (blue line Fig 1B), whereas for a resistance-inducing therapy (i.e., $\alpha > 0$) a pulsed treatment regimen (blue line Fig 1C) reduced tumor cell number over time.

We employ an experimental *in vitro* model system of MDA-MB-231 triple negative breast cancer cells exposed to the chemotherapeutic doxorubicin. By applying a range of 24-hour pulse treatments (Fig 1D), we can estimate the effective dose ($u(t)$) for each treatment (Fig 1E) and measure the total cell number over time using time-lapsed microscopy on the 6 replicate wells for each dose (Fig 1F)(see Methods: Longitudinal Treatment Response Monitoring). The mean and 95% confidence intervals of cell number in time are shown in Fig 1F. The measurements of total cell number in time acquired experimentally can be compared directly to the model predicted cell number in time. However, while we may not feasibly be able to measure the resistant and sensitive cell number longitudinally, we will demonstrate how we can estimate the "phenotypic composition"; the proportion of cells in the sensitive state $\phi_S(t)$ (or simply $\phi(t)$), throughout treatment response using lineage-traced transcriptomic data. Model outputs of $N(t)$ and

$\phi(t)$ can be used directly to compare to measurements of cell number in time and phenotypic composition in time following a drug treatment (Supp Fig S1). A full description of the parameters in the modeling workflow are described in Table 1, and their values and confidence intervals are listed in Supp Table S1.

Parameter	Description	Units	Determination
$N(t)$	Total cell number over time, measured directly and predicted by the model	Number of cells	Directly measured
$\phi(t)$	Phenotypic composition: the fraction of sensitive cells over time, estimated from scRNA-seq data and predicted by the model	Cell fraction	Estimated from classifier output from scRNA-seq data
r_S, r_R	Growth rate of sensitive and resistant cell subpopulations	hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
α	Drug-induced rate of transition from sensitive to resistant state	nM^{-1} x hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
d_S, d_R	Death rate of sensitive and resistant cell populations due to drug, $d_R < d_S$	nM^{-1} x hour $^{-1}$	Fit from $N(t)$ & $\phi(t)$ data
ϕ_0	Initial proportion of sensitive cells	number of cells	Fit from $N(t)$ & $\phi(t)$ data
K_N	Carrying capacity for the longitudinal treatment experiment performed in a 96 well plate to measure $N(t)$	number of cells	Fit from $N(t)$ untreated control
K_ϕ	Carrying capacity of the scRNAseq experiment performed in a 10 cm dish to measure $\phi(t)$	number of cells	Fixed
k_1	Scaling factor to non-dimensionalize concentration in nM of doxorubicin	nM^{-1}	Fixed
k_2	Estimated rate of decay of effect of doxorubicin after pulse-treatment	hour $^{-1}$	Fixed

Table 1. Description of model parameters to describe resistance dynamics. Descriptions of the parameters either from measured data (Data), fit of the model to the $N(t)$ (Fit from $N(t)$) or $\phi(t)$ (Fit from $\phi(t)$), the model assumptions (Fixed), or predicted from the parameter estimation from the fitted model (Predicted). We fit for six free parameters in the calibration scheme, as listed by the first four rows of the table.

Integrated Modeling Workflow for Estimating the Phenotypic Composition from scRNA-seq Data

The combined experimental-computational workflow (Fig 2) starts by tagging individual cells with unique barcodes that are integrated into the genome and expressed as sgRNA's; this COLBERT cell barcoding platform has been described previously (23). The barcode-labeled cell population is expanded to generate the naïve population for these studies (305 unique barcodes represents 305 clonal subpopulations). Cells are then treated with doxorubicin (LD95, 550 nM) for 48 hours and allowed to recover; scRNA-seq is performed prior to treatment and from two parallel replicates after the population had regrown following the pulse treatment, corresponding to seven and ten week post-treatment timepoints.

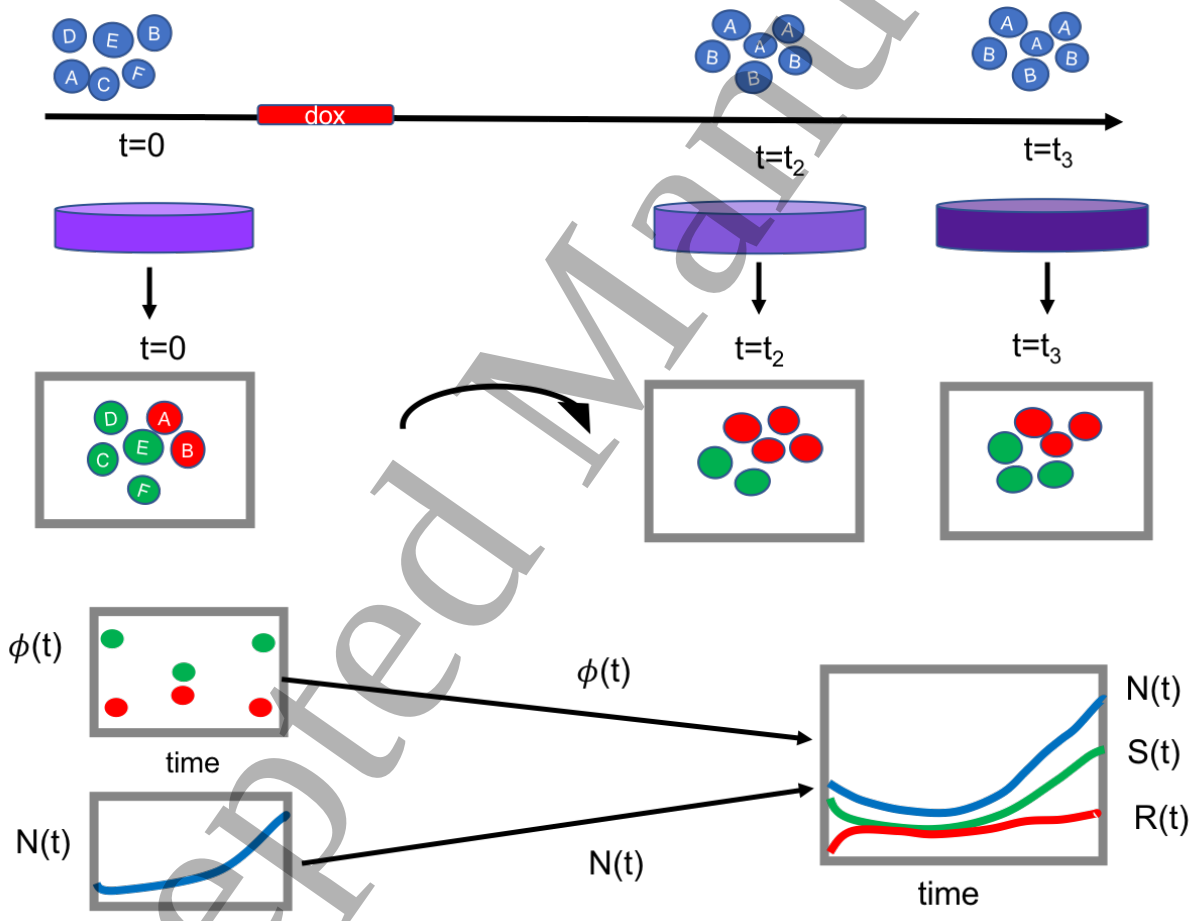


Fig 2. Schematic of the workflow for identifying model parameters from data. At *t*=0 wks prior to treatment, individual cells are tagged with a unique, heritable, expressed COLBERT barcode. Cells are treated with a pulse treatment of doxorubicin and allowed to recover from treatment, at which time the barcode abundance is quantified. Lineages whose barcode abundance increased from pre- to post-treatment are assumed to have been in a phenotypic state at *t*=0 wks that conferred them more resistant to drug than cells whose barcodes significantly decreased in abundance after treatment. Samples of the population were taken before and from parallel replicates sampled at two different time points after treatment for scRNA-seq. The transcriptomes in the pre-treatment samples of the cells are assigned resistant or sensitive if they fall on the extreme tails of this distribution and are used as the

labeled training set. Using the gene-cell matrix and labeled class identities of sensitive or resistant from the pre-treatment time point only, a classifier is built using Linear SVM to distinguish between sensitive and resistant cells. The classifier is applied to the remainder of transcriptomes of the cells, resulting in a prediction for each cell as either sensitive or resistant. These machine learning outputs are made actionable as state variables by using them to quantify the proportion of sensitive cells ($\phi(t)$) at the three time points. This is combined with separate experiments of longitudinal treatment response dynamics ($N(t)$) of the bulk population of the same cell type, and both serve as measured data to be compared to model predicted outputs for parameter estimation.

The transcribed barcode sequence indicating lineage identity is measured alongside other transcripts in scRNA-seq in each cell. Cells from the pre-treatment time point whose lineage abundance increased by any amount after treatment were designated as “resistant”, and cells whose lineage abundance decreased by more than 5% were designated as “sensitive” (Fig 3A). These thresholds were chosen because they represent the tail ends of the distribution of cells with changes in lineage abundance, and therefore were assumed to be most likely to be in a phenotypically drug-sensitive or drug-resistant state at pre-treatment. This training set consisting of 47 resistant and 768 sensitive cells and their expression levels of 20,645 genes (Fig 2, Fig 3A) was used to build a classifier capable of predicting whether a newly observed cell of unknown identity (Fig 3B) is more likely to be in a resistant or sensitive state based on its gene expression levels alone. See Methods: Machine Learning of Cell Phenotypes for full description of building of the classifier. The type of classifier was chosen by comparing the accuracy of classification of labeled cells, using 5-fold cross validation on the pre-treatment training set, for two types of classifiers: principal component analysis (PCA) with k-nearest neighbors (KNN) and linear support vector machine (Linear SVM) (Supp Fig S2). These two methods were chosen because both methods return not only estimates of a cells most likely class, but also the gene weightings used to make this estimate, making the results interpretable in the context of differential gene expression analysis. The Linear SVM classifier model was shown to be most accurate (Supp Fig S2D) and was used going forward to classify all of the remaining cells based on their gene expression levels alone, and UMAPs were used to visualize the high-dimensional cell transcriptomes (Fig 3C). The PCA+KNN classifier generated similar results in terms of estimates of $\phi(t)$(Supp Fig S3). One of the advantages of using Linear SVM as a classifier is that we can examine the highest weighted genes in the classifier

to reveal new mechanistic insights into the phenotypes relevant to functional treatment resistance. Although this is not the focus of this manuscript, our results comparing expression levels for specific genes associated with resistance can be found in Supp Fig S4. A goal of future work is to further investigate mechanistic underpinnings behind how these genes might drive resistance and find targets for these genes to identify novel therapeutic combination strategies.

For each of the data sets from the three time points, the estimates of the class of each cell were used to quantify the proportion of cells labeled as sensitive ($\phi(t)$) (Fig 2, Fig 3G). This phenotypic composition estimate at three time points can then be combined with bulk time-course data from drug treatments at different concentrations, compared to corresponding model outputs, and serve to calibrate the mathematical model of drug-induced resistance (Fig 2, Supp Fig S1).

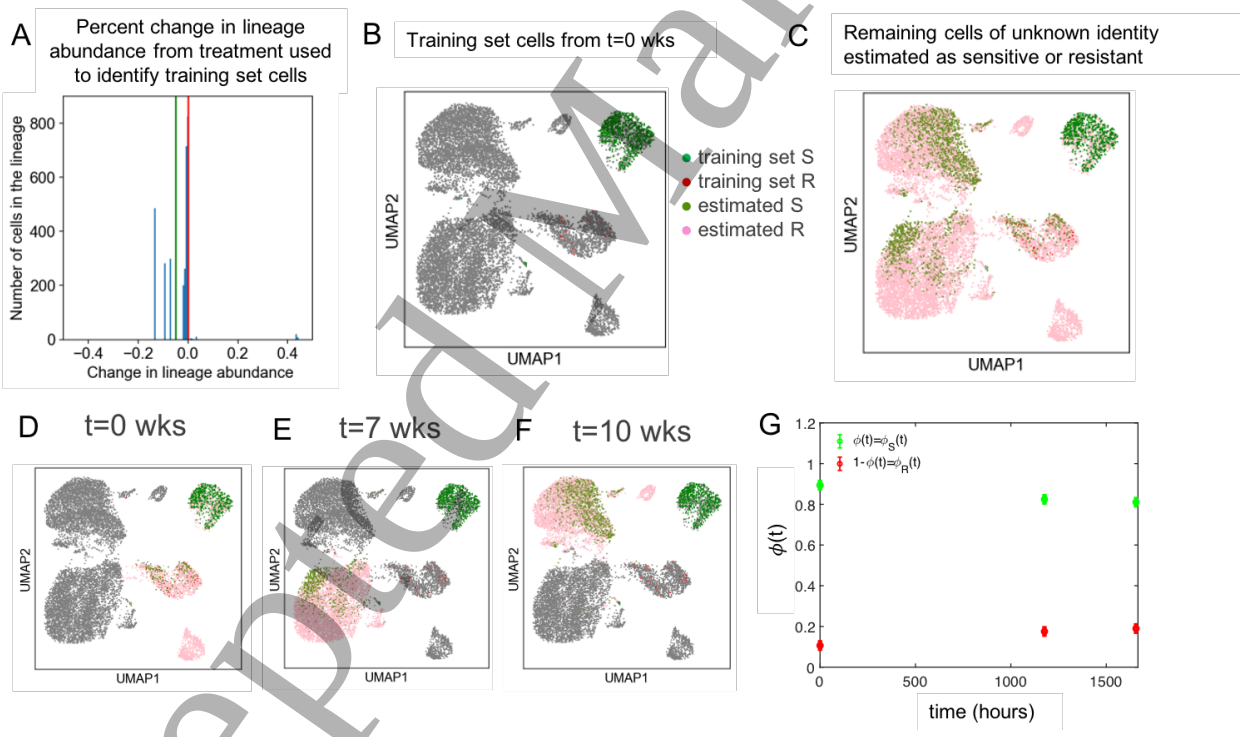


Fig 3. Functional Read-out of Changes in Lineage Abundance Allows Mapping of Phenotypes to Transcriptome A. Distribution of changes in lineage abundance from pre- to post-treatment indicates separation of lineages whose cells survive and proliferate and those that are more likely to have been killed by the drug treatment. B. Lineage-abundance guided training set of sensitive (S, green) and resistant (R, red) cells visualized using UMAP projections. C. Cells of unknown drug sensitivity identity are estimated as sensitive (pink) or resistant (olive green) based on their transcriptome using a Linear SVM classifier. D. Cells from pre-treatment (t=0 wks) labeled as and estimated as sensitive and resistant E. Cells from t=7 wks post-treatment estimated as sensitive

and resistant. F. Cells from t=7 wks post-treatment estimated as sensitive and resistant. G. Proportion of cells that are classified as sensitive (green) and resistant (red) at each time point.

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

To utilize all possible pieces of information available about the treatment response of this experimental system, we sought to develop an integrated model calibration scheme that is capable of integrating information from multimodal data sources. Here, we expect there to be a trade-off between goodness-of-fit in each of the two data sources: 1) from longitudinal population data, *N(t)*, sampled at a high temporal resolution and for a number of doses, and 2) machine learning outputs that estimate the phenotypic composition $\phi(t)$ at three distinct time points before and after treatment. For the following dual-objective function, we weight by the number of data points in order to assign equal weight to the cell number and phenotypic composition measurement sources. We use a weighted, non-linear, least squares as the simplest possible calibration method:

$$J(\theta) = \frac{1}{n_\phi}\sum_{j=1}^{n_\phi}\frac{\left(\widehat{\phi_j}-\phi_j(\theta,u)\right)^2}{\sigma_{\phi_j}^2} + \frac{1}{n_N}\sum_{k=1}^{n_{doses}}\sum_{i=1}^{n_N}\frac{\left(\widehat{N_{i,k}}-N_i(\theta,u_{i,k})\right)^2}{\sigma_{N_{i,k}}^2} \quad \text{(Eq. 3)}$$

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

Using the effective dose regimens (Fig. 1E) and treatment response data (Fig 1F) we calibrate the model using three of the selected doses- the untreated control (0 nM), the 50 nM dose, and the 100 nM dose. The remaining treatments will be used for validation. The results of the integrated parameter estimation from the $N(t)$ data from

these three doses and the $\phi(t)$ data from the three scRNA-seq time points, are shown in Fig 4. We compare the model fit to the experimental $N(t)$ data (Fig 4A) and the phenotypic composition estimates (Fig 4B). The overall goodness of fit between the mean cell number data and the model estimated cell number over time is shown in Fig 4C, with a concordance correlation coefficient (CCC) of 0.94. In order to compare methods, we also performed the calibration with only the longitudinal ($N(t)$) data to obtain a parameter set estimated without the additional information provided by the phenotypic composition (Supp Fig S5). We note that the goodness of fit in $N(t)$ for the model calibrated only to $N(t)$ is higher (Supp Fig S5C, CCC=0.97) than the integrated fit (Fig 4C, CCC=0.94). The trade-off in goodness of fit in $N(t)$ for the integrated calibration allows for an improvement in fit to phenotypic composition (Fig 4B, versus Supp. Fig. S5B).

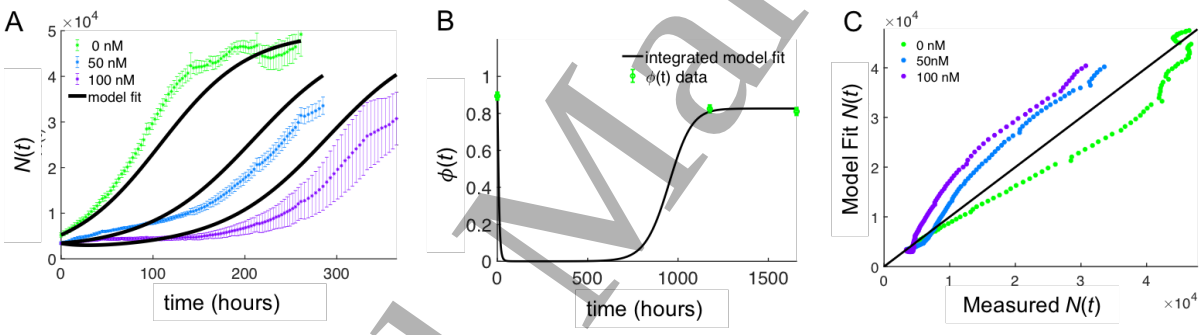


Fig 4. Integrated model calibration incorporating both measurement sources. A. Calibration results for longitudinal $N(t)$ data from the four doses (0, 50, and 100 nM) used for calibration B. Calibration results for phenotypic composition ($\phi(t)$) C. Measured cell number $N(t)$ versus model calibrated cell number, yielding a concordance in $N(t)$ of CCC = 0.94.

In the model development process, we tested that each of the parameters was sensitive to the relevant model outputs, in this case 1) the time to reach two times the initial cell number and 2) the phenotypic composition at this time, for a range of doxorubicin doses. Results from the global sensitivity analysis (See Methods: Sensitivity Analysis of Model Parameters) revealed that all parameters are globally sensitive (i.e. contribute to least 5% of the overall value) in at least one of the model outputs for at least one of the drug doses (Supp Fig S6), except for the carrying capacities (K_N *and* K_ϕ) of the two experimental systems. We used this analysis to inform our decision to set the carrying capacities from separate experiments (Supp Fig S7) and literature (24) and to fit all six remaining unknown parameters. In order to ensure the identifiability of the remaining

model parameters (Table 1), we demonstrated the structural identifiability of the system (See Methods: Structural Identifiability of Model Parameters) under the assumption of perfect data. To test for practical identifiability and obtain confidence intervals on our parameter estimates, we used bootstrapping with replacement to generate synthetic data sets and repeatedly fit for model parameters (25,26) (Supp Fig S8, Supp Fig S9, Table S1).

Model Validation Using Functional Isolation of "Sensitive" and "Resistant" Cells Predicted from Classifier

Because we rely on the machine learning classifier of cell phenotypes from transcriptomic data, we sought to validate our classifier model experimentally to ensure that cells labeled as "resistant" and "sensitive" were exhibiting these expected phenotypes. Our mathematical model assumes that sensitive cells proliferate more rapidly than resistant cells (i.e. exhibit a higher growth rate) and that resistant cells are capable of higher survival rates in response to doxorubicin treatment. To test these attributes functionally, we used the COLBERT barcoding system (23) to identify one of each lineage from the pre-treatment sample that was labeled as sensitive or resistant based on their changes in lineage abundance. The COLBERT recall system enables Fluorescence Activated Cell Sorting (FACS) isolation of specific lineages from the replicate pre-treatment population by transfection with a gene circuit to activate lineage-specific reporter expression (23) (Fig 5A). Once isolated, cells were sorted into single cell clones for functional analysis of growth dynamics and drug sensitivity. Cells from the isolated sensitive lineage grow more quickly than the isolated resistant lineage (Fig 5B), with overall growth rates of g_S=0.011 and g_R= 0.005 per hour respectively (Supp Fig S10). Drug sensitivity was assessed by dosing cells at 400 nM and 2.5 μM for 48 hours and immediately quantifying cell viability via a live-dead assay. The resistant lineage had higher percent viability at both doxorubicin concentrations, with a statistically significant difference in viability at the higher dose (Fig 5C).

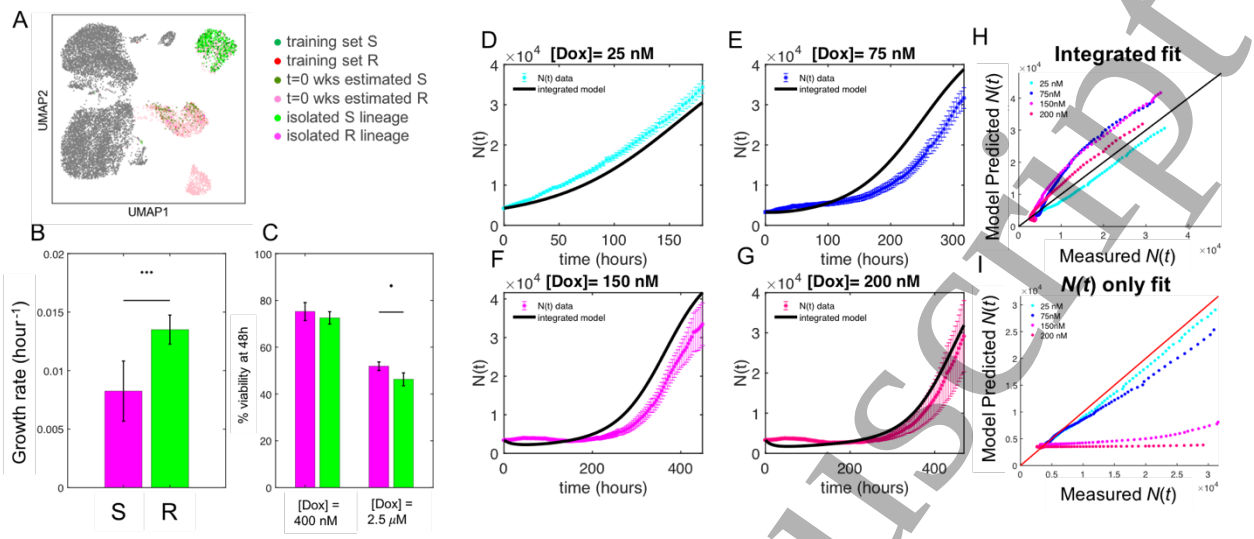


Fig 5. Combined Model Validation *via* Lineage Isolation and Prediction of Treatment Response. A. UMAP visualization of classified sensitive and resistant cells at the pre-treatment time point, with cells from an isolated sensitive lineage (AA170) in bright green, and an isolated resistant lineage (AA161) in hot pink B. Growth rates of the 12 replicate wells of each isolated lineage reveal that the resistant lineage grows significantly more slowly than the sensitive lineage (p = 2.7e-6), as is predicted from the model parameters where $r_S > r_R$. C. Functional testing of the drug sensitivity of each lineage indicates that the cells from resistant lineage (AA161, pink) have a higher resistance, measured by cell viability at 48 hours, at both 400 nM and 2.5 μM doses of doxorubicin, with p-values of p = 0.1942 and p = 0.0023, respectively. D. Prediction of treatment response at 25 nM, E. 75 nM, F. 150 nM, and G. 200nM from the integrated calibration. The mean measured cell number in time and 95% confidence interval from six replicate wells are shown for each treatment response. H. Scatterplot of model predicted $N(t)$ from the integrated calibration versus experimental $N(t)$ data for all four new treatment conditions with an overall CCC = 0.92. I. Scatterplot of model predicted $N(t)$ from longitudinal data calibration alone versus experimental $N(t)$ data for all four new treatment conditions with an overall CCC = 0.64.

Multimodal Data Sources can be Leveraged to Predict Response Dynamics to New Drug Concentration

A key advantage of leveraging multimodal data sources for parameter estimation is that we can use them to make predictions about the response dynamics to new treatment regimens. We validate the model predictions, obtained from running the model forward with the integrated calibration parameter set with input effective doses described in Fig 1E for the four remaining pulse treatment of doxorubicin that were not used to calibrate the model. The model predictions compared to the experimental measurements are shown for doses of 25 nM (Fig 5D), 75 nM (Fig 5E), 150 nM (Fig 5F), 200 nM (Fig 5G). We evaluated the overall accuracy in all the model predictions over all four not-previously-observed doses and see that we are able to predict the treatment response with reasonable accuracy (Fig 5H) with an overall CCC of 0.92 for each model predicted

and measured cell number ($N(t)$) in time. When we compare this to the prediction accuracy of the calibration performed without the phenotypic composition data, we get an overall predictive accuracy of CCC=0.64 (Fig 5I, Supp Fig S11), indicating the improvement in predictive capabilities with insight of the phenotypic dynamics. These results demonstrate the improved predictive capacity of an integrated modeling framework, in which molecular data from scRNA-seq during treatment response improves our ability to predict response to new treatments.

Discussion

Recent technological advances have enabled unprecedented, high-throughput single-cell molecular level insight of intratumor heterogeneity (27,28). The ability to precisely quantify intratumor heterogeneity (1), and illuminate key subpopulations involved in response to treatment (9), has the potential to improve both prognostic and therapeutics for cancer treatment. These genomic and transcriptomic data sets can direct the choice of specific cancer drugs and illuminate novel resistance pathways, as well as provide a prognostic marker for patients who receive it. Simultaneously, the role of mathematical modeling in oncology has been widely recognized (29) and utilized to improve both our understanding of the dynamic mechanisms of drug response (10,30,31) as well as to develop approaches to guide the design of adaptive patient-specific treatment plans (12,17,18,32,33). However, connecting the wealth of “omics” data at the molecular level with temporal dynamics used to calibrate mathematical models for adaptive therapies remains a major challenge in the field.

Recognizing the critical roles of heterogeneity in cancer dynamics, mathematical models of tumor progression often include distinct subpopulations, such as cancer stem cells (12,34,35), or drug resistant and sensitive subpopulations (15,16,19,36). However, despite these models being calibrated to observed experimental or clinical data, the underlying phenotypic composition that these model calibrations suggest cannot easily be validated, since the degree of resistance or stemness of a cancer cell population in time is not easily measured longitudinally *via* a single biomarker. A few studies utilizing multimodal imaging modalities have harnessed the ability to quantify different aspects of tumor composition—such as vasculature, necrosis, and cellularity, to develop an

integrated model calibration of multiple tumor system components (37,38). However, this integrated, multimodal approach has not explicitly included data on the composition of heterogeneous subpopulations taken from separate "omics" datasets for direct calibration of a dynamical systems model.

Here, we introduce an experimental-computational framework for utilizing transcriptomic and bulk time course data to parametrize a dynamic model of treatment response in cancer. We demonstrate the applicability of this framework when applied to clonally-resolved scRNA-seq data combined with bulk time course treatment response data from a cancer cell line and assess the ability of the model to predict treatment response dynamics. To this end, we developed a machine learning classifier built upon clonal abundance quantification which estimates the class identity of an individual cell based on its transcriptome. The output of the classifier enabled us to assign values related to the state variables in the dynamic model: the proportion of cells in the sensitive or resistant phenotypic state at each time point. We combined these estimates of phenotypic composition with population-level treatment response data to calibrate a mathematical model of drug-resistance dynamics. We validated our machine learning classifier by isolating cells from lineages labeled as sensitive or resistant and testing them functionally in growth and treatment response assays. We showed that the presence of multiple measurement sources of data allows us to more accurately predict the effect of new drug treatments on the cell population.

The power of mathematical models in oncology, especially those calibrated to real data, is that we can use them to learn about the underlying system behavior to inform decision-making (39,40). High-throughput single cell transcriptomics or other types of high throughput snapshot data can give an abundance of information about the heterogeneity and potential mechanisms of resistance of cell populations (9,41). However, the ability to use this information beyond hypothesis generation (10), but to actually inform model calibrations, is still lacking. In this work, we leverage a high-throughput "omics" data set, taken at just a few snapshots of time, to estimate the phenotypic composition and demonstrate one way to include this data alongside longitudinal data for model calibration. We by no means claim that this is the only way to integrated multimodal data sources in oncology, and present this work as an example of

one such plausible way, in hopes that it will prompt further investigation into how to incorporate experimental and clinical data from a variety of measurement sources and scales into mathematical modeling frameworks, ideally incorporating multiple "omics" data sets in future expanding work.

The functional characterization of single cells via changes in lineage abundance post-treatment enabled us to identify cells that group together based on response to treatment. While unsupervised clustering of cells by their transcriptomes can enable identification of novel cell states, these cell states are not necessarily relevant to drug-tolerance. Once can see this quite simply in scRNA-seq pipelines as failure to remove cell cycle genes from the analysis reveals that cells will often cluster by cell cycle state (42), leading them to be commonly regressed out if they are not relevant to the biological question of interest. However, we cannot regress out other unknown phenotypic subpopulations, and thus these are what can emerge from unsupervised clustering algorithms. While these can provide novel insight about population structure, they may not be what is relevant to driving changes in treatment response behavior. Thus, the ability to read-out lineage identities represents a novel functional component that enables us to zoom in at the right phenotypic state-space relevant to our question- what cells are more capable of surviving treatment and which are more sensitive to treatment, and what is driving these changes? Because we used a classifier that can output gene loadings most relevant to the separation of sensitive and resistant cells, we can look at the differences in the gene expression patterns and propose potential novel interactions and biomarkers. In this manuscript, we only demonstrate the feasibility of this endeavor; further mechanistic insight into the role of key genes and their related pathways in drug response will be a subject of future work.

We acknowledge that the modeling framework describe here has a number of limitations. In the dynamic model calibrated to the two data types, we make a number of assumptions in order for the model parameters to remain identifiable. First, we assume that the sensitive and resistant cells do not affect each other's growth rates directly, with intrinsic growth rates (r_S and r_R) independent of population composition. This does not take into account recent work in non-small cell lung cancer that has demonstrated that resistant cell growth rate was suppressed in the presence of sensitive cells, implying a

persister-like phenotype of resistant cells (43). Additionally, we do not explicitly model the reverse phenotypic transition from the resistant to the sensitive state as this would introduce an additional parameter and render parameter estimation more difficult. However, we note that a relaxation towards increased sensitivity can occur in the model as its written due to the higher growth rate of sensitive cells in the absence of treatment. In the classifier model, we acknowledge the limitation of making continuous, high-dimensional, gene expression vectors into a single binary classification scheme of sensitive and resistant. In reality, cells likely exist on a drug sensitivity spectrum, with a distribution of cells in different regions. This makes the definition of "sensitive" and "resistant" cells, which we defined via a threshold change in lineage abundance, somewhat arbitrary. We intended to overcome these limitations by validating the predictions of the dynamic model to new drug treatments and by functional characterization via isolation and functional testing of the cell phenotypes. Because of the destructive nature of scRNAseq assays, we weren't able to sample the cell population while it was depleted significantly due to drug, rendering the predicted drop in proportion of sensitive cells to lack validation. In future work, we intend to design a study with a lower dose and higher initial cell number, so that the population can be sampled at this critical intermediate time point, and used for either calibration or validation.

While scRNA-seq has limitations in the clinical setting due to its high cost, in experimental settings barcode labeling fits flexibly into existing scRNA-seq workflows and can add a critical functional component to the phenotypic read-out, as we show in this work. In the clinical setting, other types of approaches to learn more about cancer cell composition are being employed in the era of precision medicine. From radiomics to genomics, it is becoming increasingly common for patients to have access to high-throughput measurements, or at least some insight into their mutational burden at certain time points. This information may be integrated into the clinical or tumor board's decision-making process (44).

We suggest that the general approach presented here could be applied to integrate available types of data in different experimental or clinical settings, potentially with the model used here or with different models aimed at addressing a relevant question. While transcriptomic and longitudinal data have been used together in a number

of studies, this is the first work to our knowledge that allows for explicit parameter estimation using these two measurement sources of varying time resolutions. This work represents one example of the opportunities for synergy of machine learning with dynamic modeling to integrate multimodal datasets and open up new approaches to describe, predict, and ultimately optimize treatment response in cancer.

Methods

Key Resources Table

See Attached Template

Contact for Reagent and Resource Sharing

amy.brock@utexas.edu

Experimental Model and Subject Details

Cell culture

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

A subline of the MDA-MB-231 breast cancer cell line was engineered to constitutively express EGFP (enhanced green fluorescent protein) with a nuclear localization signal (NLS). Genomic integration of the EGFP expression cassette was accomplished through the Sleeping Beauty transposon system (45). The EGFP-NLS sequence was obtained as a gBlock from IDT and cloned into the optimized sleeping beauty transfer vector containing the EGFP-NLS expression cassette and the pCMV(CAT)T7-SB100 plasmid containing the Sleeping Beauty transposase was co-transfected into a MDA-MB-231 cell population using Lipofectamine 2000. mCMV(CAT)T7-SB100 was a gift from Zsuzsanna Izsvak (Addgene plasmid #34879) (46). GFP+ cells were collected by fluorescence activated cell sorting. MDA-MB-231 cells are maintained in DMEM (Gibco), 10% fetal bovine serum (Gibco) and 200 μg/mL G418 (Caisson Labs). Cells were seeded into the center 60 wells of a 96 well plate (Trueline) at about 2000 cells per well. During the monitoring and treatment, plates were kept in the Incucyte Zoom, a combined incubator and time-lapsed microscope. Cells were fed fresh media every 2-3 days for up to 5 weeks. HEK293T cells were cultured in DMEM with GlutaMAX supplemented with 10% FBS, 4.5 g/L D-glucose, 110 mg/L sodium pyruvate, streptomycin (100ug/mL) and penicillin (100 units/mL).

Longitudinal treatment response data

The EGFP-labeled subline of MDA-MB-231 breast cancer cells were used for longitudinal treatment response. Cells were passaged into the center 60 wells of 96 well plates at a density of about 2000 cells per well. Two days later, cells were treated with a 24-hour

pulse-treatment of doxorubicin at concentrations ranging from 0-200 nM (0 nM, 25 nM, 50 nM, 75 nM, 100 nM, 150 nM, 200 nM), with 6 replicate wells of each dose. Dosed media was applied to cells and treatment response was monitored using the Incucyte. After 24 hours, the dosed media was replaced with normal media and monitoring continued. Cells were fed fresh media every 2-3 days for the duration of the monitoring period (up to 2.5 weeks).

Integration, expression, and capture of COLBERT barcodes

Lentiviral Assembly

Lentiviral assembly was performed using Lipofectamine 2000 (ThermoFisher). Prior to transfection 0.25x106 HEK293T cells were plated in each well of a 6 well. 48 hours following plating, each well was transfected with 1.5 ug PsPax2 (Addgene # 12260), 0.4ug VSV-G (Addgene # 14888), 3 ug CROPSseq-BFP-WPRE-TS-hU6-N20 and 9 uL of Lipofectamine 2000 in 150 ul of Opti-mem (Thermo Fisher). Media was replaced with fresh growth medium after 18 hours of transfection. Media containing viral particles was collected at 48 and 72 hours, centrifuged for 5 minutes and passed through a 45 uM (PES) low protein Binding filter. Virus was concentrated for 1 hour at 4000g in a Vivaspin (Sartorius) filtration column then aliquoted and stored at -80 for later use.

Barcode Labeling

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

Drug Treatment of Barcoded Cells for scRNAseq and Recovery

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

scRNA-seq

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

Plasmid Assembly for Isolation of Lineages

After selecting the lineages of interest for isolation, an array of barcodes was assembled as described in (23). Briefly, oligonucleotide pairs for the barcode of interest were ordered with specific overlapping sequences to both direct assembly of barcode array and integration into the plasmid for isolation. The barcode arrays were ligated, and gel purified

to proceed with only a fully assembled array in cloning. The fully assembled barcode array was cloned into the BbsI site with standard restriction digest cloning. This double stranded barcode array was inserted into a plasmid backbone upstream of a minimal core promotor (miniCMV) and sfGFP to generate the Recall plasmid. This was repeated with individual barcodes of interest.

Recall of Isolated Sensitive and Resistant Clones by COLBERT

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

Alignment to Reference Genome

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

Filtering and Normalization

The filtered matrices produced by cellranger were loaded into scanpy (v1.4.4)(47). Cells were annotated by sample and lineage membership. Only cells meeting the following requirements were retained for further analysis: (a) a minimum of 10000 and maximum of 80000 transcript counts, (b) a maximum of 20% of counts attributed to mitochondrial genes, and (c) a minimum of 3000 genes detected. Genes detected in fewer than 20 cells were removed. Normalization was conducted based on the recommendations from multiple studies that compared several normalization techniques against each other(42,48,49). In brief, three steps were performed: (a) preliminary clustering of cells by constructing a nearest network graph and using scanpy's implementation of Leiden community detection(50), (b) calculating size factors using the R package scran(51), and (c) dividing counts by the respective size factor assigned to each cell. Normalized counts were then transformed by adding a pseudocount of 1 and taking the natural log.

Regressing Out Cell Cycle Expression Signatures

Using a list of genes known to be associated with different cell cycle phases (52), cells were assigned S-phase and G2M-phase scores. The difference between the G2M and S phase scores were regressed out using scanpy's *regress_out* function.

Quantification and Statistical Analysis

Machine Learning of Cell Phenotypes

The machine learning classifier of sensitive and resistant cell phenotypes was built from the normalized, pre-processed single cell gene expression matrix with lineage identities. For the cells in the pre-treatment sample, the lineage abundance at the pre-treatment time point (proportion of cells in each lineage) was calculated and compared to the lineage abundance from the combined post-treatment time points (7 and 10 week samples). If the lineage was not observed in the post-treatment time points, the lineage abundance post-treatment was assigned a zero. The change in lineage abundance (% post -% pre) was found for each lineage in the pre-treatment time point (See Supp. Fig. S3A). Based on this change in lineage abundance distribution, only cells on the pronounced tails of the distribution were used for classification, since these extremes were most likely to exhibit characteristics that made them significantly more or less likely to survive drug treatment. Cells from the pre-treatment timepoint whose lineage abundance increased post-treatment were labeled as resistant. Cells whose lineage abundance decreased by more than 5% were labeled as sensitive in the pre-treatment time point. These thresholds for calling a cell from the pre-treatment time point sensitive or resistant were determined based on the assumption that these cells with pronounced changes in lineage abundance represented more pronounced differences in initial drug-sensitivity phenotypes. Because drug sensitivity is not binary, but is more likely to exist on a spectrum, this threshold can in theory be shifted to encompass a wider range of phenotypes considered “sensitive” and resistant”.

The current threshold resulted in 815 cells and their corresponding 20,645 normalized gene expression levels being used to form the training set gene-cell matrix containing a cell’s gene expression vector and corresponding identity. This gene-cell matrix was then used to build a classifier capable of predicting the identity of new cells based on an individual gene expression vector. A Linear Support Vector Machine and a principal component with k-nearest neighbors were both tested as possible classifiers because of the interpretability of the output of the classifiers in terms of gene loadings. Cross validation was performed on models built using both types of classifiers, and the average accuracy and area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve were evaluated for each training-test set combination (Supp Fig S2C &D). The Linear SVM method was found to be more accurate. The ROC curves for the full training set were used to determine an optimal probability score threshold for calling a cell sensitive or resistant (Supp Fig S2 A &B). While many appeared to be reasonable, we chose a threshold value of P(resistant)= 0.9 as our cut-off for calling a cell resistant in the Linear SVM model, as this generated a realistic proportion of cells in each class at the pre-treatment time point, as we don’t expect a large proportion of the naïve cancer cell line to be resistant.

The Linear SVM classifier was built using python’s *sklearn* package *svm* function, with the gene-cell matrix as the input, and trained on the labels from the pre-treatment training set, as were all downstream analyses of the classifier’s outputs. The principal component classifier + k-nearest neighbors (PCA+KNN) was built using python’s *sklearn* package *PCA* function with the same inputs. However, for PCA+KNN, both the number of principal components used in the classifier, and the number of nearest neighbors used

to predict a cell's class based on the class of the k cells its closest to, needed to be optimized. This was done using the 5-fold CV training and testing sets and coordinate optimization was then used to iteratively find the optimal number of both nearest neighbors (*k*) and number of principal components (*n*) for correctly identifying the class of each cell. Coordinate optimization works by essentially iteratively optimizing the two variables of interest, here *k* and *n*, until they no longer change values. In this case, we first set the number of principal components to a single value and iterated through a range of nearest neighbors to find the number which gave the highest mean AUC (area under the curve) over all 5 folds of cross validation (Supp Fig 12C). Once the optimal number of neighbors was found for that number of principal components, the number of neighbors was set to that value and the optimal number of principal components was varied over a range of values, and again the highest mean AUC over all 5 folds of cross validation was found (Supp. Fig. S12D). Then we set the number of neighbors to this value and repeated the search for the optimal number of principal components. This process was repeated until the optimal number of neighbors and number of principal components no longer changed with each iteration. The percent of variance explained by each PC was recorded (Supp Fig S12A) and the cumulative variance (Supp. Fig. S12B). The entire classification and output results were performed for PCA + KNN and results are in the supplement, visualized in the space of PC1 and PC2 (Supp. Fig. S3).

Model of Drug Resistance Dynamics

The mathematical model of drug-induced resistance, in which treatment exposure directly induced phenotypic transitions into the resistant cell state, was introduced in (15). Their original model described sensitive cells (*S*) and resistant cells (*R*) independently growing according to logistic growth and independently dying due to drug treatment ($u(t)$) *via* a log-kill hypothesis. The model includes an explicit role for the transition of sensitive cells into resistant cells via a rate of drug-induced resistance (α) which is modeled as a linear function of treatment $u(t)$. Additionally, their full model included additional terms of spontaneous, treatment-independent resistance (ε) proportional to the number of sensitive cells present, as well as a resensitization term (γ) describing treatment-independent transition from the resistant to the sensitive cell state.

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

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

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

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

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

Where r_S and r_R are the sensitive and resistant subpopulation growth rates and d_S and d_R are the sensitive and resistant subpopulation death rates, assumed to be linearly proportional to the effective dose ($u(t)$). We assume that the sensitive cells grow faster

than the resistant cells so that $r_s > r_r$, as is consistent with the mechanism of action of cytotoxic therapies targeting rapidly proliferating cells (15,53). We assume $d_S > d_R$ as sensitive cells should die more quickly in response to drug than resistant cells, by definition. We modeled the effect of the pulse-treatments as single pulses of $u(t)$ whose maximum is given by the concentration of doxorubicin and whose effectiveness in time decays exponentially.

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

The constants k_1 and k_2 were chosen so that $u(t)$ is scaled between 0 and 5 and so that the effective dose decays over a time scale consistent with experimental observations of doxorubicin fluorescent dynamics *in vitro* (21,22). Numerical simulations of the forward model for a given treatment regimen were implemented in MATLAB using the backward Euler method.

Sensitivity Analysis of Model Parameters

As part of the model development process, we performed a sensitivity analysis to assess the effect of individual model parameters on the model output. Although there are a number of choices to use for model outputs, we chose to capture the broad drug response of the population using the time to reach two times the initial cell number, which we call t_{crit}, and the phenotypic composition $\phi(t=t_{crit})$ at that time, as we expect these are two outputs we would feasibly observe in an experimental setting, as the time to population rebound and the phenotype observable via scRNAseq or some other phenotypic characterization. We first performed a global sensitivity analysis on the set of parameter bounds that were well outside the parameter ranges of the calibrated parameters and their associated errors. The results of the sensitivity analysis will reveal the most important parameters of the system, causing the greatest variation in outputs. This exercise should identify any model parameters that the model is insensitive to, and therefore may present opportunities to simplify the model to capture the same dynamics while reducing uncertainty by eliminating the number of free parameters to be fit. A Sobol's global sensitivity method is applied, which is a method that utilizes the analysis of variance (ANOVA) decomposition to define its sensitivity indices (54). As a global method, random sampling is performed twice over the parameter space of the eight parameters (six free, two carrying capacities), with the number of parameters by N simulations matrices denoted by X and Z. The bounds of the global sensitivity analysis were chosen to be well outside of the 95% confidence intervals around each best fitting parameter from the profile likelihood analysis. The total effects are then calculated using the following:

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

Where σ^2 is the variance of the outputs from the first set of N random samples computed from evaluating over all x_j in X, and the function evaluations of $f(x_j)$ and $f(z_j, x_j^{-u})$ are the outputs (t_{crit} or $\phi(t=t_{crit})$) of the model at parameter values x_j compared to the function evaluated at parameter values z_j for one parameter, and x_j for all the remaining parameters. The total effects were calculated for each parameter value for outputs of both critical time (t_{crit}) and phenotypic composition ($\phi(t=t_{crit})$) for four doses ranging from 0 to 500 nM. Large sensitivity indices between parameters and model outputs characteristics indicate that small changes in the parameter values will result in large variations in the

output behavior. For this investigation, to ensure the convergences of the indices, a base simulation size of N=5000 is chosen, resulting in (5000 x 2 x 4 doses x 2 outputs x 8 parameters=640,000) simulations to generate the indices. For this study, only the total effects of the model outputs of t_{crit} and $\phi(t=t_{crit})$ are reported (Supp. Fig. S5 A & B). Specifically, the critical time and phenotypic composition at critical time is recorded for each random simulation and each dose, and per the Sobol method, the total effects indices derived from the variances of these outputs is calculated, which account for variations in individual parameters as well as additional effects resulting from the combined variation of parameters. A sensitivity cut-off of 0.05 is used, indicating parameters that cause less than 5% of the total variation of that model output.

To perform a local sensitivity analysis, we varied each parameter independently from the best fitting parameter set. To perturb each parameter, we chose a high parameter value of two times its optimal value, and a low parameter value of half its optimal value. We used these high and low parameter values, holding all other parameters constant, and ran the forward model and recorded the response over a range of doxorubicin doses from 0-200 nM, for both the effect in critical time (t_{crit}) and phenotypic composition at critical time ($\phi(t=t_{crit})$). For each independent parameter perturbation, we computed a high and low sensitivity score for the the ith parameter, for the two model outputs (t_{crit} or $\phi(t=t_{crit})$) as:

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

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

Which is the sum-squared difference between the output values (t_{crit} or $\phi(t=t_{crit})$) for each jth dose in the range of doses, for both the high and low parameter sets, for each ith parameter. The sum of the high and low sensitivity scores for each parameter were than ranked for the two outputs of t_{crit} and ($\phi(t=t_{crit})$) (Supp. Fig. S5C-F). This analysis reveals the most important parameter in driving changes in output behavior of the model locally around the best fitting parameters.

Model Fitting with Multiple Measurement Sources

To perform model fitting, we used two sources of measurement data: cell number in time ($N(t)$) in response to the pulsed doxorubicin treatments, and estimates of the phenotypic composition, $\phi(t)$, at three time points total (before and two post-treatment). The data were collected in two separate experimental settings, with two different carrying capacities, which we refer to as K_N and K_ϕ. The longitudinal cell number data was recorded in 96 well plates, resulting in a different carrying capacity than the lineage-traced single cell RNA sequencing experiment in which the population was expanded out to a 15 cm dish due to the need for large cell numbers for running on the 10x Genomics system. The carrying capacity of the longitudinal data, K_N, was found by fitting the untreated control to a logistic growth model and allowing both the effective growth rate of the total population (g_{eff}) and K_N to be fit to the data (See Supp. Fig. S6).

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

We set this carrying capacity in the model going forward for fitting the longitudinal data. For the carrying capacity of the single cell RNA sequencing experiment, K_ϕ, we used Thermo-Fisher published "Useful Numbers for Cell Culture" as an estimate (24), where the manufacturer cites the number of cells at confluency of 20 million cells. Going forward, we fit the remaining 6 parameters of θ=[ϕ_0, r_S, r_R,α, d_S, d_R] where these represent: the initial fraction of sensitive cells prior to treatment, the sensitive cell growth rate, the resistant cell growth rate, the rate of drug-induced resistance, the sensitive cell death rate, and the resistant cell death rate, respectively. All six parameters were found to be globally sensitive in one or more of the treatment conditions when looking at either t_{crit} or $\phi(t=t_{crit})$, and so we decided it was reasonable to try to fit them all from the observed data.

To estimate the model parameters θ, we used both measurement sources $N(t)$ and $\phi(t)$ and compared them to their corresponding model outputs. The data were fitted using a weighted-sum-of-squares-residual function described below:

$$J(\theta) = \frac{1}{n_\phi}\sum_{j=1}^{n_\phi}\frac{\left(\widehat{\phi_j} - \phi_j(\theta,u)\right)^2}{\sigma_{\phi_j}^2} + \frac{1}{n_N}\sum_{k=1}^{n_{doses}}\sum_{i=1}^{n_N}\frac{\left(\widehat{N_{i,k}} - N_i\left(\theta,u_{i,k}\right)\right)^2}{\sigma_{N_{i,k}}^2} \quad \text{(Eq.3)}$$

For the $N(t)$ data, the uncertainty in the data (σ^2_N) at each time point was quantified using the standard deviation of the cell number over the six replicate wells. For the uncertainty in the $\phi(t)$ estimates due to sampling a subset of cells from a population of 20 million cells, we compute the Bernoulli sample variance of

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

where n is the number of Bernoulli samples (n=3115, 5251, and 4857 cells in each time point respectively) at each of the three time points. Therefore, the maximum expected sample variance is at ϕ=0.5 and n = 3115, meaning we expect the estimate of the sample ϕ on average to be off by less than 1% from the true mean. However, this is given a true prevalence. This true prevalence is dependent on where the threshold for calling a cell sensitive or resistant is chosen to be, with any values between 0 and 1 technically possible. For this reason, we added an uncertainty term of technical noise σ_{tech}=0.01 to this estimate. In reality, the magnitude of the uncertainty in the $\phi(t)$ is not necessarily known, so we had to estimate a reasonable measurement uncertainty of this magnitude.

In this experimental set up, we have significantly higher time and dose resolution in our $N(t)$ data (472 data points) compared to our $\phi(t)$ data (3 data points), and thus chose to include normalization terms in our objective function (Eq. 3) to account for the different resolutions of the data $N(t)$ and $\phi(t)$ data, and to effectively weight them equally. Because the data come from distinct measurement sources, the robust quantification of comparative uncertainty is not known *a priori*, as we do not intuitively know whether or not the $\phi(t)$ estimates from scRNA-seq are inherently more or less reliable than the longitudinal population size data.

We use the *lsqnonlin* function in MATLAB to search for a set of parameters, θ, that minimizes $J(\theta)$. This set of parameter values was used to make predictions of new doses and also used for the local sensitivity analysis. Additionally, we also performed the calibration without the $\phi(t)$ data to compare the goodness of fit and accuracy of a more

"traditional" method. The following objective function was used for the fitting on longitudinal data only, essentially identical to the integrated calibration just without the $\phi(t)$ data.

$$J(\theta) = \frac{1}{n_N} \sum_{k=1}^{n_{doses}} \sum_{i=1}^{n_N} \frac{\left(\widehat{N_{i,k}} - N_i(\theta, u_{i,k})\right)^2}{\sigma^2_{N_{i,k}}}$$

Structural Identifiability of Model Parameters

We will demonstrate the structural identifiability of the individual model parameters using the differential algebra approach. Structural identifiability of a model and its parameters from a set of measurable outputs tells us that in theory, given perfect data, it is possible to uniquely identify model parameters. Structural identifiability is a pre-requisite for practical identifiability of model parameters from observed data. We start by presenting the non-dimensionalized model and measurement equations, assuming we can measure both N(t) and ϕ(t).

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

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

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

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

We assume all parameters are non-negative and $0 < p_r < 1$ represents the relative growth rate of the resistant population with respect to the sensitive population scaled by the carrying capacity, and $p_r < 1$ assumes that resistant cells grow more slowly than sensitive cells. In work by Greene et al (14), they demonstrate that, if they assume d_r=0, i.e. resistant cells are not killed by drug, and that the initial state of the population is completely comprised of sensitive cells (i.e. N_0=S_0), than the remaining parameters are uniquely identifiable from observations of total cell number alone.

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

We follow the same arguments outlined in (14), along with the complete explanation of the approach with illustrative examples, for the case of multiple outputs from (55). We start by formulating the dynamical system relevant to our in vitro experimental system. Of note, even though we separately measure $N(t)$ and $\phi(t)$ at discrete time points, since this analysis is for structural identifiability and assumes perfect, noise-free data, we will transform the observable outputs of $N(t)$ and $\phi(t)$ into:

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

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

Treatment is initiated at time t=0, at which we make no assumptions about the composition of the population such that $S(0) = S_0$, $R(0) = R_0$. Here $0<S_0+R_0<1$. We note

this is due to the non-dimensionalization in which we now track the proportion of confluent cells i.e. $S(t) = \frac{S\prime(t)}{K}$ and $R(t) = \frac{R\prime(t)}{K}$ (see (14)) for additional details. We can now formulate our system in input/output form as:

$$\dot{x}(t) = f\big(x(t)\big) + u(t)g\big(x(t)\big)$$
$$x(0) = x_0$$

Where *f* and *g* are:

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

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

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

$$y_1(t) = h_1\big(x(t)\big) = x_1(t)$$
$$y_2(t) = h_2\big(x(t)\big) = x_2(t)$$

A system in this form is said to be uniquely structurally identifiable if the map $(p, u(t)) \rightarrow (x(t,p), u(t))$ is injective (55–57), where p is the vector of parameters to be identified. In this instance $p = (S_0, R_0, d_s, d_r, \alpha, p_r)$, the initial states and the parameters. Local identifiability and non-identifiability correspond to the map being finite-to-one and infinite-to-one, respectively. Our objective is then to demonstrate unique structural identifiability for model system and hence recover all parameter values p from the assumption of perfect, noise-free data.

To analyze identifiability, we utilize results appearing in (14) and (55), where a differential-geometric perspective is used. For the structural identifiability, we hypothesize that we have perfect (hence noise-free) input-output data is available of the form of y_1 and y_2 and its derivatives on any interval of time. We then, for example, make measurements of:

$$y_1(0) = h_1(x_1(0))$$
$$\dot{y_1}(0) = \frac{\partial}{\partial t}\Big|_{t=0} h_1(x_1(t))$$
$$y_2(0) = h_2(x_2(0))$$
$$\dot{y_2}(0) = \frac{\partial}{\partial t}\Big|_{t=0} h_2(x_2(t))$$

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

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

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

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

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

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

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

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

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

As is outlined in (55), (58) proved that F_1=F_2, so that the iterated Lie derivatives F_2 may be considered as the set of "elementary observables". Hence, identifiability may be formulated in terms of the reconstruction of parameters *p* from elements in F_2. Parameters p are then identifiable if the map

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

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

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

$$L_f h_1 = (1 - \mathrm{x}_1 - \mathrm{x}_2)\mathrm{x}_1$$
$$L_f h_2 = p_r(1 - \mathrm{x}_1 - \mathrm{x}_2)\mathrm{x}_2$$
$$L_g h_1 = (\alpha + \mathrm{d_s})\mathrm{x}_1$$
$$L_g h_2 = \alpha x_1 - d_r x_2$$
$$L_f L_g h_2 = \alpha x_1(1 - x_1 - x_2) - d_r p_r x_2(1 - x_1 - x_2)$$

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

$$S_0 = h_1(x_0)$$
$$R_0 = h_2(x_0)$$
$$p_r = \frac{L_f h_2}{R_0(1 - S_0 - R_0)}$$
$$d_r = \frac{1}{R_0(1 - p_r)}\left(\frac{L_f L_g h_2}{1 - S_0 - R_0} - L_g h_2\right)$$
$$\alpha = \frac{L_g h_2 + d_r R_0}{S_0}$$
$$d_s = \frac{L_g h_1}{S_0} - \alpha$$

Since F_1 = F_2, all of the above Lie derivatives are observable via appropriate treatment protocols. Thus by incorporating knowledge of $\phi(t)$, all parameters in system 1 are

structurally identifiable. This represents an improvement over the identifiability with $N(t)$ alone as a measurable output and allows us to introduce a non-zero d_R parameter, which we have reason to believe based on experimental evidence, is the more biologically relevant scenario.

Author Contributions

KJ and AB designed the study; GH, DM, EB, AG, and AA performed experiments; WM curated the data; KJ, GH, DM, EB, AG, RD and WM analyzed the data; KJ performed mathematical modeling; ES, AJ, TY advised on mathematical modeling, KJ and AB wrote the manuscript with input from all authors; all authors read and approved the manuscript.

Acknowledgements

The authors are grateful for grant support from the NIH iMAT program (R21CA212928 to AB), CPRIT (RR1600005 to TEY), NCI (U01CA174706 and R01CA186193 to TEY) and NSF Grants #1716623 and #1849588 to EDS). KJ was supported by an NSF Graduate Research Fellowship 1610403. T.E.Y. is a CPRIT Scholar of Cancer Research. The authors also thank the Genomic and Sequencing Analysis Facility at the University of Texas at Austin and Dennis Wylie for advice throughout the project.

References

1. Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM. Leveraging Single-Cell RNA Sequencing Experiments to Model Intratumor Heterogeneity. Clin Cancer Informatics [Internet]. 2019;1–10. Available from: https://doi.org/10.1200/CCI.18.00074
2. Syed AK, Woodall R, Whisenant JG, Yankeelov TE, Sorace AG. Characterizing Trastuzumab- Induced Alterations in Intratumoral Heterogeneity with Quantitative Imaging and Immunohistochemistry in HER2 + Breast Cancer. Neoplasia [Internet]. 2019;21(1):17–29. Available from: https://doi.org/10.1016/j.neo.2018.10.008
3. Pyne S, Hu X, Wang K, Rossin E, Lin T, Maier LM, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci [Internet]. 2009;106(21):8519–24. Available from: www.pnas.org/cgi/doi/10.1073/pnas.0903028106
4. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11(2).
5. Guo J, Grow EJ, Yi C, Micochova H, Maher GJ, Lindskog C, et al. Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. Cell Stem Cell [Internet]. 2018;21(4):533–46. Available from: https://doi.org/10.1016/j.stem.2017.09.003

6. Kumar MP, Du J, Lagoudas G, Yang J, Sawyer A, Drummond DC, et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep [Internet]. 2018;25(6):1458-1468.e4. Available from: https://doi.org/10.1016/j.celrep.2018.10.047
7. Wang Y, Wang R, Zhang S, Song S, Jiang C, Han G, et al. iTALK : an R Package to Characterize and Illustrate Intercellular Communication. bioRxiv [Internet]. 2019; Available from: https://dx.doi.org/10.1101/507871
8. Zhao X, Wu S, Fang N, Sun X, Fan J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. 2019;00(July):1–15.
9. Al'Khafaji A, Gutierrez C, Brenner E, Durrett R, Johnson KE, Zhang W, et al. Expressed barcodes enable clonal characterization of chemotherapeutic responses in chronic lymphocytic leukemia. bioRxiv [Internet]. 2019;1–24. Available from: https://dx.doi.org/10.1101/761981
10. Smalley I, Kim E, Li J, Spence P, Wyatt CJ, Eroglu Z, et al. Leveraging transcriptional dynamics to improve BRAF inhibitor responses in melanoma. EBioMedicine [Internet]. 2019; Available from: https://doi.org/10.1016/j.ebiom.2019.09.023
11. Stumpf PS, Smith RCG, Lenz M, Schuppert A, Müller FJ, Babtie A, et al. Stem Cell Differentiation as a Non-Markov Stochastic Process. Cell Syst. 2017;5(3):268-282.e7.
12. Brady R, Nagy JD, Gerke TA, Zhang T, Wang AZ, Zhang J, et al. Prostate-Specific Antigen Dynamics Predict Individual Responses to Intermittent Androgen Deprivation. bioRxiv [Internet]. 2019; Available from: https://dx.doi.org/10.1101/624866
13. McKenna MT, Weis JA, Quaranta V, Yankeelov TE. Variable Cell Line Pharmacokinetics Contribute to Non-Linear Treatment Response in Heterogeneous Cell Populations. Ann Biomed Eng [Internet]. 2018/02/26. 2018 Jun;46(6):899–911. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29484528
14. Greene JM, Sanchez-Tapia C, Sontag ED. Mathematical Details on a Cancer Resistance Model. bioRxiv [Internet]. 2018;1–42. Available from: https://dx.doi.org/10.1101/475533
15. Greene JM, Gevertz JL, Sontag ED. Mathematical Approach to Differentiate Spontaneous and Induced Evolution to Drug Resistance During Cancer Treatment abstract. JCO Clin Cancer Informatics. 2019;42–9.
16. Gevertz JL, Greene JM, Sontag ED. Validation of a Mathematical Model of Cancer Incorporating Spontaneous and Induced Evolution to Drug Resistance. bioRxiv [Internet]. 2019;1–15. Available from: http://dx.doi.org/10.1101/2019.12.27.889444
17. Gatenby RA, Silva AS, Gillies RJ, Frieden BR. Adaptive therapy. Cancer Res. 2009;69(11):4894–903.
18. Prokopiou S, Moros EG, Poleszczuk J, Caudell J, Torres-roca JF, Latifi K, et al. A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation. Radiat Oncol [Internet]. 2015;1–8. Available from: http://dx.doi.org/10.1186/s13014-015-0465-x
19. Howard GR, Johnson KE, Ayala AR, Yankeelov TE, Brock A. A multi-state model

of chemoresistance to characterize phenotypic dynamics in breast cancer. Sci Rep [Internet]. 2018;(July):1–11. Available from: http://dx.doi.org/10.1038/s41598-018-30467-w

20. Pisco AO, Brock A, Zhou J, Moor A, Mojtahedi M, Jackson D, et al. Non-Darwinian dynamics in therapy-induced cancer drug resistance. Nat Commun. 2013;4(2467).
21. McKenna MT, Weis JA, Quaranta V, Yankeelov TE. Variable Cell Line Pharmacokinetics Contribute to Non-Linear Treatment Response in Heterogeneous Cell Populations. Ann Biomed Eng. 2018;46(6):899–911.
22. McKenna MT, Weis JA, Barnes SL, Tyson DR, Miga MI, Quaranta V, et al. A Predictive Mathematical Modeling Approach for the Study of Doxorubicin Treatment in Triple Negative Breast Cancer. Sci Rep [Internet]. 2017;7(1):1–14. Available from: http://dx.doi.org/10.1038/s41598-017-05902-z
23. Al'Khafaji AM, Deatherage D, Brock A. Control of Lineage-Specific Gene Expression by Functionalized gRNA Barcodes. ACS Synth Biol. 2018;
24. Thermo Fisher Scientific. Useful Numbers for Cell Culture [Internet]. [cited 2020 Feb 11]. Available from: https://www.thermofisher.com/us/en/home/references/gibco-cell-culture-basics/cell-culture-protocols/cell-culture-useful-numbers.html
25. Efron B. Better Bootstrap Confidence Intervals. 1987;82(397):171–85.
26. Press WH, Teukolsky S a., Vetterling WT, Flannery BP. Numerical Recipes in Forttran 77: the art of scientific computing. Cambridge University Press. Numerical Recipes Software; 1992. 684–694 p.
27. Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell. 2019;75(1):7–12.
28. Levitin HM, Yuan J, Sims PA. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends in Cancer. 2018 Apr;4(4):264–8.
29. Rockne RC, Hawkins-daarud A, Swanson KR, Sluka JP, Glazier JA, Macklin P, et al. The 2019 mathematical oncology roadmap The 2019 mathematical oncology roadmap. 2019;
30. McKenna MT, Weis JA, Brock A, Quaranta V, Yankeelov TE. Precision Medicine with Imprecise Therapy: Computational Modeling for Chemotherapy in Breast Cancer. Transl Oncol [Internet]. 2018;11(3):732–42. Available from: https://doi.org/10.1016/j.tranon.2018.03.009
31. Jarrett AM, Lima EABF, Hormuth DA, McKenna MT, Feng X, Ekrut DA, et al. Mathematical models of tumor cell proliferation: A review of the literature. Expert Rev Anticancer Ther [Internet]. 2018 Dec 2;18(12):1271–86. Available from: https://doi.org/10.1080/14737140.2018.1527689
32. Poleszczuk J, Enderling H. The Optimal Radiation Dose to Induce Robust Systemic Anti-Tumor Immunity. Int J Mol Sci. 2018;19(11).
33. Zhang Y, Huynh JM, Liu G, Ballweg R, Aryeh KS, Paek AL, et al. Designing combination therapies with modeling chaperoned machine learning. PLoS Comput Biol [Internet]. 2019;15(9):1–17. Available from: http://dx.doi.org/10.1371/journal.pcbi.1007158
34. Badri H, Pitter K, Holland EC, Michor F, Leder K. Optimization of radiation dosing schedules for proneural glioblastoma. J Math Biol. 2016;72(5):1301–36.

35. Poleszczuk J, Enderling H, Poleszczuk J, Enderling H. Cancer Stem Cell Plasticity as Tumor Growth Promoter and Catalyst of Population Collapse. Stem Cells Int [Internet]. 2016;2016:1–12. Available from: http://www.hindawi.com/journals/sci/2016/3923527/
36. Greene JM, Levy D, Fung KL, Souza PS, Gottesman MM, Lavi O. Modeling intrinsic heterogeneity and growth of cancer cells. J Theor Biol [Internet]. 2015;367:262–77. Available from: http://dx.doi.org/10.1016/j.jtbi.2014.11.017
37. Jarrett AM, Bloom MJ, Ekrut DA, Yankeelov TE. Mathematical modelling of trastuzumab-induced immune response in an in vivo murine model of HER2 + breast cancer. 2018;2:1–30.
38. Hormuth DA, Jarrett AM, Feng X, Yankeelov TE. Calibrating a Predictive Model of Tumor Growth and Angiogenesis with Quantitative MRI. Ann Biomed Eng. 2019;47(7):1539–51.
39. Yankeelov TE, Atuegwu N, Hormuth D, Weis JA, Barnes SL, Miga MI, et al. Clinically Relevant Modeling of Tumor Growth and Treatment Response. 2013;5(187):1–6.
40. Yankeelov TE, Quaranta V, Evans KJ, Rericha EC. Toward a science of tumor forecasting for clinical oncology. Cancer Res. 2015;75(6):918–23.
41. Ma K-Y, Schonnesen AA, Brock A, Van Den Berg C, Eckhardt SG, Liu Z, et al. Single-cell RNA sequencing of lung adenocarcinoma reveals heterogeneity of immune response–related genes. JCI Insight [Internet]. 2019 Feb 21;4(4). Available from: https://doi.org/10.1172/jci.insight.121387
42. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis : a tutorial. Mol Syst Biol. 2019;15(e8746).
43. Nam A, Mohanty A, Bhattacharya S, Kotnala S, Achuthan S. Suppressing chemoresistance in lung cancer via dynamic phenotypic switching and intermittent therapy. 2020;bioRxiv. Available from: https://doi.org/10.1101/2020.04.06.028472
44. He J, Ahuja N. Personalized Approaches to Gastrointestinal Cancers: Importance of Integrating Genomic Information to Guide Therapy. Surg Clin North Am. 2015;95(5):1081–94.
45. Kowarz E, Loescher D, Marschalek R. Optimized Sleeping Beauty transposons enable robust stable transgenic cell lines. Biotechnol J. 2015;41:647–53.
46. Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat Genet. 2009;41(6):753–61.
47. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol [Internet]. 2018;19(1):15. Available from: https://doi.org/10.1186/s13059-017-1382-0
48. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods [Internet]. 2019;16(1):43–9. Available from: https://doi.org/10.1038/s41592-018-0254-1
49. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun [Internet]. 2019;10(1):4667. Available from: https://doi.org/10.1038/s41467-019-12266-7
50. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep [Internet]. 2019;9(1):5233. Available from:

https://doi.org/10.1038/s41598-019-41695-z
51. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol [Internet]. 2016;17(1):75. Available from: https://doi.org/10.1186/s13059-016-0947-7
52. Tirosh I, Izar B, Prakadan SM, Ii MHW, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (80-). 2019;352(6282).
53. Anderson ARA, Weaver AM, Cummings PT, Quaranta V. Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. Cell. 2006;127(5):905–15.
54. Jarrett AM, Liu Y, Cogan NG, Hussaini MY. Global sensitivity analysis used to interpret biological experimental results. J Math Biol [Internet]. 2015;71:151–70. Available from: http://dx.doi.org/10.1007/s00285-014-0818-3
55. Sontag ED. Dynamic compensation, parameter identifiability, and equivariances. PLoS Comput Biol. 2017;13(4):1–17.
56. Eisenberg MC. Input-output equivalence and identifiability: some simple generalizations of the differential algebra approach. arXiv. 2019;1–25.
57. Brouwer AF, Meza R, Eisenberg MC, Arbor A. A systematic approach to determining the identifiability of multistage carcinogenesis models. Risk Anal. 2017;37(7):1375–87.
58. Wang Y, Sontag ED. On two definitions of observation spaces. Syst Control Lett. 1989;13:213–8.

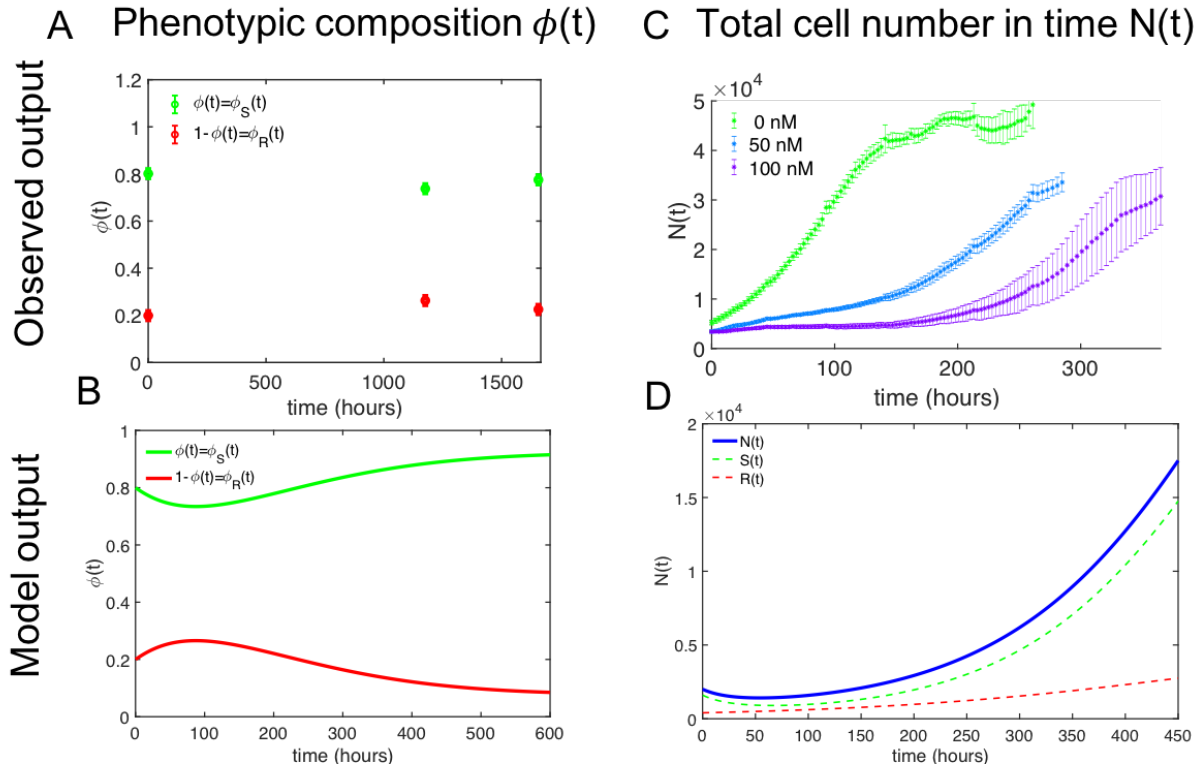
Supplemental Information

Integrating multimodal data sets into a mathematical framework to describe and predict therapeutic resistance in cancer

Kaitlyn Johnson, Daylin Morgan, Eric Brenner, Andrea Gardner, Russ Durrett, Grant Howard, Eduardo Sontag, Angela Jarrett, Thomas E. Yankeelov, Amy Brock

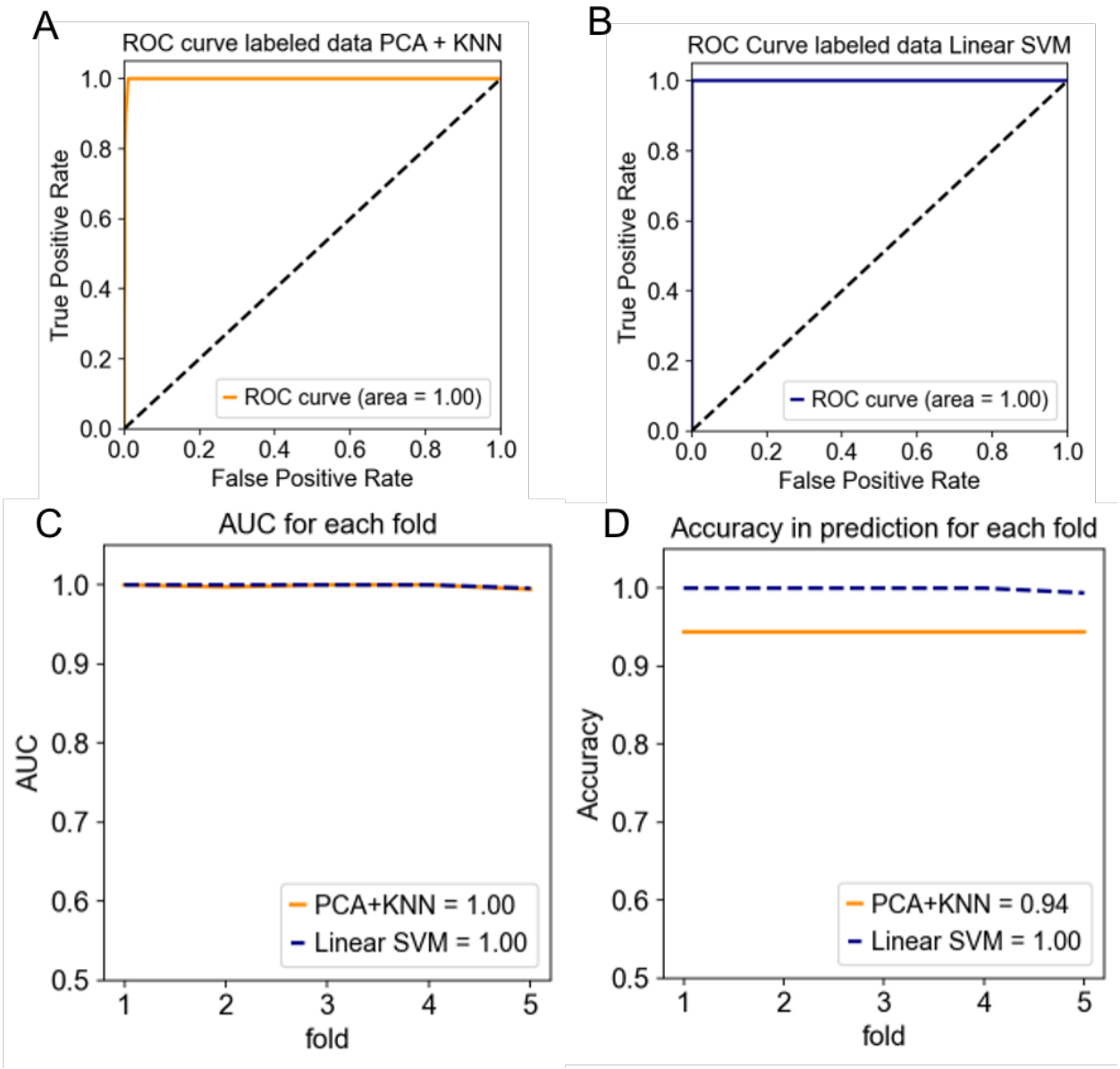
	ϕ_0	r_s	α	r_r	d_s	d_r	K_N	K_ϕ	k	k_{drug}
Integrated fit	0.8896 [0.8696, 0.9092]	0.0212 [0.0207, 0.0213]	0.0178 [0.0077, 0.0230]	0.0056 [0.0037,0.0100]	0.0621 [0.0564, 0.0731]	0.0935 [0,0.0239]	4.965e4	2e7	0.5	0.13175
N(t) only fit	0.8389 [0.6848, 0.9063]	0.0269 [0.0213, 0.0236]	0.0157 [0.0137, 0.0307]	0.0134 [0.055, 0.0797]	0.0183 [0.188, 0.188]	4.5847e-16 [0,0.0047]	4.965e4	2e7	0.5	0.13175

Supplementary Table S1. Estimated parameter values from the integrated model fit (using N(t) and phi(t) and using N(t) only. The first six parameters are calibrated to data, the last four are set (and thus are the same for both calibration schemes). Confidence intervals from fit parameters are estimated using bootstrapping parameter estimates.

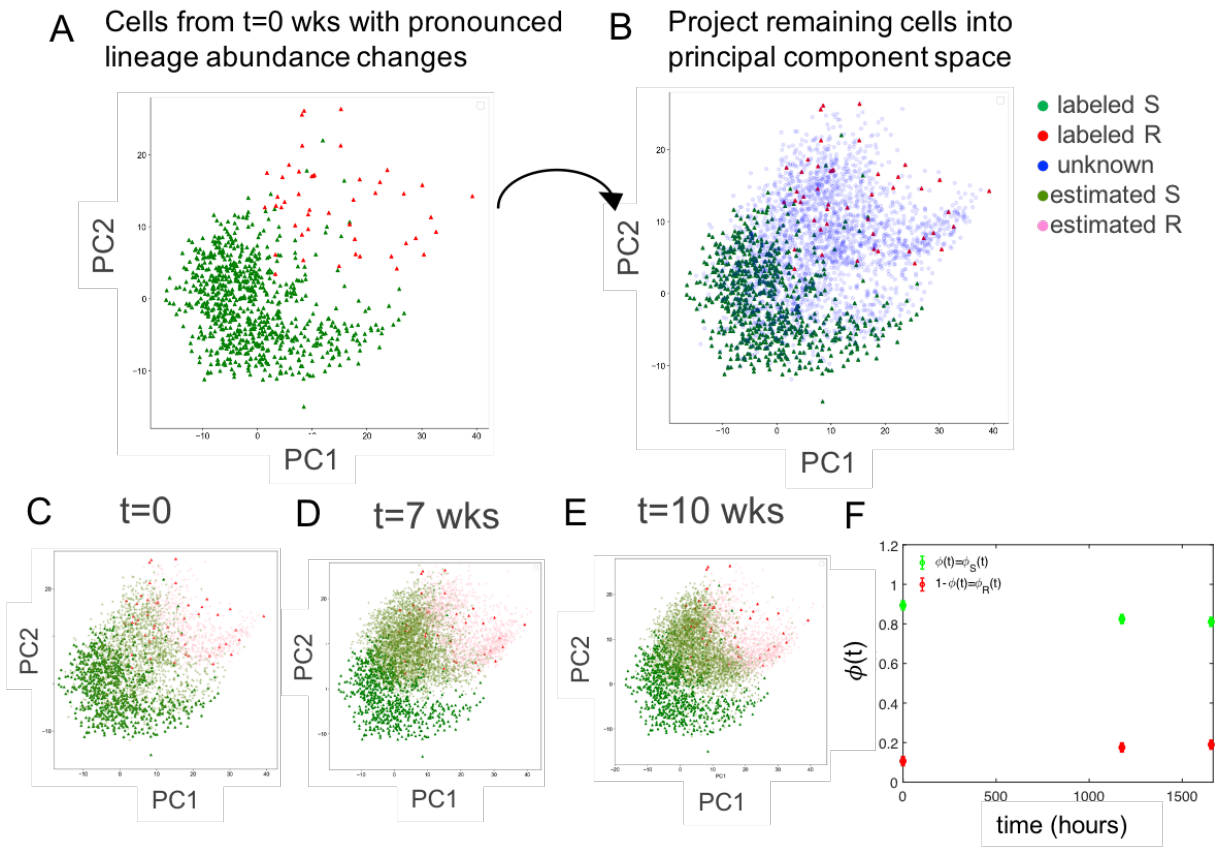


Supplementary Figure S1. Measured and model predicted outputs to be used for parameter estimation from observed data A. Observed estimated fraction of sensitive cells (green) and resistant cells (red) from scRNAseq classifier at three time points ϕ(t). B. Model predicted output of sensitive cell fraction dynamics (green) and resistant cell fraction dynamics (red) for an example parameter set. C. Observed number of tumor cells

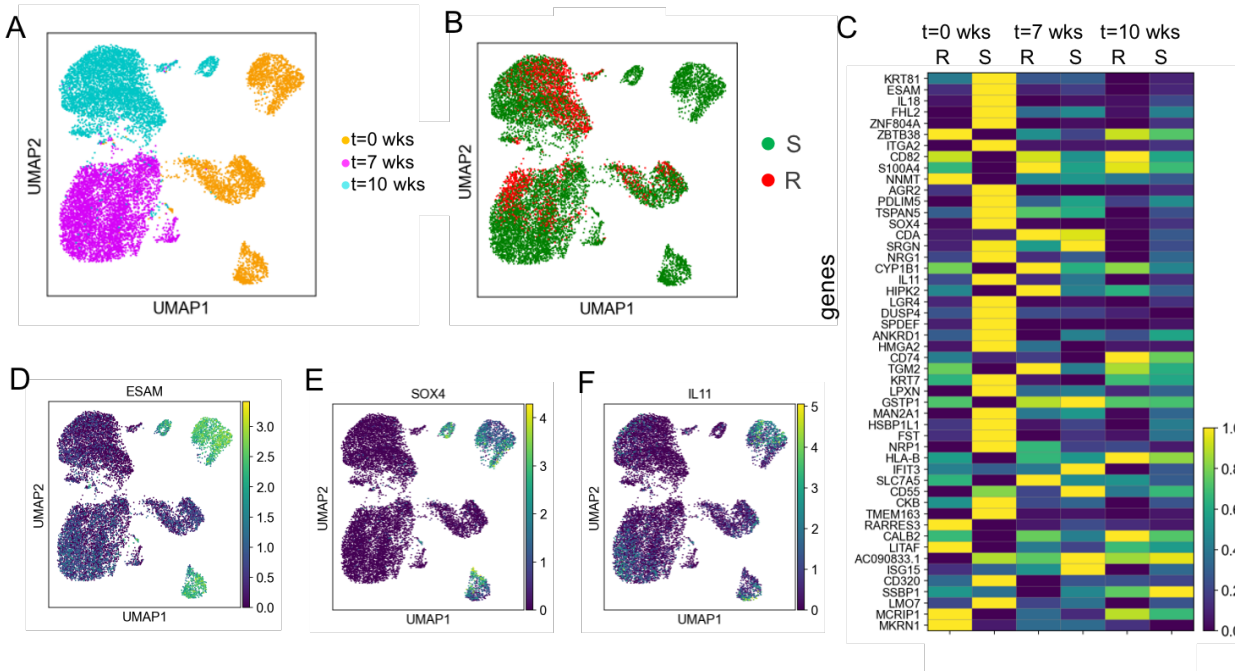
in time for pulse treatments of doxorubicin at 0, 50, and 100 nM, the doses used for model calibration. D. Model predicted output of total cell number in time for a single pulse treatment simulated from the model and an arbitrary example parameter set.



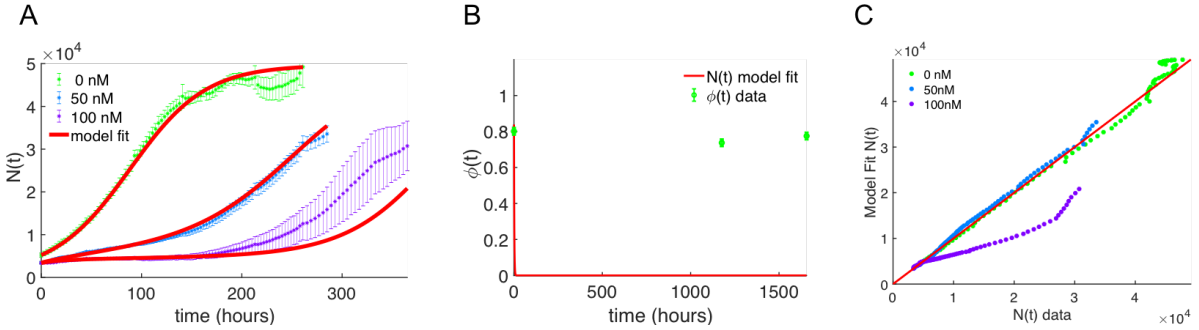
Supplementary Figure S2. Comparison of classifiers for estimating sensitive and resistant cells. A. ROC curve from PCA + KNN classifier B. ROC curve from Linear SVM classifier C. AUC from ROC curve for each of 5 folds cross validation data sets D. Accuracy of classification of the testing set data in each fold of cross validation reveals Linear SVM is consistently more accurate than PCA + KNN



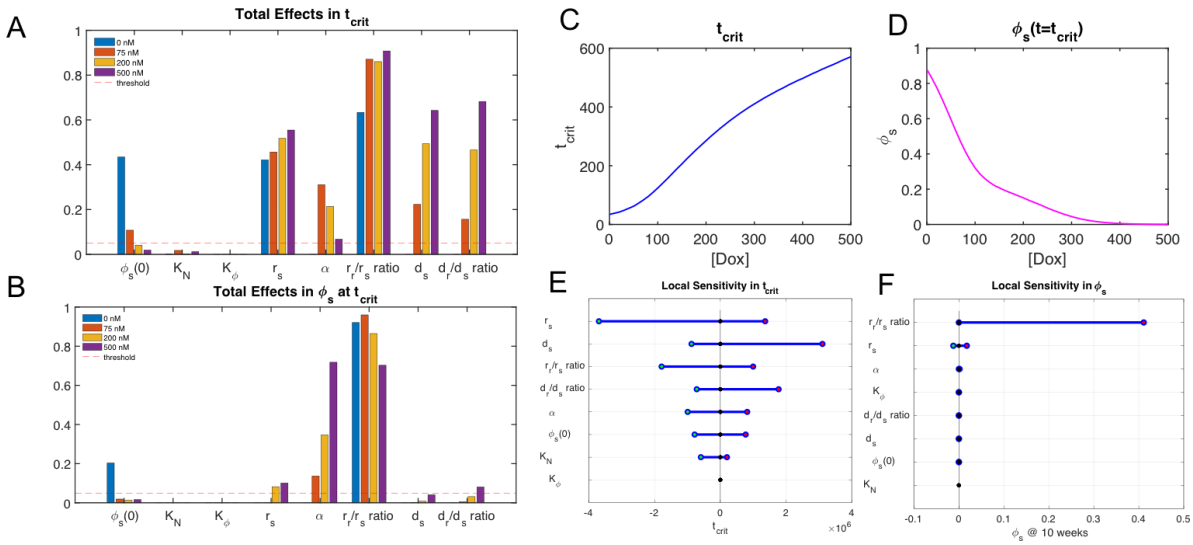
Supplementary Figure S3. Single cell transcriptomes from each time point projected into principal component space and classified using nearest neighbors A. Lineage-abundance guided "labeled" cells projected into principal component space separate along components (PC1 and PC2 shown here for visual effect). B. Unknown cells are projected into the principal component space of the labeled cells. C. Remaining cells from t=0 projected onto labeled cells in PC space and estimated as sensitive (olive) or resistant (green). D. Cells from t=7 weeks projected alongside labeled cells. E. Cells from t=10 weeks projected alongside labeled cells. F. Proportion of cells in each time point that are estimated or labeled as sensitive (green), or resistant (red).



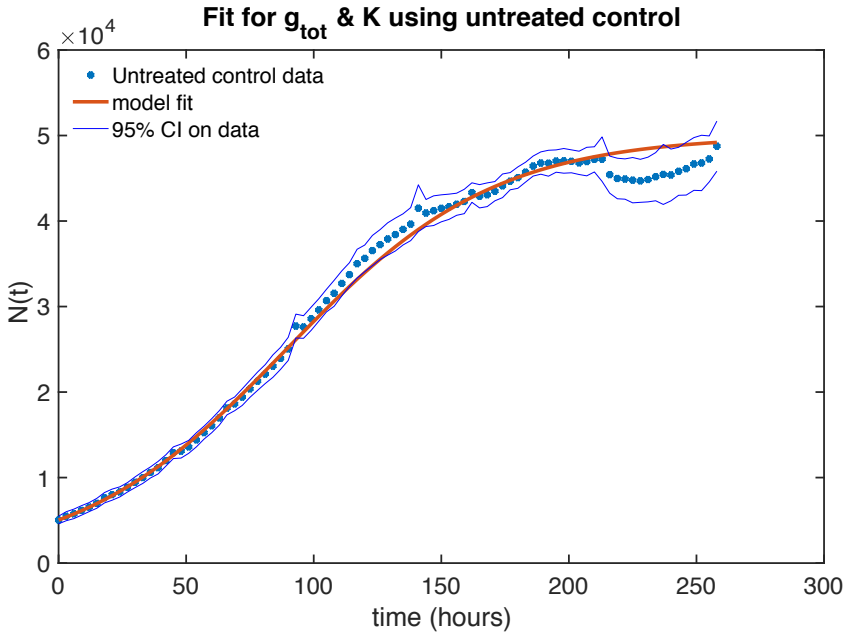
Supplementary Figure S4. Differential Gene Expression Analysis Provide Molecular Insight into Drug Resistance Interactions A. UMAP projection of single cell transcriptomes colored by time point B. Single cells colored by sensitive and resistant cell labels visualized via UMAP projections indicates drug sensitivity phenotypes cluster together, but not exclusively by the apparent UMAP clustering n C. Heat map of the top 50 gene weights in the Linear SVM, comparing the average expression across the sensitive and resistant cell groups in the three time points. The colorbar is scaled within each gene (row). D. UMAP projections of cells colored by expression level of ESAM indicates high expression of UBE2S is associated with sensitivity. G. UMAP projections of cells colored by expression level of SOX4 indicates that low expression of SOX4 is associated with sensitivity. I. UMAP projections of cells colored by expression level of IL11 indicates that high expression of IL11 is associated with sensitivity.



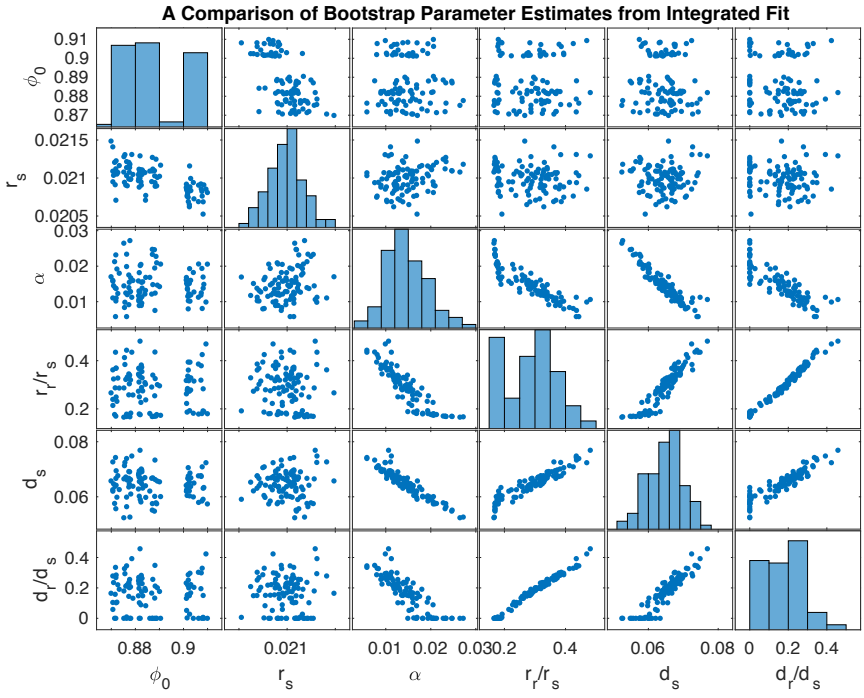
Supplementary Figure S5. Model calibration using only N(t) data. A. Calibration results for longitudinal $N(t)$ data from the four doses (0, 50, and 100 nM) used for calibration B. Comparison of model fit to estimates of phenotypic composition ($\phi(t)$). This information was not used for calibration, hence why the error is extremely large. C. Measured cell number $N(t)$ verses model calibrated cell number, yielding a concordance in $N(t)$ of CCC = 0.975.



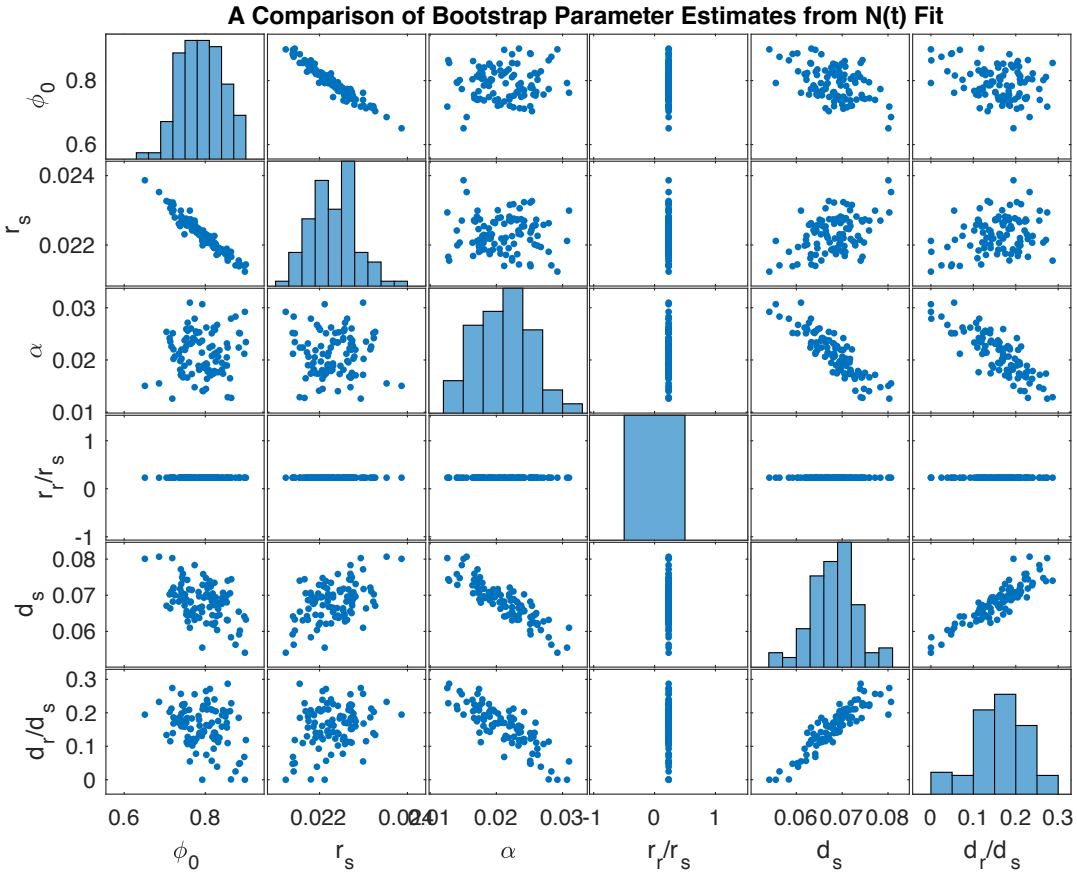
Supplementary Figure S6. Sensitivity Analysis of Model Parameters Reveals All Parameters are Locally and Globally Sensitive Under Treatment. A. Sobol's total effects of each parameter globally on critical time for 0,75, 200, and 500 nM pulse treatments reveals that all fit parameters are above the threshold of sensitivity for at least one of those doses (the parameter contributes at least 5% to the critical time for at least one of the doxorubicin concentrations). B. Sobol's total effects of each parameter globally on sensitive cell fraction for 0, 75, 200 and 500 nM pulse treatments reveals that most fit parameters are above the threshold of sensitivity for at least one of the doses. The carrying capacity of the single cell RNA sequencing experiment (K_2) is the only parameter that is not above the threshold for any sensitivity analysis output or dose, and for this reason supports our decision to set that carrying capacity from a literature value (the expected number of 231 cells at confluence in a 10 cm dish, which the cells were expanded up to). C. An example of the model predicted critical time as a function of doxorubicin concentration, taken from the selected parameter set in red in Fig 5A. Critical time is chosen as an output for model sensitivity because it evaluates treatment response and drug sensitivity in of a cell population:drug concentration combination without biasing for response dynamics that might vary from system to system, and because it is most relevant to what we experimentally are able to observe (i.e. the cells rebounded to 2 times their initial cell number on this day). D. An example of the model predicted sensitive cell fraction at the critical time as a function of doxorubicin concentration, again for the selected parameter set in red in Fig 5A. This was chosen again because of its relevance to experimental workflows, as the time at which the population rebounds to 2 the seeding population might be a good time at which we could perform an experimental analysis of the tumor cell composition (i.e. scRNAseq). E. Local sensitivity in critical time produced by varying the selected parameter set by 50% above and below its value and recording the resulting change in critical time trajectory over a doxorubicin range of 0 to 500 nM. F. Local sensitivity in sensitive cell fraction at critical time produced by again varying the selected parameter set by 50% above and below its value and recording the resulting change in sensitive cell fraction over a doxorubicin range of 0 to 500 nM.



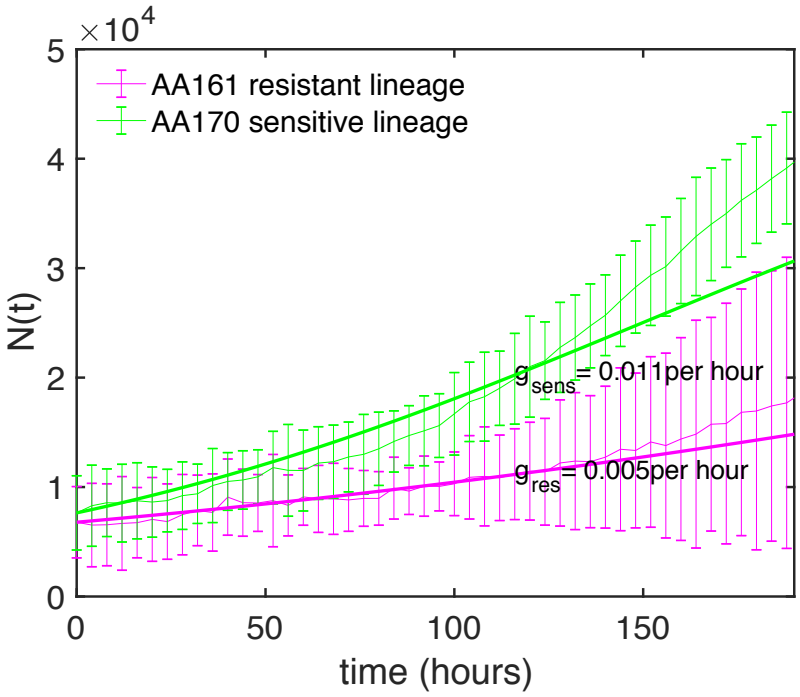
Supplementary Figure S7. Fit to untreated control to find carrying capacity (K_N) of MDA-MB- 231 cells in a 96 well plate.



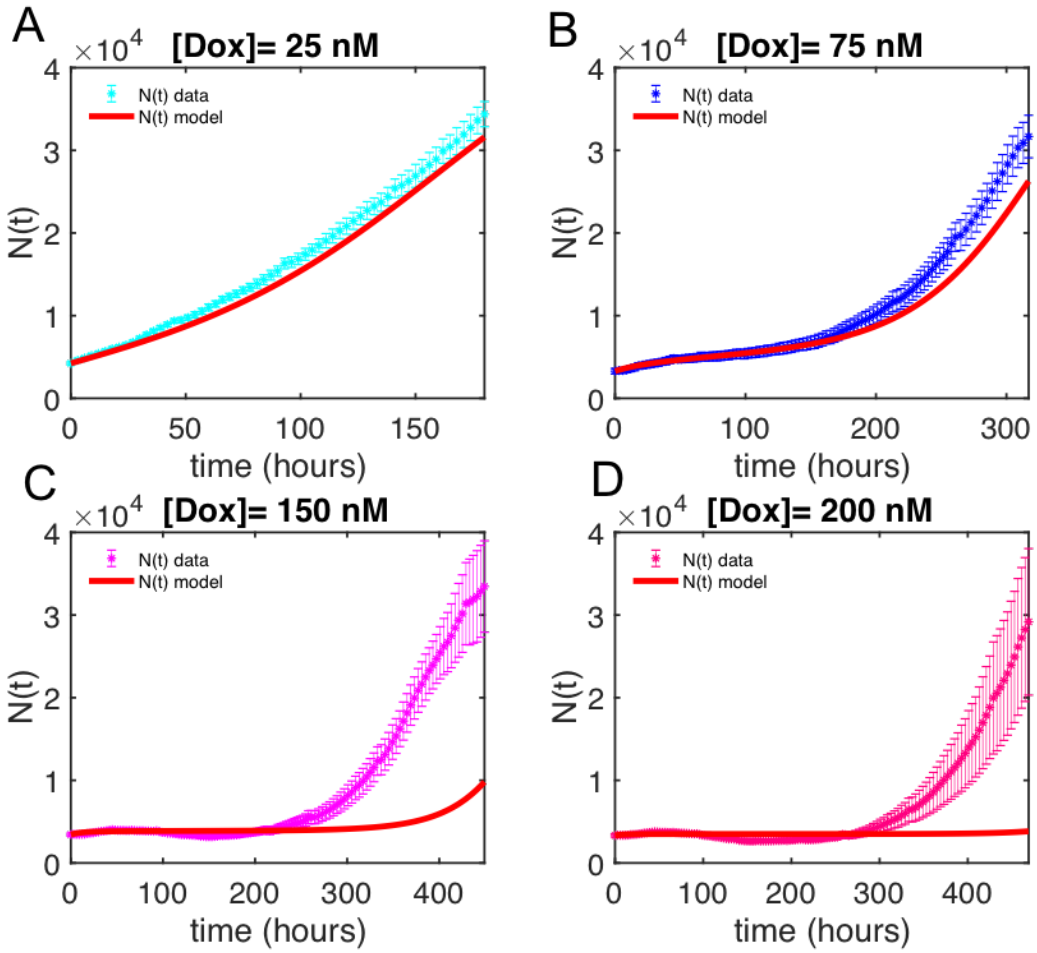
Supplementary Figure S8. **Visualization of the distribution of parameter estimates in the bootstrapped parameter set for the integrated calibration (from N(t) and ϕ(t))**. For each parameter, the 2.5th and 97.5th percentiles were found from 100 simulated data sets to construct the 95% confidence intervals around each parameter value.



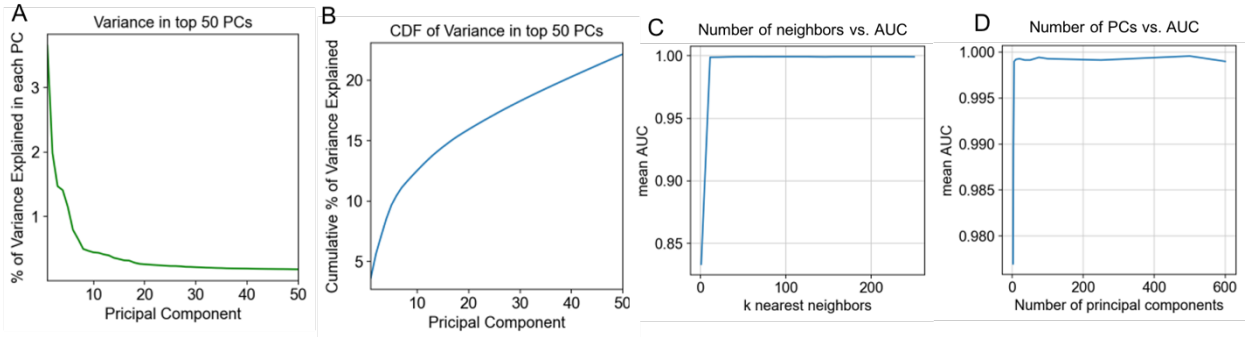
Supplementary Figure S9. **Visualization of the distribution of parameter estimates in the bootstrapped parameter set for calibration from N(t) data only**. For each parameter, the 2.5th and 97.5th percentiles were found from 100 simulated data sets to construct the 95% confidence intervals around each parameter value. It is evident that the growth rate parameter is not identifiable as it doesn't change from the initial guess. This is likely due to insufficient data for the N(t) calibration scheme to fit to the 6 free parameters of interest. If we were only able to use this data, we would need to set some parameters from literature or other experiments in order to obtain identifiability.



Supplementary Figure S10. Growth dynamics of isolated sensitive and resistant cell lineages indicates that sensitive cells growth on more quickly than the resistant cells, validating our modeling assumptions.



Supplementary Figure S11. Model predicted treatment response from longitudinal N(t) calibration only. Prediction of treatment response at A. 25 nM B. 75 nM C. 150 nM and D. 200nM from the N(t) calibration using the other doses. No phenotypic composition data was used to calibrate the model parameters that were used to predict the new treatment response.



Supplementary Figure S12. Variance explained in each PC and hyperparameter optimization for PCA + KNN. A. Proportion of variance explained by the top 50 principal components PCs B. Cumulative variance in each successive principal component for the top 50 PCs. C. Number of nearest neighbors used in the classifier versus mean AUC from

5-fold CV to determine optimal number of neighbors of k=73. C. Number of principal components used in the classifier versus mean AUC from 5-fold CV to determine optimal number of components, n=500. D. ROC curve from classifier with optimized number of nearest neighbors and components for separating labeled cells.