Combining mathematical and statistical modeling to simulate time course bulk and single cell gene expression data in cancer with CancerInSilico

Thomas D Sherman[1], Luciane T Kagohara[1], Raymon Cao[1], Raymond Cheng[2], Matthew Satriano[3], Michael Considine[1], Gabriel Krigsfeld[1], Ruchira Ranaweera[4], Yong Tang[5], Sandra A Jablonski[6], Genevieve Stein-O'Brien[1,7], Daria A Gaykalova[8], Louis M Weiner[6], Christine H Chung[4], and Elana J Fertig[1]

[1] Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD USA
[2] Science, Math and Computer Science Magnet Program, Poolesville High School, Poolesville, MD USA
[3] Department of Mathematics, University of Waterloo, Waterloo, Ontario, Canada
[4] Moffitt Cancer Center, Tampa, FL, USA
[5] Salubris Biotherapeutics, Inc, Gaithersburg, MD, USA
[6] Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC USA
[7] Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD USA
[8] Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD USA

Contact tsherma4@jhu.edu, ejfertig@jhmi.edu

Abstract

Motivation: Bioinformatics techniques to analyze time course bulk and single cell omics data are advancing. The absence of a known ground truth of the dynamics of molecular changes challenges benchmarking their performance on real data. Realistic simulated time-course datasets are essential to assess the performance of time course bioinformatics algorithms.

Results: We develop an R/Bioconductor package CancerInSilico to simulate bulk and single cell transcriptional data from a known ground truth obtained from mathematical models of cellular systems. This package contains a general R infrastructure for cell-based mathematical model, implemented for an off-lattice, cell-center Monte Carlo mathematical model. We also adapt this model to simulate the impact of growth suppression by targeted therapeutics in cancer and benchmark simulations against bulk *in vitro* experimental data. Sensitivity to parameters is evaluated and used to predict the relative impact of variation in cellular growth parameters and cell types on tumor heterogeneity in therapeutic response.

Availability and Implementation: CancerInSilico is implemented in an R/Bioconductor package by the same name. Applications presented are available from https://github.com/FertigLab/CancerInSilico-Figures.

Introduction

Time course bioinformatics analysis techniques are emerging to delineate cellular composition and pathway activation from longitudinal genomics data [1,2]. However, benchmarking their performance is challenged by a lack of ground truth of the processes occurring in those datasets. For example, even relatively simple covariates, such as cellular density and proliferation rates impact experimental measures at a given time point, such as therapeutic sensitivity in cancer [3]. The interactions between these processes will introduce even greater complexity in gene-level dynamics. Simulated data can enable such benchmarking of bioinformatics analysis methods for omics data. Statistical methods that utilize expected gene expression profiles from reference datasets to model the error distribution of bulk and single cell sequencing data are prominent [4–6]. Yet, there are few time course omics datasets to use as a benchmark and even fewer with known cellular-molecular dynamics. Therefore, new simulation systems with known ground truth are needed to benchmark the performance of emerging time course bioinformatics algorithms for bulk and single cell datasets.

Mathematical models of cellular dynamics are maturing in systems biology and can be used to track the unmeasurable state of the processes occurring in each cell in complex biological systems, such as cancer [7–13]. Some models simulate cell growth at a cellular level, where the population behavior is driven by the laws governing the individual cells and their interactions [14,15]. To further capture the complexity of biological systems, numerous multiscale and hybrid models linking cellular signaling to the equations of the cellular composition are emerging [16–18]. These models often require numerous parameters to simulate high throughput proteomic and transcriptional data. Therefore, these models that simulate high throughput data often have similar complexity to real biological systems. Thus, it is nearly as intractable to benchmark the performance of time course bioinformatics algorithms against these simulated data as it is for high throughput time course data generated from biological experiments. Another challenge to the adoption of mathematical models for bioinformatics benchmarking is a lack of software implementing them in the R language widely used by the bioinformatics community. Creating a mathematical model-based simulation of genomics data in R enables direct application to use these simulations as a ground truth for time-course bioinformatics techniques for bulk and single cell data.

In this paper, we present a new software package to simulate time course transcriptional data. The model is implemented in the R/Bioconductor package CancerInSilico and includes extensions to simulate single cell RNA-seq data. The underlying mathematical model for cellular growth in this model is an off-lattice, cell-center Drasdo and Höhme Monte Carlo mathematical model [14]. We develop an extension to this model that also include a distribution of cellular growth rates and cell types. We then simulate pathway activity from our new model based upon the simulated distribution of growth factor, state in the cell cycle, and cellular type. We tailor the applications of this model to cancer, by further extending the Drasdo and Höhme model [14] to also include a targeted therapeutic that represses the growth factor signaling pathway and thereby alters cellular growth rates. Finally, we couple this mathematical

model with a statistical model from [19] to simulate transcriptional data based upon simulated pathway activity and curated gene sets [20,21]. We simulate data from microarrays, RNA-seq, and single cell RNA-seq using established platform-specific error distribution models [4,19,22]. Finally, we benchmark the resulting cellular behavior and transcriptional data relative to time course *in vitro* data of four cancer cell lines treated with cetuximab. We explore the impact of intra-tumor heterogeneity of growth rates and cellular composition on the model reflective of growth curves for a head and neck cancer patient derived xenograft with and without cetuximab treatment. The package is developed generically, using S4 classes to enable users to readily add custom cellular behavior and signaling pathways in the simulation of time course transcriptional data.

Materials and Methods

CancerInSilico package and cellular growth model

The core cellular growth model in CancerInSilico is an off-lattice, stochastic cell-based model. The model inputs data from three distinct classes of parameters: (1) cellular population distributions, (2) cellular properties, and (3) cellular mechanics. Each of the parameters for these classes is described in detail in **Supplemental File**. Briefly, cellular population distributions represent the relative density of cells and experimental conditions such as cellular synchronization or presence of a boundary. Cellular properties describe the expected cell cycle length and cell size, whereas cellular mechanics encode parameters for cellular movement and cellular response to contact inhibition. The former may be freely set by the user to reflect different cellular types and therapeutic response, while the latter are fixed from parameterization in previous studies [14]. CancerInSilico is run through the core function *inSilicoCellModel* for a specified run time, and the model outputs the size, shape, and cell cycle state for each simulated cell (**Figure 1**). The implementation in the CancerInSilico model simulates cellular growth in two dimensions. The code provides an R class structure called *CellModel* that generalizes to all classes of cell-based models and can be readily extended to model three-dimensional tumor growth. Users may develop their own code based upon this framework implement alternative modeling strategies for any of these components of the model, including notably cellular division and cellular mechanics. New models encoded with this class structure can directly utilize functions from CancerInSilico to run the simulation, visualize cells, and simulate omics data from other mathematical models. In our case, we utilize this infrastructure to add cancer-specific extensions to the model in Drasdo and Höhme [14] such as therapeutic effect and multiple cellular types. All simulations presented here use the CancerInSilico Bioconductor package version 2.0. The code to reproduce these simulations is available from https://github.com/FertigLab/CancerInSilico-Figures.

Pathway simulation

The connection between the cell model and the gene expression simulation happens through user defined pathways that are then associated with gene expression changes

in a corresponding set of genes annotated to that pathway (**Figure 1**). We define an intermediate variable (*P*) that is a continuous value between zero and one that records how active each biological pathway is within each modeled cell. For this study, we model four pathways related the phase transition from G to S and G to M, growth factor signaling, and contact inhibition. For the G to M and G to S pathways, cells are either 0 or 1 at the current time point depending on if they transitioned between cell cycles within a specified time window from the current time. In this simulation, the transition to S phase occurs when the cell is half its maximum size and transition to M phase when the cell divides. Activity in the growth factor signaling pathway is assumed to be proportional to the expected cell cycle length of each cell, which differs from the true cell cycle length that occurs in the stochastic simulation. Specifically, every cell in the simulation has an inherent expected growth rate that is based on its cell type. This is an intrinsic property of each cell, and the distribution of these growth rates is an input parameter to the model. We scale the expected growth rates by assuming they are exponentially distributed with a mean length of 48 hours. This scaling ensures that simulations with different distributions of growth rates are directly comparable, for example enabling comparison of simulated omics data from treated and untreated cells, without the introduction of artificial maximum or minimum cell division times. The pathway activity for growth genes is a negative exponential function of this inherent growth rate. The contact inhibition activity is defined by the “local density” of the cell, which is the proportion of area within 50% of the minimum cell radius of the reference cell that is occupied by other cells.

Simulating gene expression data

Once pathways are simulated, gene expression data is simulated based upon annotated gene sets for each pathway (**Figure 1**) generalizing methods developed previously [19]. Unless otherwise specified, the gene set for G to S is obtained from the E2F pathways in the Hallmark and PID pathways from MSigDB [21], G to M from PID PI3KCI_AKT, TNF, TGFBR, and RB1 pathways in MSigDB, contact inhibition gene set from the KEGG Hippo pathway and targets of the TEF family transcription factors from TRANSFAC professional 2014 version [20], and growth factor signaling gene set from targets downstream obtained from TRANSFAC curated to the EGFR signaling network previously [23].

Each gene has a pre-specified expression range (*Gmin* to *Gmax*), determined either from a reference dataset or set according to a specified distribution. If a gene is annotated to only one pathway with activity *P*, then its expression value is given as $G = Gmin + P^{*}(Gmax - Gmin)$. The effect of cell type is also modeled as a pathway whose activity is binary for each cell, with a gene set specific to that cell type. If a gene is regulated by multiple pathways, its expression is determined by combining all pathways using a user-defined function. By default, gene expression is set to be the maximum expression of that gene across all pathways. Additional genes unaffected by any pathway are also simulated with a pre-specified gene expression value. In bulk data, *P* is determined by computing the average value for pathway activity in a random set of *N* sampled cells, whereas in single cell data the value of *P* for each of the *N* sampled cells

is used directly. Technical error is then simulated in all genes appropriate to the measurement platform. A normal error model is used to simulate log transformed microarray data and negative binomial error model adapted from the code for LIMMA voom [22] is used to simulate bulk RNA-sequencing data. Technical noise for simulated single cell RNA-sequencing data are generated using the error and drop out models from Splatter [4]. T-SNE plots are generated with the R package Rtsne.

Cell proliferation and microarray gene expression analysis of BxPC-3 cells

For cell proliferation, BxPC-3 cells were seeded at a density of 6000 cells per well in 96-well plates and treated the next day with varying concentration of cetuximab (0.01, 0.1, 1, 10, and 100 µg/ml). Cell viability was measured at indicated time points using GF-AFC substrate (Promega, Madison, WI) according to the manufacturer's protocol.

For gene expression analysis, BxPC-3 cells were cultured in DMEM with 2% FBS and seeded at a density of 10^5 cells per well in 12-well plates after starving (DMEM with 0.2% FBS). 10µg/ml cetuximab was added the next day. Cell samples were taken out at indicated time points and stored in RNAlater (Qiagen, Germantown, MD). Total RNA was isolated from cells pellets and extracted using the RNeasy plus Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. After the extraction, RNA concentration and purity (OD260/280) were measured using the NanoDrop ND-1000 spectrophotometer (Thermo Fischer Scientific, Waltham, MA), and the RNA integrity number (RIN) was determined using an Agilent 2100 Bioanalyzer Instrument (Agilent, Santa Clara, CA). Gene expression profiling was performed using the Illumina HumanHT-12 v4 Expression BeadChip platform containing 47,000 probes covering RefSeq and Unigene annotated genes (Illumina Inc., San Diego, CA). 495ng total RNA was labeled using the Illumina TotalPrep-96 RNA Amplification Kit (Ambion, Austin, TX). Briefly, the protocol features a first- and second-strand reverse transcription step, followed by a single in vitro transcription (IVT) amplification that incorporates biotin-labeled nucleotides to generate biotinylated cRNA. The purified cRNA is then quantified and the fragment size was determined using an Agilent 2100 Bioanalyzer Instrument (Agilent, Santa Clara, CA). Then, 750ng of labeled biotinylated cRNA probe was hybridized overnight to the Illumina HumanHT-12 v4 Expression BeadChip. The hybridization, washing, and scanning were performed according to the manufacturer's instructions.

The BeadChips were scanned using a HiScanSQ System (Illumina, San Diego, CA). The microarray images were registered and extracted automatically during the scan using the manufacturer's default settings. Data are cyclic loess normalized with LIMMA [24] and available from GEO (GSE114375). Data for probes annotated to a single gene and technical replicate samples are averaged to obtain a single measurement per gene and experimental condition for visualization in heatmaps with the R package gplots.

Cell proliferation and RNA-seq gene expression analysis of HNSCC cell lines SCC25, UM-SCC-1 and UM-SCC-6

An additional set of experiments were performed to measure cetuximab sensitivity and gene expression with RNA-seq on a set of head and neck squamous cell carcinoma (HNSCC) cell lines containing SCC25, UM-SCC-1, and UM-SCC-6. Cell line authenticity was confirmed using the GenePrint 10 (Promega) for short tandem repeat (STR) profile. SCC25 and UM-SCC-1 cells were grown in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS and 0.4µg/mL (SCC25) or 1.0µg/mL hydrocortisone. UM-SCC-6 cells were cultured in DMEM high glucose supplemented with 10% and 1X NEAA. All cells were maintained at 37^{o}C in a humidified incubator with 5% CO_2.

For cetuximab treatment and proliferation assay, 12,500 cells (SCC25) or 25,000 (UM-SCC-1 and UM-SCC-6) were plated in quintuplicates in 6-well plates. Cetuximab (Bristol-Myers Squibb, Princeton, NJ) was purchased from Johns Hopkins Pharmacy. All cell lines were treated with cetuximab (100nM) or PBS (untreated control) for 5 days starting 24 hours after seeding. Media containing cetuximab or PBS was changed every day. Proliferation was measured using alamarBlue Cell Viability assay (Thermo Scientific), as described by the manufacturer. Briefly, 10% sample volume of alamarBlue reagent was added to each well and fluorescence (excitation 544nm, emission 590nm) was measured after 4 hours of incubation at 37^{o}C. A media only well was used as blank control.

RNA isolation and sequencing were performed for the SCC25, UM-SCC-1 and UM-SCC-6 cells for each day under cetuximab or PBS treatment at the Johns Hopkins Medical Institutions (JHMI) Deep Sequencing & Microarray Core Facility. Total RNA was isolated from at least 1000 cells using the TRIzol Reagent (Invitrogen) according to manufacturer's instructions. The RNA concentration was determined by the spectrophotometer Nanodrop (Thermo Fisher Scientific, Waltham, MA) and quality was assessed using the 2100 Bioanalyzer (Agilent, Santa Clara, CA) system. An RNA Integrity Number (RIN) of 7.0 was considered as the minimum to be used in the subsequent steps for RNA-seq. Library preparation was performed using the TrueSeq Stranded Total RNAseq Poly A1 Gold Kit (Illumina, San Diego, CA), according to manufacturer's recommendations, followed by mRNA enrichment using poly(A) enrichment for ribosomal RNA (rRNA) removal. Sequencing was performed using the HiSeq platform (Illumina) for 2X100bp sequencing. Reads were aligned to hg19 with salmon [25]. Gene-level counts were calculated with the R/Bioconductor package tximport [26] and rlog normalized with DESeq2 [27] for visualization. All RNA-seq data from this study are available from GEO (GSE114375).

HNSCC Patient–Derived Xenograft (PDX)

Tumor tissues measuring approximately 2x2 mm were collected from surgically resected HNSCC patients under the auspices of a tissue bank protocol approved by Johns Hopkins University Institutional Review Board. All animal studies and care were approved by the Institutional Animal Care and Use Committee of the Johns Hopkins University. Following HNSCC tumor resection, de-identified patient samples were implanted into athymic nude mice (F1) and passaged to F2 mice at the Johns Hopkins

University Tissue Core (IACUC #: MO011M210). Dr. Chung's laboratory was given excised tumors from F2 mice to propagate for treatment experiments (F3 mice) under the IACUC protocol MO13M114. Briefly, athymic nude mice (Crl: NU-Foxn1nu, 4–6 weeks old; 20 g; Harlan Laboratories, Indianapolis, IN) were anesthetized using 5% isofluorane and a small 2mm incision was made in the flank. Tumor tissue measuring approximately 2x2 mm was implanted under the skin and the wound closed with sterile surgical clips and analgesic was administered. Animals were monitored for tumor growth twice a week and tumor volume calculated using the formula (length x width^2)/2. Tumors were allowed to reach a volume of 200 mm^3 before starting treatment with therapeutic agents. Animals were euthanized when tumor volume reached 1500 mm^3 or they had lost >25% of their pre-implant body weight. Euthanasia was performed by CO_2 inhalation and tumors were collected for further analysis.

Results

Parameter sensitivity analysis demonstrates that untreated cells from a single cell type are expectedly most sensitive to presence of a boundary, cellular density, and cell cycle length

We first perform parameter sensitivity analysis of our default model, which uses core parameters of the Drasdo and Höhme model [14] and models a single cell without treatment. This baseline simulation enables us to benchmark the performance of the model under conditions in which the population dynamics are readily interpretable and predictable. We perform a complete sensitivity analysis for all input parameters (described in the **Supplemental File**). The presence of a boundary, density, and cell cycle length are the set observed to have the greatest impact on qualitative model performance (**Figure 2**).

We implement the Drasdo and Höhme model [14] to simulate cancer cell growth in conditions with and without a boundary to model constrained *in vitro* and unconstrained *in vivo* growth, respectively. Presence of a boundary models the difference between whole population growth in a dish and local population growth within a region that enables unconstrained growth. Without a boundary, cells initially grow exponentially and then growth rates decrease below exponential due to contact inhibition. In contrast, cellular growth in the presence of a boundary is logistic (**Figure 2a**). In both cases, the effective growth rate of the number of cells is inversely proportional to the expected cell cycle length parameter. The total number of cells is also inversely proportional to the expected cell length parameter in cases without a boundary. In the case with the boundary, decreasing the expected cell cycle length increases the rate at which the maximum number of cells is reached. Total cellular density is capped in simulations with and without a boundary (**Figure 2b**). The maximum density is lower in simulations without a boundary than those that have one. In the unbounded case, the model has a higher probability of cellular expansion to low density regions than dividing cells within high density regions thereby controlling cellular density. The bounded simulations do not have that option and so they must continue to divide and pack in a denser formation.

The cell-based model also includes mechanical parameters that have been tuned to optimize model performance previously [14]. These parameters impact the technical architecture of the software and are not directly measurable experimentally. We perform additional sensitivity analysis for these parameters to assess their impact on model behaviors. We observe that changes to the technical parameters produce artificial effects in the model that mirror sensitivity to interpretable, cellular parameters. For example, changing the Monte Carlo probability thresholds decreases cellular growth in a similar manner to increasing cell cycle length (**Supplemental Figure 1**). We note that changing the physical, measurable parameters can be performed directly from experimental data with similar outcomes. Therefore, it is optimal to alter cellular growth parameters directly and retain well-tuned technical parameters for simulation in the software.

Modeling growth inhibition by targeted therapeutics

We extend the model to incorporate targeted therapeutics that block oncogenic pathways. We simulate their effect by assuming that they decrease cellular growth rates in cancer cells, which has been reported as the mechanism of action of FDA approved growth factor receptor inhibitors such as the epidermal growth factor receptor (EGFR) inhibitor, cetuximab. These therapeutic effects are encoded in our model by increasing the length of cell cycle when the drug is present to simulate the delay in proliferation that they cause. The code enables users to input a distribution of impacts on cell cycle lengths in the cellular population. This distribution enables therapeutics to impact the growth rate of each cell differently. Drugs may also be introduced to the population at a specified simulation time to simulate cancer cellular growth prior to treatment. Because we model therapeutic efficacy through cell cycle length, a simulation that treats cancer cells from the initial time point is equivalent to a simulation of untreated cells with an altered initial expected cell cycle length in our model.

To test our model, we compare growth rates of simulated cells with a boundary to *in vitro* growth rates observed in two different experiments. Biological parameters from *CancerInSilico* are tuned to minimize the sum of squares error of the measured growth curve relative to the simulated growth curve in the untreated case. In the treated case, only the expected cell cycle length was changed, keeping all other parameters constant from the untreated simulations. We observe that increasing the length of the cell cycle as a surrogate for therapeutic concentration accurately models the growth rates that are observed in the *in vitro* experiments. The growth curves from the BxPC-3 cell line were fit for controls and treatment with 10 µg/ml and 100 µg/ml of cetuximab using the same parameters and tuning only the expected cell cycle length to simulate the impact of treatment (**Figure 3a)**. In the second experiment, the model parameters were tuned to fit the growth curves of the SCC1, SCC6, and SCC25 cell lines in PBS controls (**Figure 3b**) and treatment with 100nM of cetuximab was fit by increasing only the expected cell cycle length (**Figure 3c)**.

Modeling pathway interactions enables simulated time-course gene expression data to mirror real gene expression data

CancerInSilico uses the cellular states in the mathematical model as the basis to simulate gene expression data (**Figure 4**). Specifically, the model assumes that there are pathways that are activated at each time point corresponding to: (1) the percentage of cells in G to S, (2) the percentage that have successfully completed division, (3) the local density of cells ("contact inhibition"), and (4) expected length of the cell cycle ("activation growth factor pathway"). The modeled effect of drug changes only the expected length of the cell cycle, yet this change will also impact the state of the cell cycle indirectly due to interdependencies of model parameters (**Figure 1**). Pathway activity is modeled as a value between zero and one. The pathway is zero if it is not active in any cell and one if it is active in all cells, with intermediate values depending on the pathway (**Supplemental Figure 2**). Pathway values are used to simulate gene expression in target genes from bulk data with a generalization of previously described methods [21]. We apply this model to simulate gene expression data from the fitted simulations described in the previous section (**Figure 3**). For the BxPC-3 cell line, microarray data was simulated for the PBS and 100 µg/ml experiments (**Supplemental Figure 3**). For the SCC1, SCC6, and SCC25 cell lines, bulk RNA-seq data was generated for both the treated and untreated experiments (**Supplemental Figures 4-6)**. The default transcriptional data simulated with *CancerInSilico* has a sharper separation of cellular signaling pathways than the real gene expression data (**Figure 4a,b, Supplemental Figures 5,6**), specifically in the G to S pathway. We note that the BxPC-3 cell line experiments are performed with microarray and synchronized, whereas the SCC1, SCC6, and SCC25 cell lines were performed with RNA-sequencing and are not synchronized. The default, simulated gene expression data appears to have more cell synchronization consistent with the BxPC-3 data than in the real data for the other cell lines.

The discrepancy between real and simulated data likely arises from one of three sources: (1) incomplete annotations of gene sets to pathway activity, (2) insufficient variation in the cell cycle length of cells in the modeled population, or (3) incomplete modeling of the interactions between the modeled pathways. Assessing the accuracy of gene sets is beyond the scope of this study. The impact of both (2) and (3) on the qualitative behavior of simulated gene expression data can be explored computationally. To minimize the effect of cell synchronization, an exponential distribution was introduced for the expected cell cycle length. This provided additional variance between the individual cell cycles and the resulting gene expression data was smoother (**Supplemental Figures 3-6**)**.** This increase in smoothness better models the qualitative gene expression changes in the unsynchronized cell lines than the synchronized BxPC-3 cell line.

Interactions between pathways will further have a significant impact on the simulated data. Notably, changes to the growth factor receptor signaling pathway are expected to impact cellular growth. In this case, the genes in the gene set annotated to the growth factor receptor signaling would also be associated with pathway activity in the G to S

and G to M transitions. To simulate this interaction, we assign the genes annotated to the growth factor to G to S and also to G to M (**Figure 4c,d, Supplemenatl Figures 3-6**). In these simulations, the increase in variation of cell cycle length was simulated only for the SCC1, SCC6, and SCC25 cell lines to minimize cell cycle synchronization consistent with the experimental design. Modeling such pathway overlap reduced the effect of the therapeutic on transcription and the separation of the cell signaling pathways so that the simulated data more closely reflects the qualitative behavior of the real expression data. Together, these results suggest that both heterogeneity in the cellular population and gene set overlap may be critical parameters to vary when testing the performance of bioinformatics algorithms based upon the simulated data. This model more accurately simulates the dynamics of genes that are activated than those that are repressed, suggesting that alterative models will be essential to benchmark the performance of algorithms to model repressive processes.

Heterogeneity in therapeutic response arises from simulating multiple cell types

Tumors and cancer cell lines are mixtures of distinct cell types. In the case of tumors, this may include diverse sets of cells with diverse growth properties that will impact tumor growth. *CancerInSilico* enables us to test the impact of cellular heterogeneity as modeled by heterogeneity in the cellular properties of a cell type. We recall that each cell in our simulations has an expected growth rate. *CancerInSilico* allows users to sample the values for these growth rates from a distribution. Increasing the variance of expected growth rates provides one means to model increased intra-tumor heterogeneity. However, a set of 200 simulations with mean expected cell cycle length of 24 hours and standard deviation of 1 hour has only modestly lower variation of simulated cellular behavior than does a comparable set of simulations with standard deviation of expected cell cycle length of 10 hours (**Figure 5a,b**).

Another mechanism that *CancerInSilico* can use to model heterogeneity is to label each cell as being from a distinct cell type. The model encodes a set of parameter distributions that is specific to each modelled cell type. We apply this framework to model two distinct cell types, one with an expected cell cycle length of 12 hours and standard deviation of 4 hours (type A) and one with a mean cell cycle length of 36 hours and standard deviation of 4 hours (type B). We observe that varying the initial proportion of cells of each type across simulations has a greater impact on heterogeneity of simulated cell behavior than is observed by varying expected cell cycle length alone (**Figure 5c,d**). Qualitatively, we observe similar variation in growth curves of a head and neck squamous cell carcinoma patient derived xenograft model, suggesting that replicates of this biological model were seeded with different initial subsets of cell types from the heterogeneous primary tumor (**Figure 5f**).

CancerInSilico also encodes an option to simulate single cell RNA-sequencing data to generate omics data that reflects the heterogeneity of the sample population. This simulation models the “pathway” activity and corresponding gene expression changes for each cell with a negative binomial error model and dropout model adapted from Splatter [4]. The model then randomly samples a pre-specified number of cells. We

apply this technique to simulate single cell RNA-sequencing data from the simulation with a population of cells equally distributed between the two types described above (**Figure 6**). Each cell type is labeled as a pathway with binary values for activity to activate a gene set that corresponds to cellular identity. Because these cell types are computational, we define this gene signature by splitting the set of genes in the growth factor receptor pathway equally between the two cell types. We note that these genes are only used as identifiers to define a signature and are not associated with a particular functional cell type. Therefore, analysis of these simulated data can serve to benchmark the ability to separate cellular identity and not specific gene function. In this simulated single-cell RNA-seq data, we observe strong separation between cell types (**Figure 6a**) and time (**Figure 6b**) and observe a mixture between cell cycle phases (**Figure 6c**).

Discussion

We develop a new R/Bioconductor package *CancerInSilico* that extends the mathematical model from Drasdo and Höhme [14] to simulate treatment with growth factor receptor inhibitors, cellular-level variation in model parameters, and multiple cell types. We demonstrate that our new model mirrors real *in vitro* cancer cells with and without cetuximab treatment and *in vivo* growth of a patient derived xenograft tumor model. Coupling this mathematical model with statistical models of technical noise and expression changes in gene sets annotated to cell signaling pathways [4,19,22] enables simulation of *in vitro* time course omics data. The modeling of individual cells in this system also enables simulation of time course, single cell RNA-sequencing data. The outputs of both bulk and single-cell transcriptional data can readily be input into simulation tools such as Polyester [5] or flux simulator [6] to model reads from high-throughput sequencing data. Thus, this package provides software that can generate simulated time course omics data that can be used to benchmark the performance of methods for time course bioinformatics analysis from a known ground truth that is lacking in real data.

Currently, *CancerInSilico*, has implemented an off-lattice cell-based model based on previous work [14]. Simulating single-cell gene expression requires cell-level information and a cell-based model is the most natural way to achieve this. An off-lattice model provides detailed geometric information about cell size and shape, which is important for pathways that are affected by the geometry of the cell population, i.e. contact inhibition. Tracing each individual cell also readily enables simulation of a distribution of model parameters for each cell and multiple cell types. Modeling cellular growth as a distribution, rather than a single parameter, enabled accurate modeling of the transcriptional changes in cell cycle pathways observed in real *in vitro* RNA-seq gene expression data from non-synchronized cells. However, introducing further variation by modeling multiple cell types was essential to model the wide range of growth curves from multiple replicates of the same patient derived xenograft model. This modeling result demonstrates that the initial population of tumor cells may have a critical impact on long-term *in vivo* behavior, motivating study with single cell rather than bulk technologies.

The flexibility of the off-lattice model comes at the price of computation time. Efficient algorithms and data structures work more naturally with lattice-based models of tumor growth [12]. A lattice-based model will significantly alter the geometry of the cell population and thereby significantly alter the simulated omics data. It is essential to compare these differences relative to real *in vitro* and *in vivo* data in future work. Moreover, the R/Biconductor framework for *CancerInSilico* is designed to accommodate multiple types of cell models including these lattice-based models. In the future, this will allow us to analyze how sensitive simulated gene expression data are to the choice of mathematical model. Moreover, such future work will also enable simulations of gene expression data based upon the cellular dynamics simulated from the model that is most appropriate a specific biological system.

The statistical models used to simulate gene expression data from pathway activity in *CancerInSilico* mirror the process by which omics tools estimate that activity. Namely, pathways are assumed to activate discrete sets of genes annotated to a common function based upon the modeled cellular state. Default parameters for the model yield omics profiles with strong separation between signaling pathways that are greatly simplified relative to those observed in real data. Therefore, applying omics algorithms to these default simulations may result in under estimation of the accuracy of their performance for real time course data. Mathematical models of gene regulatory networks have been developed to model the dynamics of regulatory networks that lead to transcriptional changes [28,29]. Hybrid, multi-scale approaches that combine these network-based models with the cellular-scale models more accurately model the complexity of system-wide dynamics [16–18] and are a promising area for future work to simulate time course omics data. However, the complexity of these gene regulatory models and extensive parameterization will limit the straightforward validation of omics algorithms that is possible from the simplified statistical models employed in *CancerInSilico*. Moreover, we find that including interactions between cell signaling pathways by merely overlapping gene sets and increasing modeled technical noise enables the statistical model in *CancerInSilico* to accurately simulate the qualitative behavior of time course gene expression from *in vitro* microarray and RNA-seq data with and without cetuximab treatment. Thus, we recommend benchmarking time-course omics data on simulated data generated from a wide range of these parameter values to fully assess their performance. In the RNA-seq data, the model more accurately the qualitative transcriptional dynamics of genes which are activated over time than those that are repressed. Future work is needed to develop additional statistical models to model these repressive processes, which are critical to the function of real biological systems.

The primary goal of this study was to benchmark the performance of *CancerInSilico* as a simulation system rather employ simulations to discover novel biology from this model. The ability to model real omics data only by incorporating overlap between the growth factor receptor and cellular growth pathways is consistent with complex cross-talk between cellular signaling pathways in biological systems. For example, our own data have shown that complex feedback mechanisms activate the growth factor receptors after their inhibition with cetuximab [30]. Taken together, these observations

demonstrate the challenges in annotating gene sets to a single and independent biological function. Still, each of these biological processes occurs at distinct temporal scales. Therefore, increasing the number of time points available for omics data will enhance our ability to resolve these processes as we observed in time course data from cetuximab resistance [31]. Tools to assess optimal time points for such measurements, such as the Time Point Selection (TPS) method [32], are critical to enable optimal analysis. Once these data have been collected, accurate delineation of molecular processes for precision medicine will require bioinformatics techniques that are able to account for usage of genes in multiple biological processes.

Acknowledgements

We thank Ludmila V Danilova, Alexander V Favorov, Emily Flam, Dylan Kelley, Feilim Mac Gabhann, Cristian Tomasetti, and members of NewPISlack for critical comments and feedback on this project.

Funding

This work was supported by NIH Grants CA177669, CA006973, CA212007, CA50633, CA51008, DE017982, and SPORE DE019032. This work was also supported by The Cleveland Foundation Helen Masenhimer Fellowship, Johns Hopkins University Catalyst and Discovery Grants, and Johns Hopkins School of Medicine Synergy Award. This project has been made possible in part by grant number 2018-183444 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

References

1. Liang Y, Kelemen A. Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. Brief. Bioinform. 2017;
2. Liang Y, Kelemen A. Computational dynamic approaches for temporal omics data with applications to systems medicine. BioData Min. 2017; 10:
3. Hafner M, Niepel M, Chung M, et al. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. Nat. Methods 2016; 13:521–527
4. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017; 18:
5. Frazee AC, Jaffe AE, Langmead B, et al. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics 2015; 31:2778–2784
6. Griebel T, Zacher B, Ribeca P, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012; 40:10073–10083
7. Barish S, Ochs MF, Sontag ED, et al. Evaluating optimal therapy robustness by virtual expansion of a sample population, with a case study in cancer immunotherapy. Proc. Natl. Acad. Sci. 2017; 114:E6277–E6286
8. Tran PT, Bendapudi PK, Lin HJ, et al. Survival and Death Signals Can Predict Tumor Response to Therapy After Oncogene Inactivation. Sci. Transl. Med. 2011; 3:103ra99-103ra99
9. Chmielecki J, Foo J, Oxnard GR, et al. Optimization of Dosing for EGFR-Mutant Non-Small Cell Lung Cancer with Evolutionary Cancer Modeling. Sci. Transl. Med. 2011; 3:90ra59-90ra59
10. Picco N, Sahai E, Maini PK, et al. Integrating Models to Quantify Environment-Mediated Drug Resistance. Cancer Res. 2017; 77:5409–5418
11. Rockne R, Alvord EC, Rockhill JK, et al. A mathematical model for brain tumor response to radiation therapy. J. Math. Biol. 2009; 58:561–578
12. Szabó A, Merks RMH. Cellular Potts Modeling of Tumor Growth, Tumor Invasion, and Tumor Evolution. Front. Oncol. 2013; 3:
13. Ghaffarizadeh A, Heiland R, Friedman SH, et al. PhysiCell: an Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems. 2017;
14. Drasdo D, Höhme S. Individual-based approaches to birth and death in avascular tumors. Math. Comput. Model. 2003; 37:1163–1175
15. Gallaher J, Anderson A. The role of contact inhibition in intratumoral heterogeneity: An off-lattice individual based model. 2016;
16. Rejniak KA, Anderson ARA. Hybrid models of tumor growth. Wiley Interdiscip. Rev. Syst. Biol. Med. 2011; 3:115–125
17. Clancy CE, An G, Cannon WR, et al. Multiscale Modeling in the Clinic: Drug Design and Development. Ann. Biomed. Eng. 2016; 44:2591–2610
18. Yankeelov TE, An G, Saut O, et al. Multi-scale Modeling in Clinical Oncology: Opportunities and Barriers to Success. Ann. Biomed. Eng. 2016; 44:2626–2641
19. Fertig EJ, Favorov AV, Ochs MF. Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobioscience 2013; 12:142–149

20. Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003; 31:374–378
21. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 2005; 102:15545–15550
22. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15:R29
23. Fertig EJ, Ren Q, Cheng H, et al. Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. BMC Genomics 2012; 13:160
24. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47–e47
25. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 2017; 14:417–419
26. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 2015; 4:1521
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:
28. Chasman D, Fotuhi Siahpirani A, Roy S. Network-based approaches for analysis of complex biological systems. Curr. Opin. Biotechnol. 2016; 39:157–166
29. Trairatphisan P, Mizera A, Pang J, et al. Recent development and biomedical applications of probabilistic Boolean networks. Cell Commun. Signal. 2013; 11:46
30. Fertig EJ, Ozawa H, Thakar M, et al. CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network. Oncotarget 2016; 5:
31. Stein-O'Brien G, Kagohara LT, Li S, et al. Integrated time course omics analysis distinguishes immediate therapeutic response from acquired resistance. 2018;
32. Kleyman M, Sefer E, Nicola T, et al. Selecting the most appropriate time points to profile in high-throughput studies. eLife 2017; 6:

List of Figures

Figure 1. Overview of gene expression simulation with CancerInSilico. The model inputs a set of parameters that represent the distribution of the initial cellular population, behavior of each cell, and mechanical parameters for the off-lattice, cell based model. These parameters are then used as the basis to simulate cellular growth as a function of time. A statistical model is used to estimate pathway activity from the properties of each cell and geometric distribution of cells at each time point from the mathematical model. Finally, these pathway values are used as the basis to simulate single cell (left) and bulk (right) gene expression data for members of gene sets annotated to each simulated pathway.

Figure 2. Model sensitivity to expected cell cycle length and initial cellular density. (a) Number of cells as a function time simulated for varying expected cell cycle length in simulations with and without a boundary as indicated in the figure legend. (b) Population density as a function of time simulated for varying initial cellular density in simulations with and without a boundary as indicated in the figure legend.

Figure 3. Comparison of simulated and *in vitro* cellular growth with and without cetuximab treatment. (a) Simulated growth curves (lines) and *in vitro* growth (dots) for BxPC-3 cells treated with 10 µg/ml or 100 µg/ml of cetuximab and PBS controls. (b) Simulated growth curves (lines) and *in vitro* growth (dots) for three head and neck cancer cell lines, SCC1, SCC25, and SCC6 in PBS controls. (c) As for (b) in cells treated with with 100 nM of cetuximab.

Figure 4. Comparison of simulated and *in vitro* gene expression for cetuximab treated cells and PBS controls. (a) Heatmap of microarry gene expression data for BxPC-3 cells in 100 µg/ml cetuximab and untreated controls treated daily for a period of 5 days. Cells were synchronized 24 hours prior to starting treatment at time 0. (b) Heatmap of RNA-seq gene expression data for SCC1 cells in 100 nM cetuximab and untreated controls treated daily for a period of 5 days without synchronization. (c) Simulated gene expression data for BxPC-3. In this simulation, cells annotated to the growth factor receptor pathway are also assigned to the G to M and G to S pathways to model pathway interactions. (d) As for (c) for SCC1 cells. In this simulation, an exponential model is used to introduce variation in the expected cell cycle length to model the lack synchronization of cells *in vitro*.

Figure 5. Comparison of simulated cell growth with multiple cell types and *in vivo* tumor growth of a patient derived xenograft model. (a) Cellular growth over time in a set of 200 simulations containing a single cell type with expected cell cycle length of 24 hours and standard deviation of 1 hour. (b) As for (a) with a standard deviation of 10 hours. (c) Cellular growth over time in a simulation containing a two cell types, called A and B. Cell type A has an expected cell cycle length of 12 hours and standard deviation of 4 hours and cell type B has an expected cell cycle length of 36 hours and standard deviation of 4 hours. Plots represent 40 simulations with initial proportion of cells of type A selected from a uniform between 0.45 and 0.55. (d) As for (c) with initial proportion of

cells of type A selected from a uniform between 0 and 1. (e) Proportion of cells of type A in the initial population vs proportional of cells of type A in the final population at 168 hours. (f) Tumor volume of triplicates for a single patient derived xenograft model.

Figure 6. T-SNE of simulated time course single cell RNA-sequencing data for a population with cells of types A and B. Points colored by (a) cell type, (b) time, and (c) cell cycle phase.

Figure 1

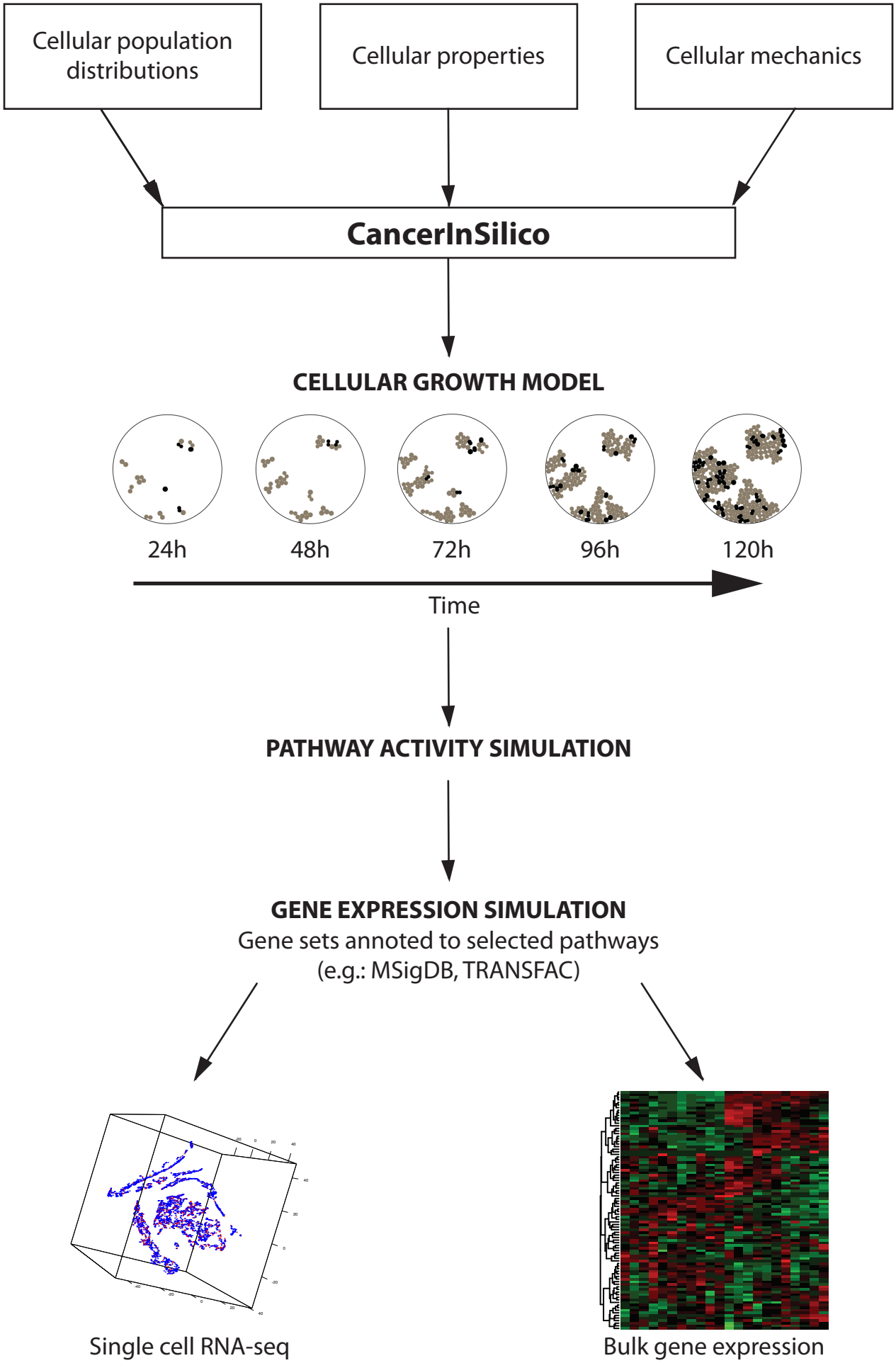


Figure 2

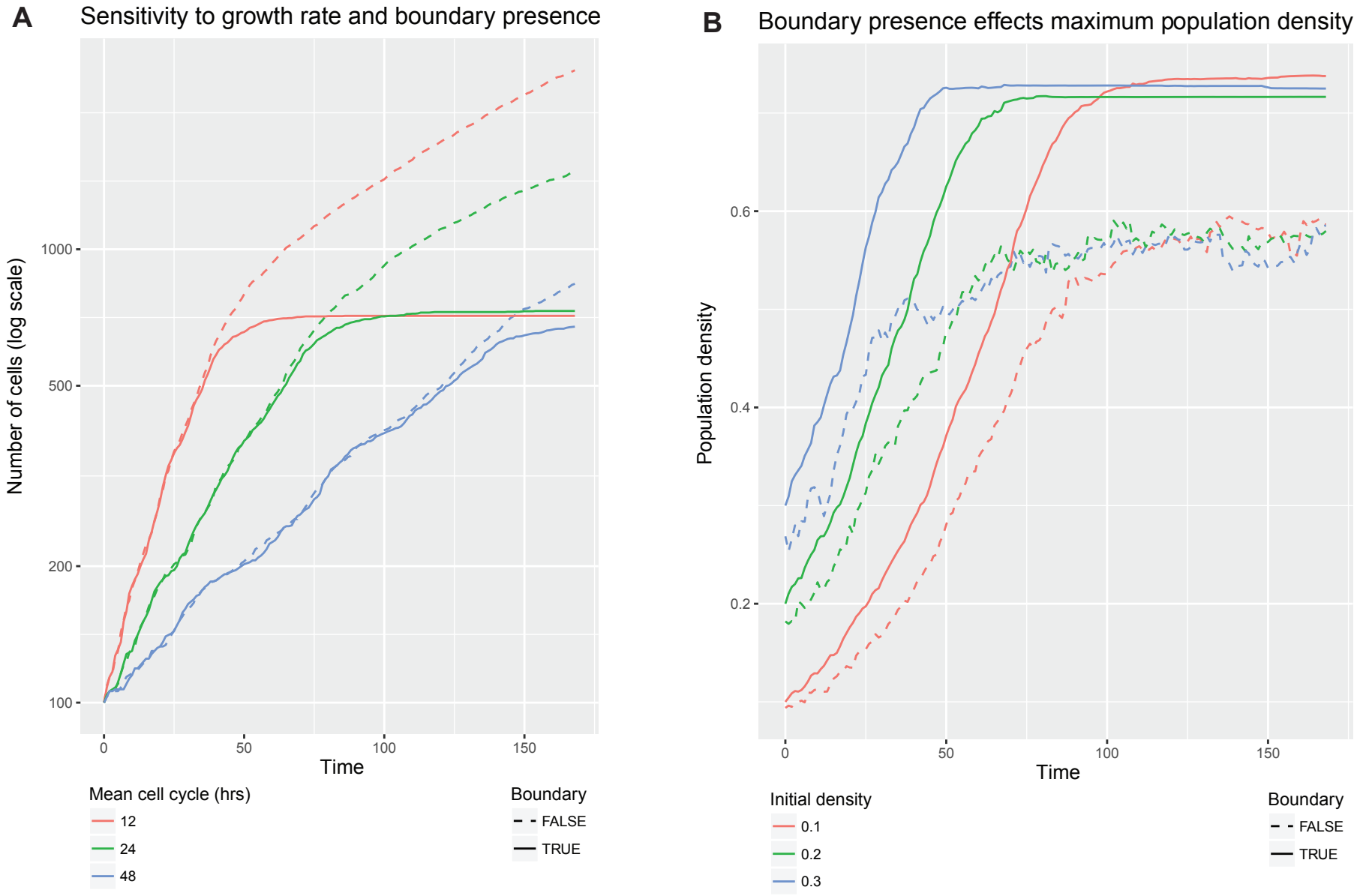
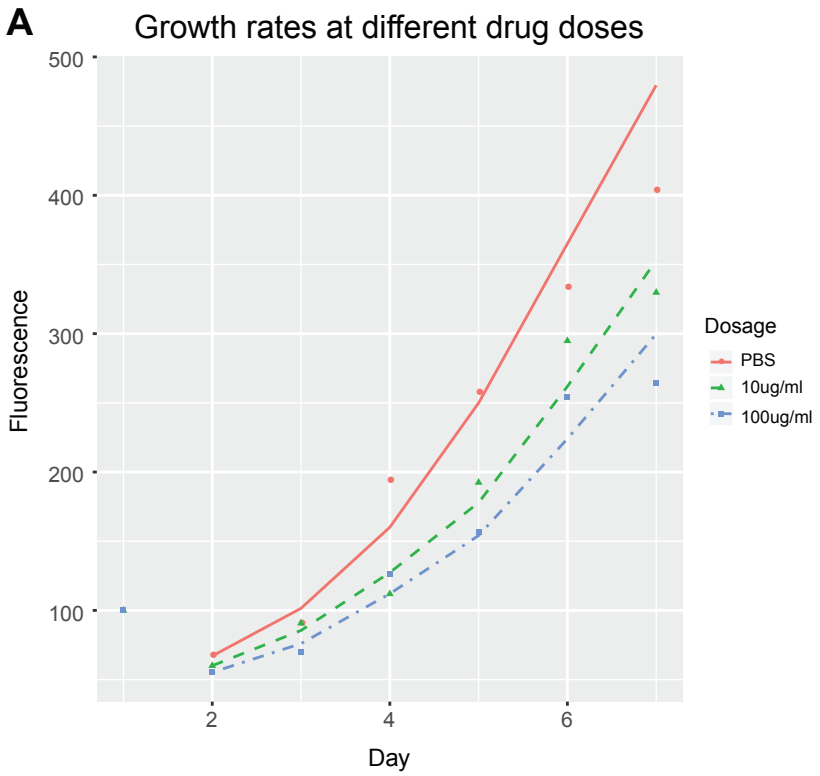
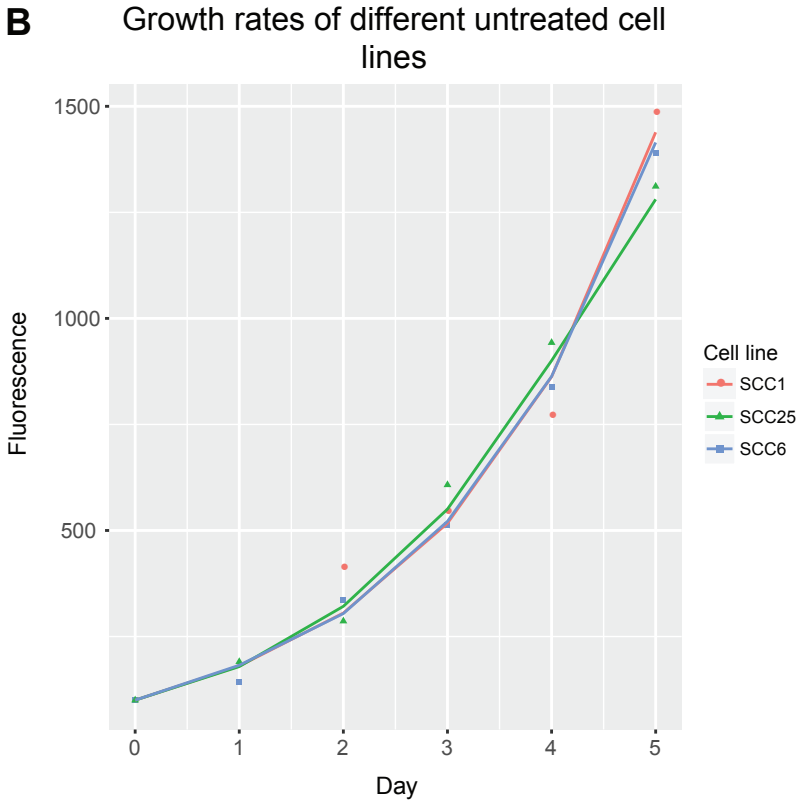

A
Sensitivity to growth rate and boundary presence
Number of cells (log scale)
1000
500
200
100
0
50
100
150
Time
Mean cell cycle (hrs)
12
24
48
Boundary
FALSE
TRUE
B
Boundary presence effects maximum population density
Population density
0.6
0.4
0.2
0
50
100
150
Time
Initial density
0.1
0.2
0.3
Boundary
FALSE
TRUE

Figure 3


A
Growth rates at different drug doses
Fluorescence
500
400
300
200
100
2
4
6
Day
Dosage
PBS
10ug/ml
100ug/ml


B
Growth rates of different untreated cell lines
Fluorescence
1500
1000
500
0
1
2
3
4
5
Day
Cell line
SCC1
SCC25
SCC6

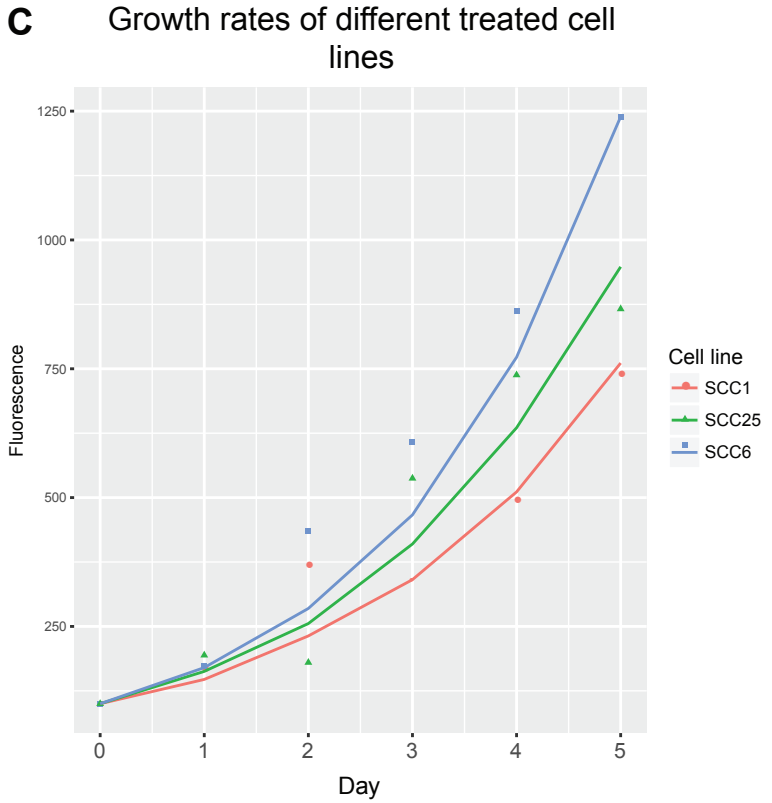

C
Growth rates of different treated cell lines
Fluorescence
1250
1000
750
500
250
0
1
2
3
4
5
Day
Cell line
SCC1
SCC25
SCC6

Figure 4

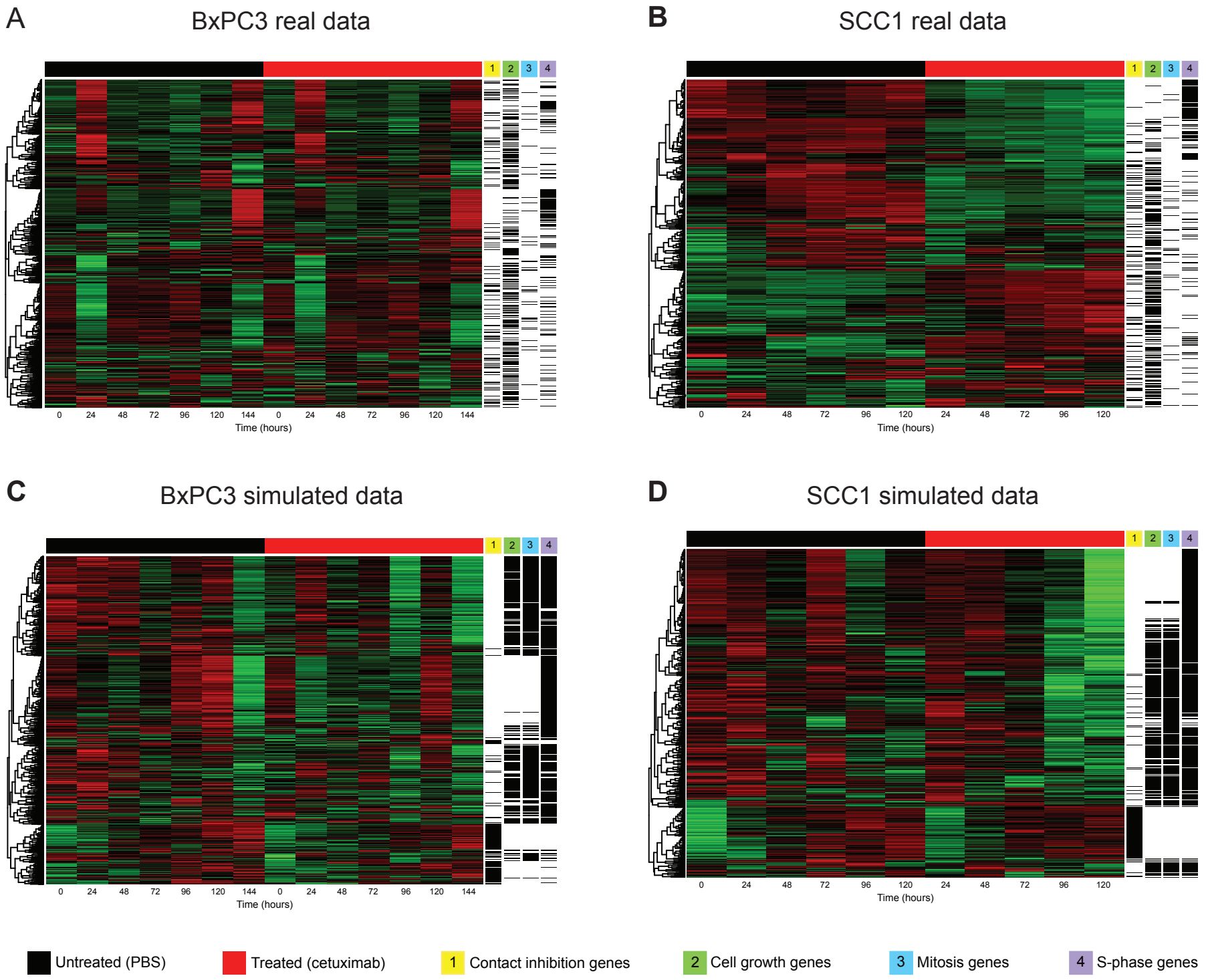

A
BxPC3 real data
B
SCC1 real data
C
BxPC3 simulated data
D
SCC1 simulated data
1
2
3
4
0 24 48 72 96 120 144 0 24 48 72 96 120 144
Time (hours)
0 24 48 72 96 120 24 48 72 96 120
Time (hours)
Untreated (PBS)
Treated (cetuximab)
1 Contact inhibition genes
2 Cell growth genes
3 Mitosis genes
4 S-phase genes

Figure 5

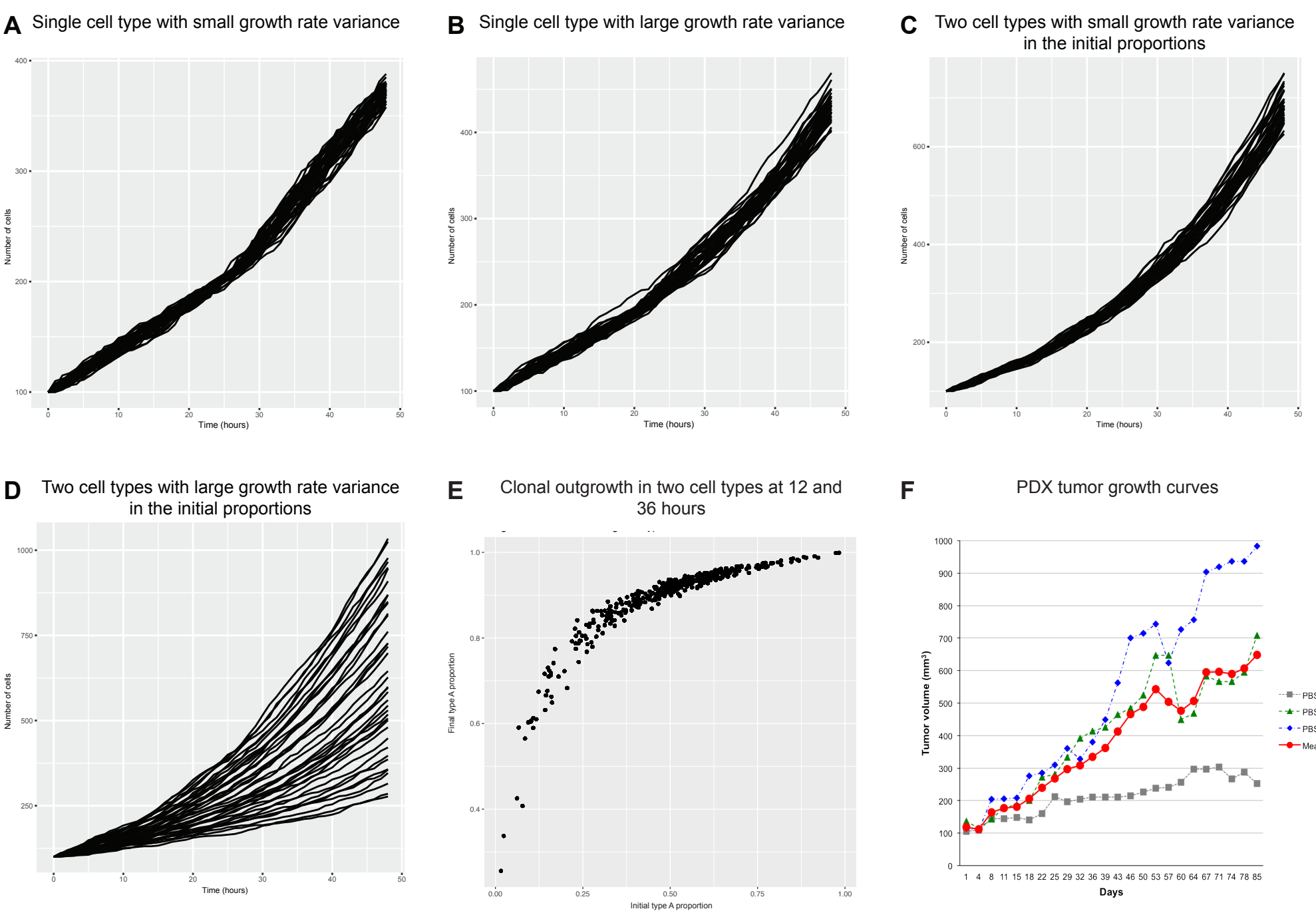

A
Single cell type with small growth rate variance
Number of cells
Time (hours)
B
Single cell type with large growth rate variance
Number of cells
Time (hours)
C
Two cell types with small growth rate variance in the initial proportions
Number of cells
Time (hours)
D
Two cell types with large growth rate variance in the initial proportions
Number of cells
Time (hours)
E
Clonal outgrowth in two cell types at 12 and 36 hours
Final type A proportion
Initial type A proportion
F
PDX tumor growth curves
Tumor volume (mm³)
Days
PBS1
PBS2
PBS3
Mean

Figure 6