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

This manuscript was automatically generated on June 19, 2021.

Authors

- **Farnaz Mohammadi**
  ORCID 0000-0002-0197-4670 · Github farnazmdi
  Department of Bioengineering, University of California, Los Angeles

- **Shakthi Visagan**
  ORCID 0000-0001-9295-2188 · Github shak360
  Department of Bioengineering, University of California, Los Angeles

- **Sean M. Gross**
  ORCID 0000-0002-9621-8551
  Department of Biomedical Engineering, Oregon Health and Science University, Portland

- **Luka Karginov**
  ORCID 0000-0002-2455-1558 · Github lkargi
  Department of Bioengineering, University of Illinois, Urbana Champaign

- **JC Lagarde**
  ORCID 0000-0002-3738-3119 · Github jclagarde
  Department of Bioengineering, University of California, Los Angeles

- **Laura M. Heiser**

  Department of Biomedical Engineering, Oregon Health and Science University, Portland

- **Aaron S. Meyer***
  ORCID 0000-0003-4513-1840 · Github aarmey · twitter aarmey
  Department of Bioengineering, University of California, Los Angeles; Department of Bioinformatics, University of California, Los Angeles; Jonsson Comprehensive Cancer Center, University of California, Los Angeles; Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles
  *Corresponding author: a@asmlab.org

Abstract

Cell plasticity, or the ability of cells within a population to reversibly alter their phenotype, is an important feature of tissue homeostasis during processes such as wound healing and cancer. Plasticity operates alongside other sources of cell-to-cell heterogeneity such as genetic mutations and variation in signaling. Ultimately these processes prevent most cancer therapies from being curative. The predominant methods of quantifying tumor-drug response operate on snapshot population-level measurements and therefore lack evolutionary dynamics, which are particularly critical for dynamic processes such as plasticity. Here we apply a tree-based adaptation of a hidden Markov model (tHMM) that employs single cell lineages as input to learn the characteristic patterns of single cell heterogeneity and state transitions in an unsupervised fashion. This model enables single cell classification based on the phenotype of individual cells and their relatives for improved specificity in pinpointing the structure and dynamics of variability in drug response. Integrating this model with a modular interface for defining observed phenotypes allows the model to easily be adapted to any phenotype measured in single cells. To benchmark our model, we paired cell fate with either cell lifetimes or individual cell cycle phase lengths (G1 and S/G2) as our observed phenotypes on synthetic data and demonstrated that the model successfully classifies cells within experimentally tractable dataset sizes. As an application, we analyzed experimental measurements of cell fate and phase duration in cancer cell populations treated with chemotherapies to determine the number of distinct subpopulations. In total, this tHMM framework allows for the flexible classification of single cell heterogeneity across lineages.

Summary points

- A lineage tree-based hidden Markov model (tHMM) quantifies cell-to-cell variability and dynamic population changes.
- A modular interface to define phenotypes allows the model to work with any single cell phenotypic measurements.
- The model accurately classifies cells within experimentally tractable dataset sizes.
- Classifying cells based on their phenotypic heterogeneity and relationships uncovers dynamics hidden at the population level.

Author Summary

Heterogeneous traits, such as resistance or susceptibility to a drug, can be correlated in related cells because of shared or inherited factors. These acquired traits may be the result of the microenvironment, epigenetics, or mutations. Using cells' relationships we capture these dynamic transitions between different cell states and arrive at a more accurate quantification of cell heterogeneity within tumor populations compared to clustering such as K-means that ignores any relationship between the individuals. Our computational approach employing a modified hidden Markov model provides greater specificity by identifying intratumor heterogeneity on a single cell level based on lineage histories and can identify dynamic changes in population structure upon treatment.

Introduction

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

Large-scale profiling studies can find molecular features that associate with drug response using population-level samples.[10,11] These associations, while valuable, can miss the contribution of cell-to-cell heterogeneity, and especially stochastic changes in individual cell states that have significant effects on overall tumor drug response.[3,12,13] The most common methods for quantifying drug response are metrics of tumor cell population expansion or contraction.[14–17] Recent research has made efforts to track phenotypic measurements of fitness at the single cell level;[18,19] however, even single cell measurements are typically performed with snapshots that subsequently miss the contribution of individual cells in the overall population response.[20]

Measurements accompanied by lineage relationships are uniquely valuable for studying inherited phenotypes within families of people. This value is evident in linkage studies wherein relatives are used to identify or refine the genetic determinants of disease.[21–23] Notably, linkage studies can identify genetic determinants with greater power than even large association studies because relatives essentially serve as internal controls.[24] Linkage studies also start with the phenotype of individuals, rather than grouping based on molecular differences. Defining heterogeneity phenotypically has previously helped to identify subpopulations that would be hidden if using molecularly defined clusters.[25] While the inherited factors are different between cells (e.g., proteins, RNA) and people (DNA), such approaches are likely to be similarly useful in populations of cells. Recently, constructing phylogenic trees of cancer cells using lineage tracing and single cell sequencing has helped to characterize the directionality of metastatic seeding.[26] Lineage-resolved data has also demonstrated particular value in uncovering cell-to-cell heterogeneity due to transient differences outside of cancer.[27,28] Therefore, tools to analyze and explore these data will be critical.

Hidden Markov models (HMMs) provide an efficient strategy to infer discrete states from measurements when a series of co-dependent observations are made. An example of this is their wide-spread use in time series analysis, where each measurement is dependent on those that came before.[29,30] Recognizing this co-dependence allows HMMs to make accurate inferences even in the presence of extremely noisy measurements since each neighboring measurement provides accumulating evidence for a prediction.[31] HMMs have been adapted to lineage trees (tHMMs) so that each measurement across the tree can similarly provide accumulating evidence for a prediction. Just like with time-series data, these models can provide very accurate predictions despite noisy measurements and limited information by recognizing the co-dependence between related measurements.[32,33] tHMMs have been used in a multitude of applications, from image classification to comparative genomics.[34,35] In cells, these models have been fit to lineages collected from stem cells and bacteria colonies but have always required tailor-made implementations.[36,37] Improvements in cell tracking and high-throughput imaging promise to make these models valuable techniques for

studying the plasticity of heterogeneous cell populations. However, widespread use will require usable implementations that can be readily adapted to different experimental measurements, examples of successful tHMM applications that will encourage increased usage, and standards for experimental validation.

Here, we develop a generic implementation of tHMMs with a defined interface for integrating diverse types of measurements on cell lineage trees. We use the relationship between the cells to analyze how populations of breast cancer cells respond to therapy with a cell cycle reporter. Single cell measurements of the cell cycle revealed extensive variation in drug response that is not captured in population-level measurement. This model allows us to quantify the dynamics and phenotypic features of heterogeneity in drug response. We are able to predict the most likely number of phenotypically distinct subpopulations, the characteristics of those subpopulations, transition probabilities from one state to another, and each cell's expected state. This work, therefore, provides a flexible phenotype-driven route to discovering cell-to-cell variation in drug response, demonstrates an overall strategy for quantifying the dynamics of cell heterogeneity, and implements a very general software tool for the widespread use of tHMM models.

Results

Lineage information provides unique information about the source and dynamics of intratumoral heterogeneity

Single cells grow and then divide into two daughter cells, eventually forming a binary genealogical tree, also known as a lineage. We collected the single-cell measurements in the form of lineage trees to track these relationships. The life cycle of each cell before division includes phases that have to pass one after another. Cell division is allowed upon a successful through each cell cycle phase. To illustrate the unique value of lineage measurements in analyzing intra-tumoral and drug response heterogeneity, we plotted two sets of sample lineages from measurements of the breast cancer cell line AU565 (Fig. 1 a-b). The single cell lineages reveal a striking difference in cell cycle phase durations and cell division dynamics between the two samples. Population-level measurements, on the other hand, would be unable to identify this difference as the starting and ending cell numbers are not different. Measurements that record the history of cells (e.g., CFSE staining, Luria-Delbruck experiment) can help to identify these variations within cell populations but must make assumptions about the dynamics of heterogeneity.[13,28] Lineage measurements, by contrast, provide sufficiently rich temporal information to quantify the specific structure of the phenotypic heterogeneity.

As a further exploration of the data, we randomly sampled lineages from the gemcitabine-treated cells (Fig. 1 c). Gemcitabine is a chemotherapy agent that disrupts DNA replication and results in extension of and apoptosis in S phase.[38] The S/G2 phase lengths in treated cells are noticeably extended compared to control cells, and cell division is diminished. Individual lineages show striking variation, including anywhere from zero to three cell divisions, but tightly shared behavior among cells and their close relatives. By observing individual cells, a variety of phenotypic measurements such as inter-mitotic times, cell cycle phase durations, motility, cell death and division, morphology, and protein and transcription factor levels can be characterized in parallel.[26]

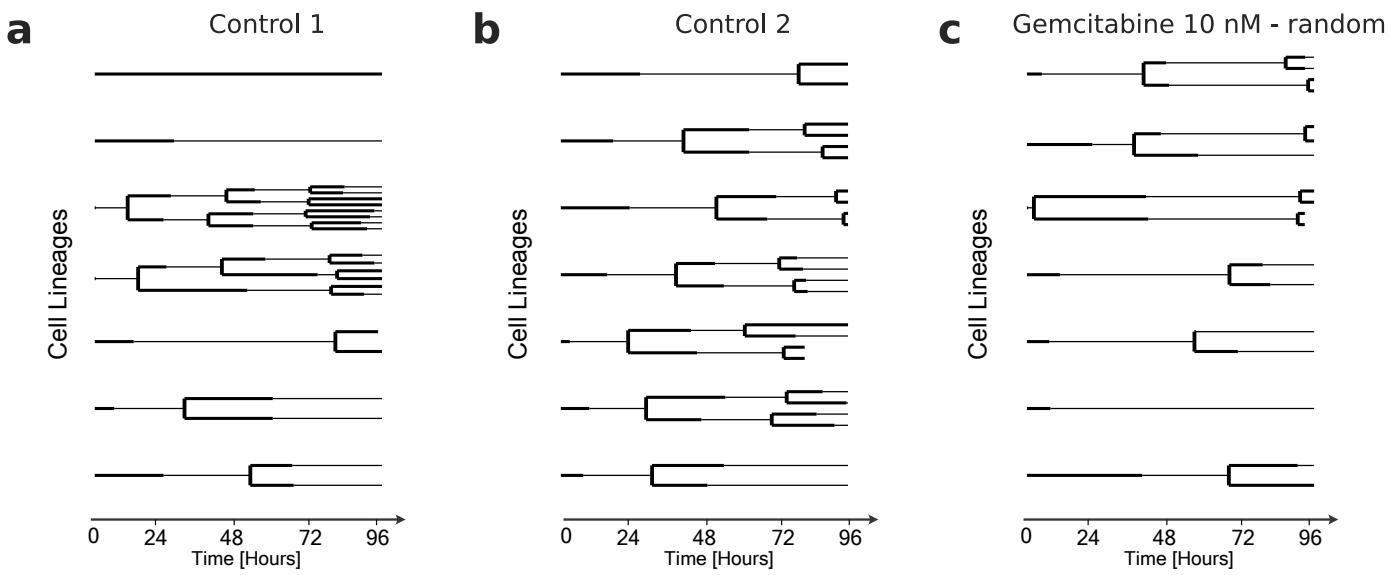

Figure 1: Total cell number is insufficient to distinguish the structure of heterogeneous populations. (a/b) Selected lineages of untreated AU565 cells. (c) Randomly selected lineages of AU565 cells treated with 5 nM gemcitabine. Each line indicates the lifetime of one cell. The line branching into two lines indicates cell division. The G1 and S/G2 phase durations are shown in solid thick and thin lines, respectively.

A lineage tree-based hidden Markov model infers the state of cells given measurements on lineage trees

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

After fitting, the model can provide a variety of information (Fig. 2c). First, it infers the starting and transition probabilities of each state. Second, the distribution of cells' phenotypes in each state are estimated and can be compared to distinguish how cells of each state behave. For instance, if we use the growth rates of cells as their phenotype, we may observe a subpopulation of cells with shorter times to division, and another with longer times. Moreover, the state of each individual cell can be predicted from the fit data or new measurements. Finally, the model provides a likelihood of each cell's observations and therefore the data overall. This last quantity can be used, for example, to estimate the number of distinguishable cell states. When implementing these processes, we ensured that a cell's measurements are defined through a modular interface, allowing other forms of data to be easily integrated.

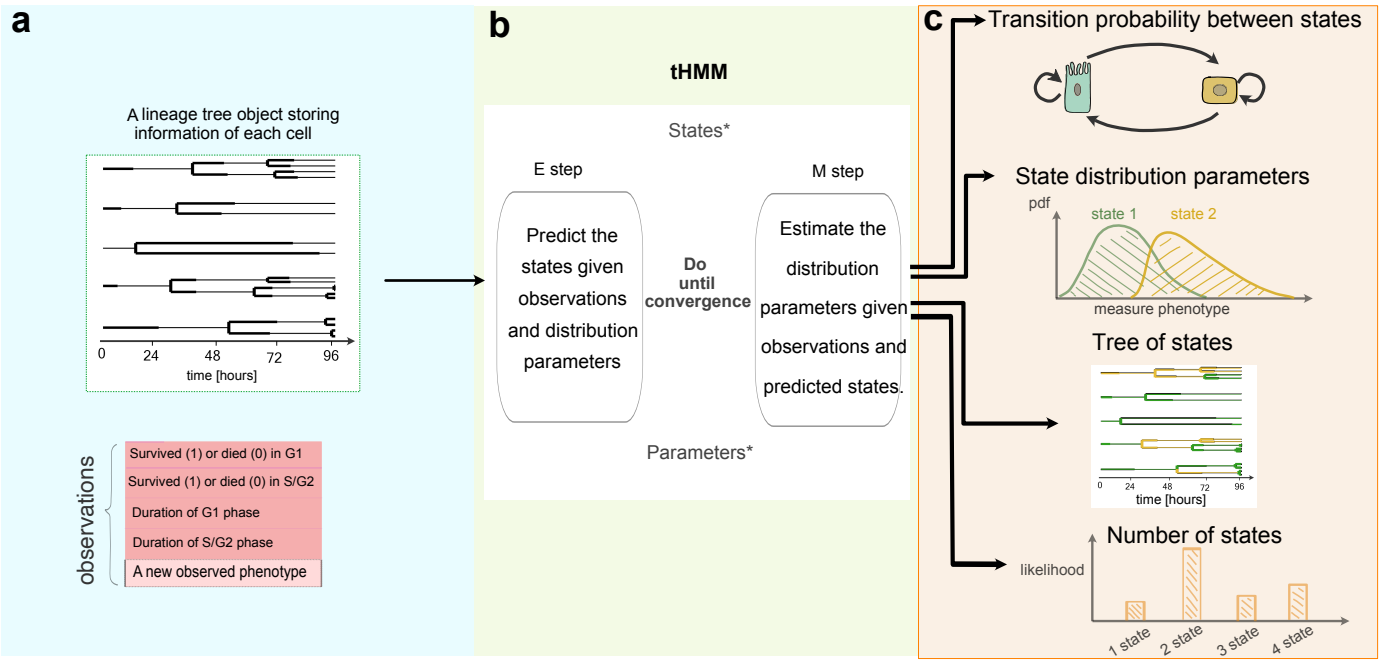

Figure 2: The tHMM interface. (a) Input data takes the form of single cell measurements across time, where the lineage relationship between cells is known. (b) The fitting process includes expectation and maximization steps, where model parameters are iteratively updated until convergence. (c) Output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

Experiments of finite time necessitate corrections for experimental censorship

The phenotypic measurements used in this study, either synthetic or experimental, were in the form of cell lifetime and fate. Modeling the duration of each cell's lifetime is complicated by the influence of experimental parameters. Specifically, cells at the beginning or end of an experiment persist beyond its duration and so, while we observe them, we do not know their exact duration. Previously, this has been addressed by simply removing incompletely observed cells.[39] However, doing so results in a systematic bias, where longer-lived cells are preferentially eliminated. On the other hand, ignoring the truncation of these values also creates a systematic bias; by creating an upper bound on the cells' lifetimes.

To correct for this effect in our model, we marked cells that encountered the start or end bounds of the experiment. When estimating the properties of a cell's lifetime or the probability of a cell's survival versus death for these cells we instead used a censored estimator or the survival function of the distribution.[40] Using synthetic data, we verified that this correction resulted in accurate phenotype estimations of 95% or higher (Figs. 3, S3, S10). As shown, using the censored estimator as opposed to the uncensored model on incompletely observed cells results in higher cell state assignment accuracy. Thus, taking into account the cells that outlive the bounds of the experiment through a censored model ensures that we can accurately infer cell states despite the contribution of experiment duration.

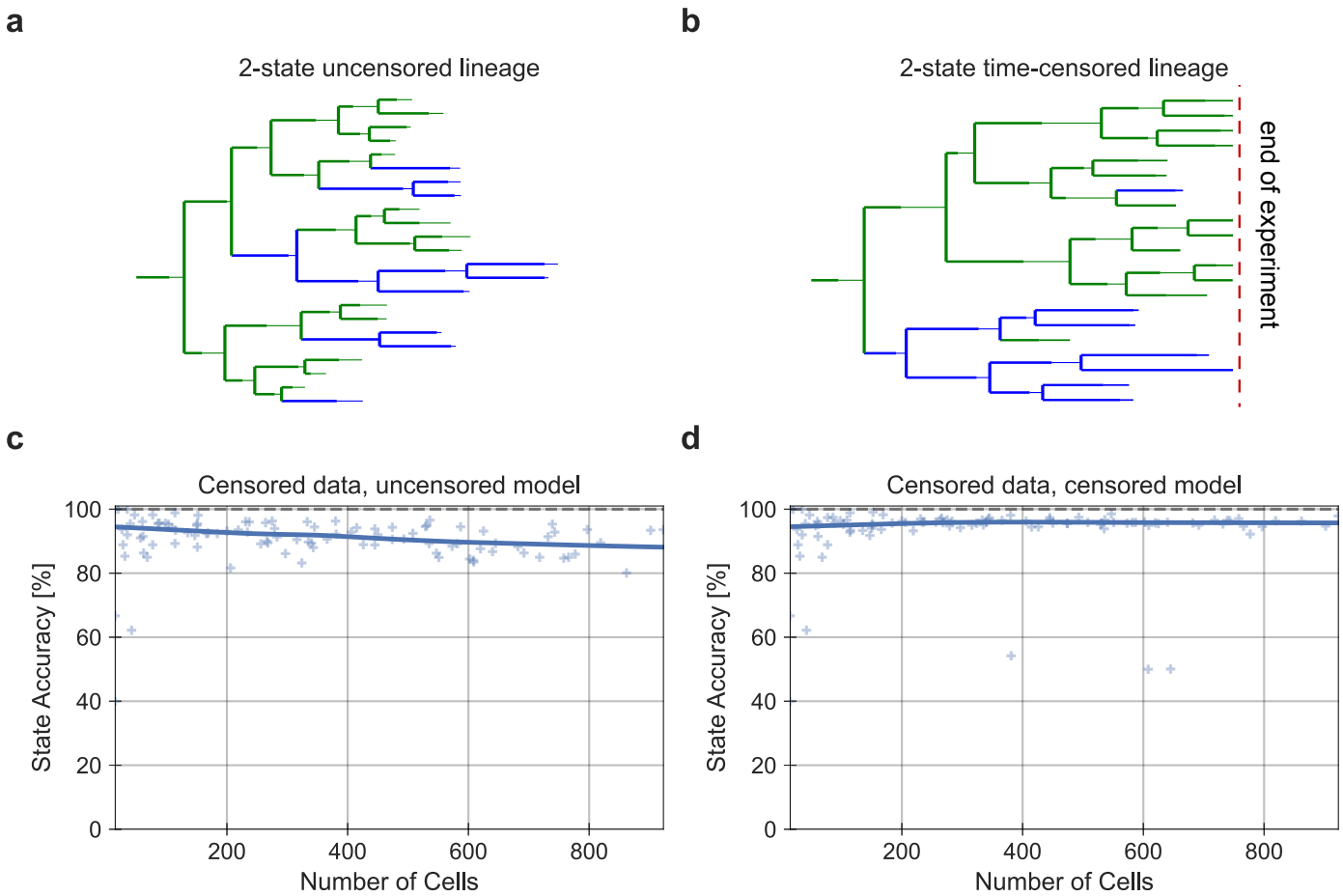

Figure 3: Experiments of finite time necessitate data censorship corrections. (a) An example, synthetic, uncensored two-state lineage. (b) An example, synthetic, censored two-state lineage. State 0 and 1 cells are shown in green and blue. (c) State assignment accuracy with censored lineages using an uncensored model. (d) State assignment accuracy with experiment time and cell death-censored lineages. Each scatter point represents the state assignment accuracy of a lineage with the shown cell number that has been fitted to the model. The solid lines show the Lowess trendline of the individual run accuracies.

Synthetic lineage benchmarks show a tHMM can accurately estimate population behavior

To evaluate how accurately a tHMM model could infer the behavior of multi-state cell populations, we used synthetic populations of cells in a wide variety of configurations, such as various populations sizes and abundance of the present states. In each case, we determined that our implementation of a tHMM model could accurately infer the hidden states and parameters of a population given at least 100 cells. This synthetic data included situations that were both not censored (Figs. S1, S2, S8, S9) and censored due to cell death and experimental duration to be representative of experimental lineages (Figs. 4, S10, S3). In addition to varying the number of cells in a population, we benchmarked populations with varied cell state percentages (Figs. S4, S5) and varied degrees of phenotypic differences (Figs. S6, S7, 5). This benchmarking consistently showed that the tHMM model would provide accurate results across a range of circumstances.

More specifically, one of the benchmarking studies we performed was with data matching our measurements of AU565, where G1 and S/G2 phase durations, represented by a Gamma distribution, and their corresponding cell fate, represented by a Bernoulli distribution, were quantified (Fig. 4).

Although the tHMM model was fit with no information about the true underlying parameters of the simulated cells, it distinguished the pre-assigned two underlying cell states' phenotypes (Fig. 4 b–d) and member cells with more than 95% accuracy (Fig. 4e). The Wasserstein distance metric was used to represent the difference between the true and estimated Gamma distributions to show the accuracy of parameter estimation (Fig. 4 d). On the population level, the distance between the true and estimated transition probabilities calculated by the Frobenius norm was less than 0.1 for 100 cells or more. Starting probabilities were compared to their corresponding true values in terms of Euclidean distance and showed less than a 0.2 error for populations with 10 lineages or more (Fig. 4f–g). Thus, we are confident that with similar experimental data, we should derive accurate results.

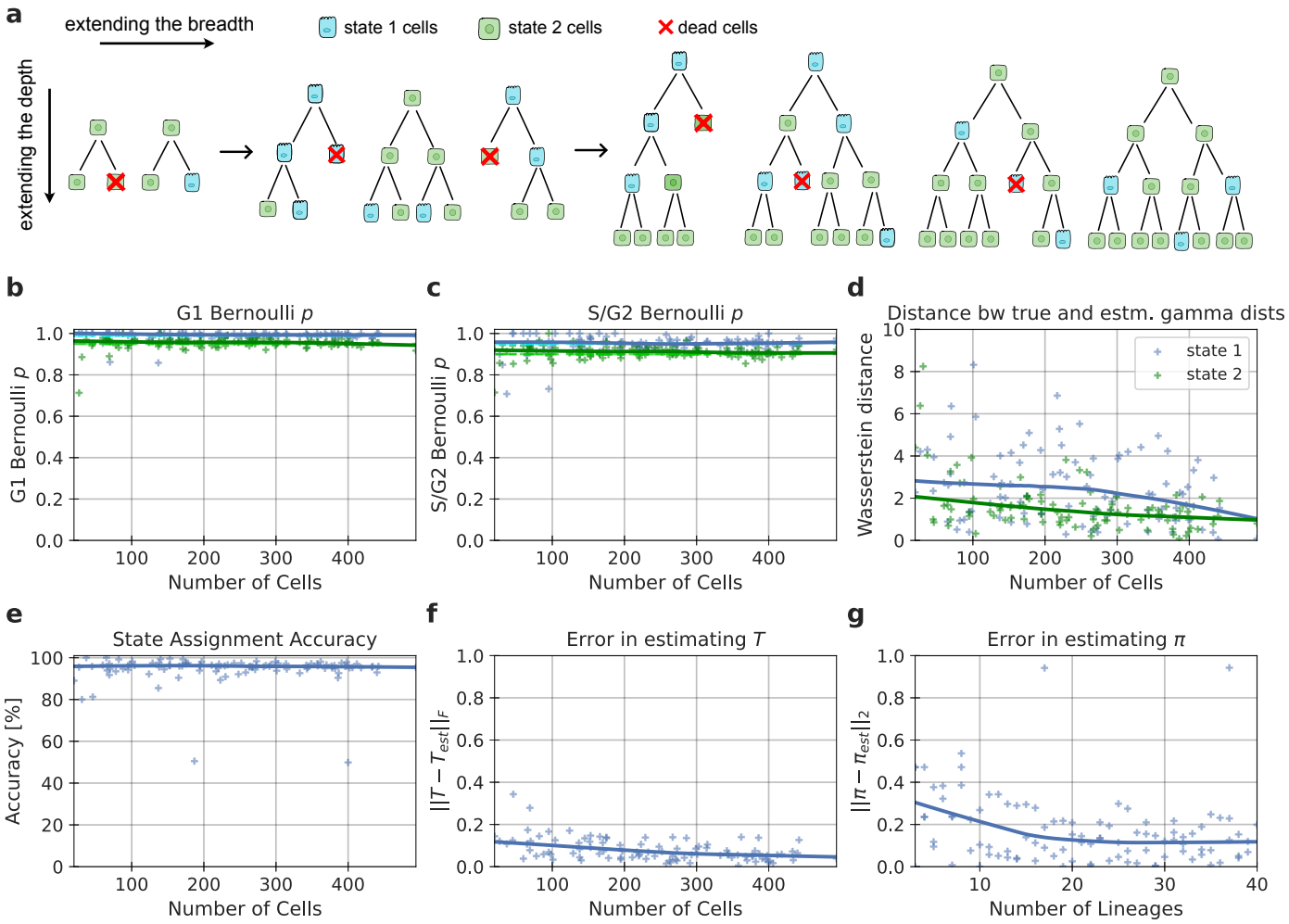

Figure 4: Model performance on censored lineages of increasing breadth and depth. (a) Two-state populations of increasing breadth (increasing number of initial cells and therefore lineages) and of increasing depth (increasing experiment time) are analyzed. The states are shown in green and blue colors and red shows cell death. Each point in the scatter plots represents the inferred value for a model evaluation trial with the number of cells shown in the x-axis. The solid lines are the Lowess trendline across the individual trials. The accuracy in estimating the Bernoulli parameters for G1 and S/G2 phase, respectively, is shown in (b-c). The distance between the true and estimated Gamma distributions associated with phase lengths for the two states is depicted in (d). The model performance is also evaluated based on the state assignment accuracy (e), the error between the estimated and the true transition rate matrix (f), and the initial probability vector (g). Note that the Wasserstein distance between the true and estimated distributions for each state is much lower than the distance between two distributions that are quite similar (Fig. 5b).

Lineage information improves cell state identification with heritable phenotypes

Cells of even very distinct molecular states can have partly overlapping phenotypes due to non-heritable variation. Therefore, we sought to evaluate how different two states would need to be for us to accurately identify them as distinct (Fig. 5a). We varied the G1 phase duration of two states from identical to very distinct (Wasserstein distance greater than 70) (Fig. 5b) and quantified the accuracy of our model. While the phenotypic observation of a given state had to be different for our model to accurately assign cells, even moderately overlapping phenotypes could be distinguished by using the lineage relationships of cells (Fig. 5c). As a baseline comparison, we used K-means clustering to cluster cells based on their phenotype without using their lineage relationships (Fig. 5c). A tHMM consistently outperformed this approach. The model performance in censored and uncensored populations were similar (Fig. S6, S7). This shows that the heritability of a cell state can help to more accurately identify cells with partially overlapping phenotypes.

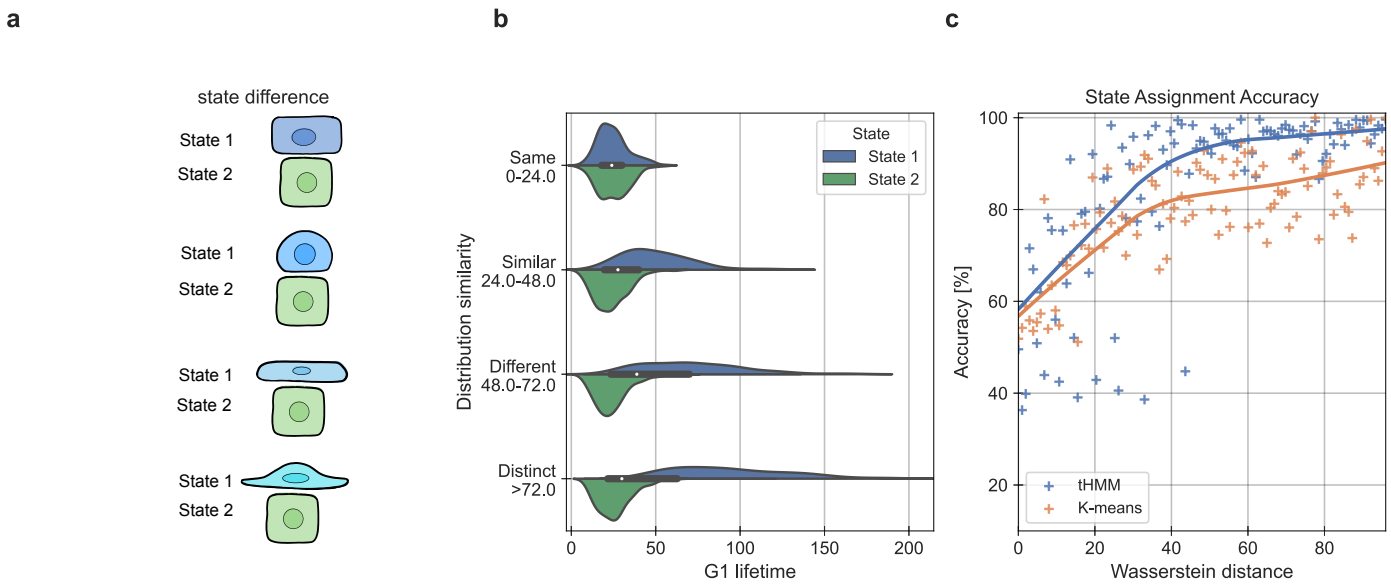

Figure 5: Model performance versus the Wasserstein distance between states. (a) Cartoon of how two states can vary in their phenotypic similarity. On the top, cells might be virtually indistinguishable (here based on shape). On the bottom, they might be so different that looking at one cell is sufficient to identify its state. (b) The distribution of G1 duration is varied in state 1 (blue) while the other state is kept constant. (c) State assignment accuracy versus the Wasserstein distance between state phenotypes. Each point represents the accuracy of state assignment for a lineage created by a set of parameters that yield the shown Wasserstein distance between the two state distributions. Either a K-means model was used with the phenotypes of each individual cell (orange) or the full tHMM model was applied (blue). The solid lines show a Lowess trendline of the model accuracy.

Likelihood-based model selection can effectively identify the number of distinct states

One does not usually know the number of distinct cell states within a population. Further, the number of distinct states may depend on the environmental context of the cells, particularly because we use phenotypic measurements.[41,42] To test whether we could infer the number of phenotypically-distinct states, we performed model selection using the Bayesian information criterion (BIC) while varying the number of states in synthetic data (Fig. 6). We normalized the BIC values such that zero corresponds to the state with the highest likelihood. The synthetic populations that were used included approximately 250 to 650 cells with known cell phase fate and phase lengths. The predicted number of cell states was

correct, and the few incorrect predictions still centered around the true answer, for both uncensored and censored lineages (Fig. 6). This indicated that model selection can help to identify the appropriate number of cell states for a set of measurements.

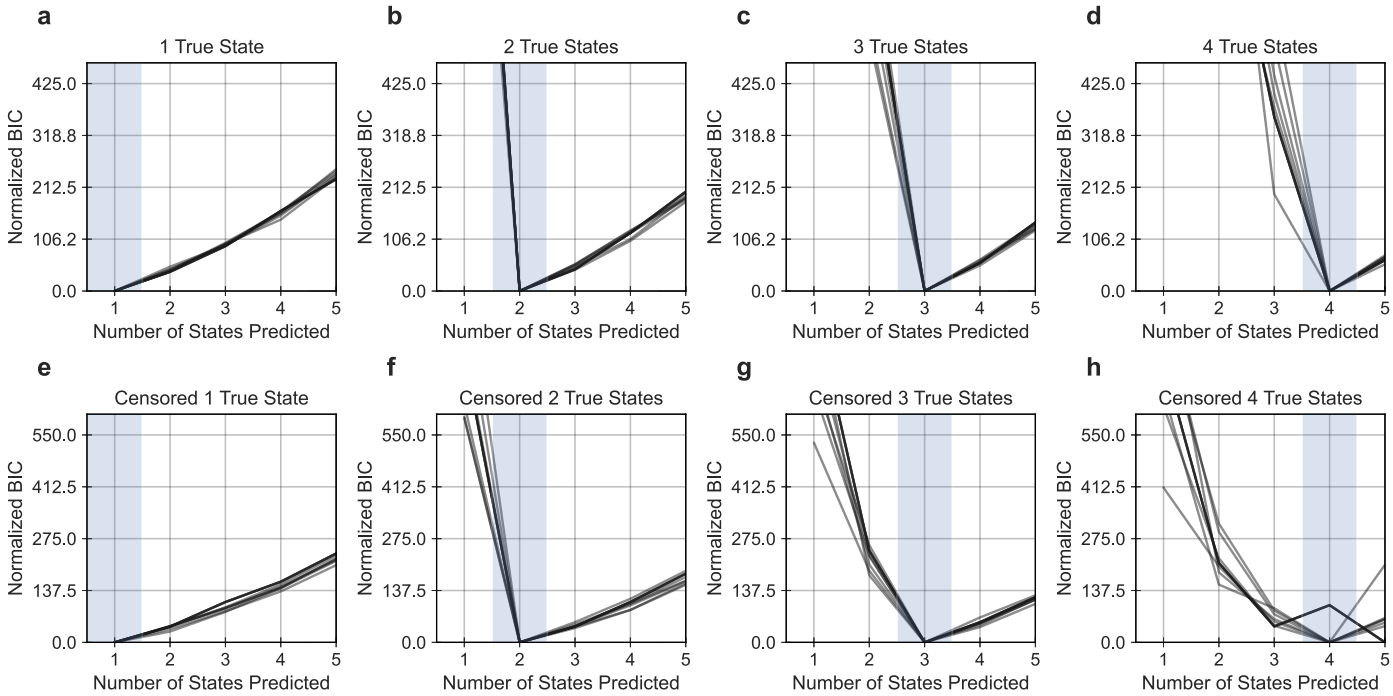

Figure 6: Model selection effectively identifies the number of distinct states. (a)-(d) Model BIC for uncensored lineages with 1–4 states. (e)-(h) Model BIC for censored lineages with 1–4 states. BIC values are normalized such that the optimum is equal to 0. Histograms are shown for the minimal BIC over 10 repeated analyses.

tHMM infers multiple distinct subpopulations in experimental drug response data

As an application of our model, we used phenotypic measurements of the G1 and S/G2 phase durations and terminal cell fates of AU565 cells in a control condition and when treated with 3 concentrations of gemcitabine and lapatinib. We imaged cells every 30 minutes for 96 hours and then tracked them over time to assemble their lineage relationships. The lapatinib and gemcitabine-treated populations including control contained a total of 5290 and 4537 cells, respectively, with lineages that had 1–5 generations of cells. The model was then fit to each drug's data across all conditions, holding the initial and transition probabilities constant across concentrations while allowing the phenotype emissions to vary. We forced the duration shape parameter to be shared and the scale parameters to vary monotonically with drug concentration. The Bernoulli cell fate parameters were estimated without constraints. We compared models of 1–7 states using the BIC, where the lowest BIC value across numbers of states indicates the optimal model correcting for complexity (Fig. 7a-b). The data for each compound indicated the presence of multiple inherited states.

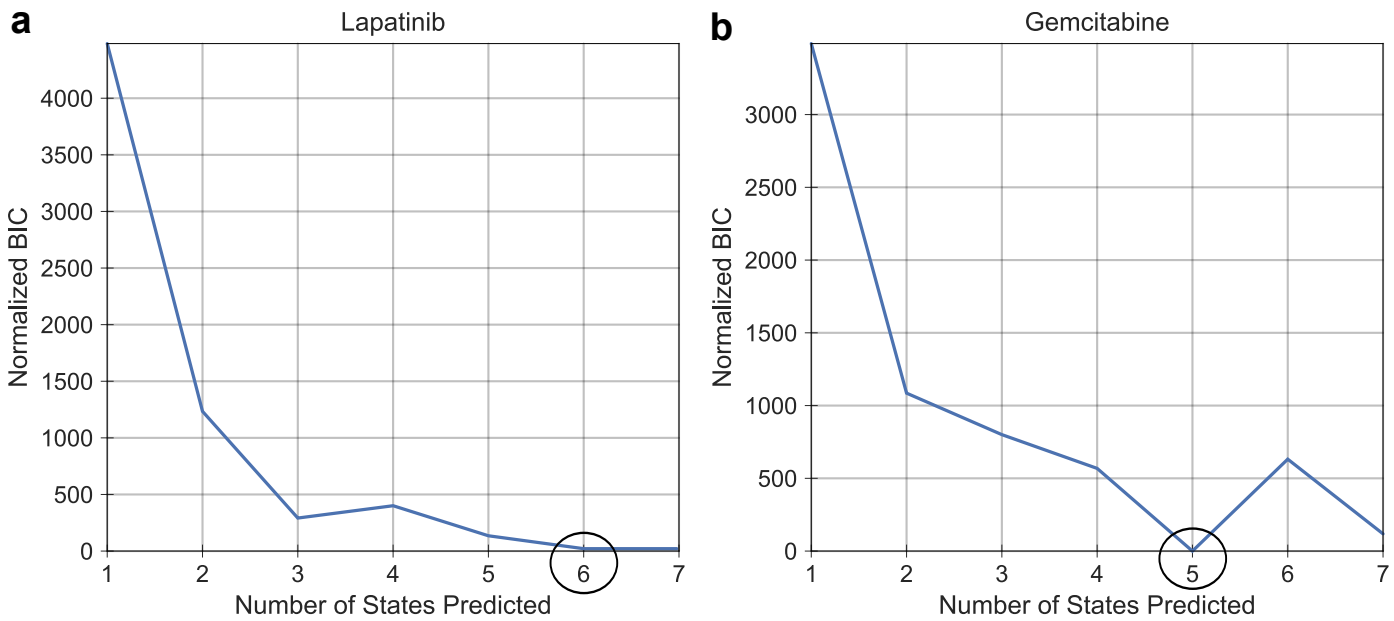

Figure 7: BIC-based model selection infers the likely number of phenotypically distinct subpopulations. Normalized BIC values for (a) cells in control and treated with 5 nM, 25 nM, and 250 nM of lapatinib and (b) cells in control and treated with 5 nM, 10 nM, and 30 nM of gemcitabine. For both drugs, the BIC values for all conditions were normalized to zero based on the minimum value.

Lapatinib treatment leaves stable and cyclical states in the cell population

We fitted the lapatinib treated data to the model with 6 states based on our BIC-based model selection (Fig. 7a). Fitting revealed states of widely varying persistence over generations, from a 0.02 probability of remaining in state 6 to a 0.91 probability of remaining in state 2 (Fig. 8a). Interestingly, state 5 and 6 formed a cycle, wherein the most probable transition was between these two states (Fig. 8a).

Examining the phenotypes of each state revealed distinct drug responses. Lapatinib is a EGFR/HER2 inhibitor that induces cell cycle arrest in G1 phase.[43] Every state showed an increase in G1 phase lifetimes, and states were most distinct in their G1 durations (Figs. 8b-f, S11). State 4 showed the most drastic response to treatment, going from one of the fastest-growing states to most arrested (Fig. 8b,d,f). In contrast to G1 phase, there was almost no change in the mean S/G2 duration except for cells in state 5 (Fig. 8g). The probability of cell death in G1 phase was higher in state 2 and 3, and in S/G2 phase, only cells in state 3 had a chance of cell death (Figs. 8h-i). Cell death appeared only at the highest concentrations of lapatinib. States were separated based on whether they responded to drug treatment with cell death or phase duration extension. For example, states 2 and 3 had the highest probabilities of cell death in G1 phase, but also the shortest duration (Figs. 8f-i). This likely reflects differences in the dynamics of death versus cell cycle progression outcomes.

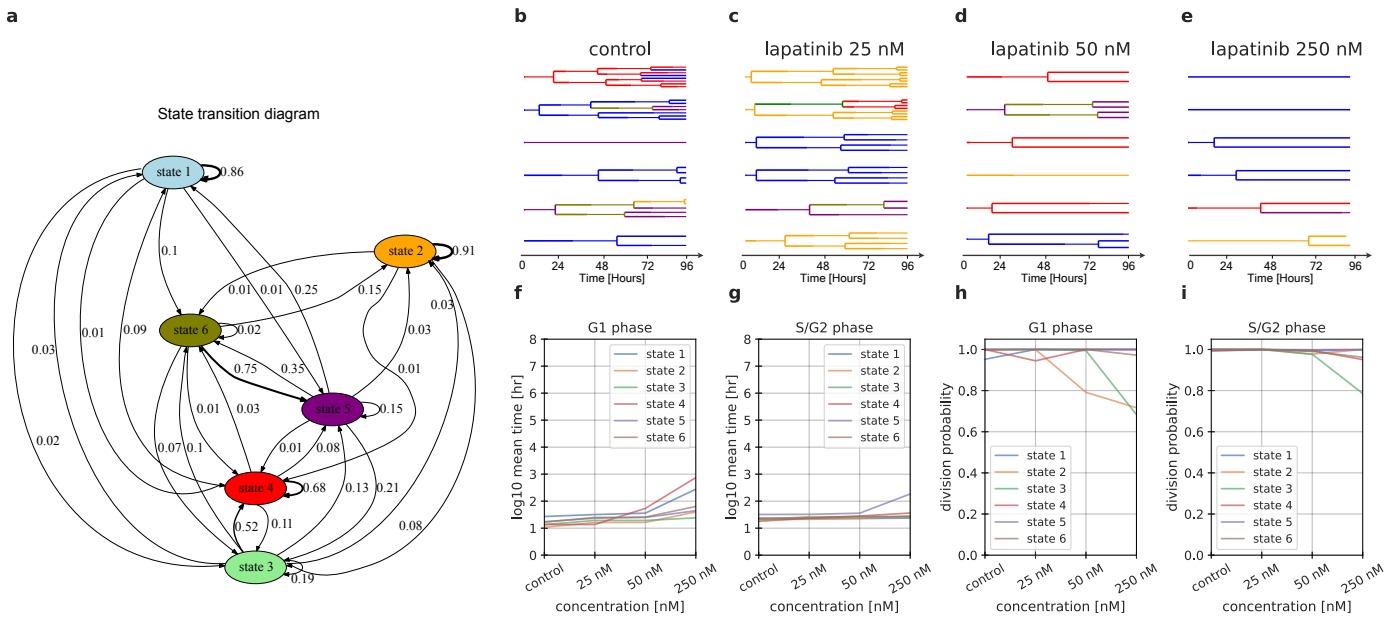

Figure 8: State-specific emissions of the lapatinib-treated data. (a) State transition graph showing the probability of state transitions among the three predicted states. (b-e) Lineage trees after fitting the model and state assignment (control, 25 nM, 50 nM, and 250 nM). (f-g) The log10 of fit mean time of G1 and S/G2 phase durations for different concentrations. Most effects of lapatinib treatment is observed in G1 phase. (h-i) Probability of dividing versus dying in the G1 and S/G2 phase for different concentrations.

tHMM identifies more stable states in gemcitabine-treated populations

Gemcitabine, as mentioned previously, is a chemotherapy agent that induces cell cycle phase extension and apoptosis in S phase by disrupting DNA repair. The AU565 cells were treated with 5 nM, 10 nM, and 30 nM of gemcitabine, and model selection inferred 5 states in the population (Fig. 7b). Examining the 5-state fit revealed relatively stable states 1, 3, and 4 (Fig. 9a–e, S12). States 2 and 5 had higher rates of transition between one another in a partial cycle.

The phenotypic effects of drug treatment were evident in both cell cycle phases, though most prominently in S/G2 phase, where the phase duration of states 2 and 4 doubled (Fig. 9f,g). States 2 and 4 were also similar in having a dose-dependent decrease in their survival at the end of S/G2 phase, but few effects in G1 phase (Fig. 9h-i). Interestingly, one state, state 1, maintained a normal growth rate at even the highest concentration of gemcitabine, but with higher rates of cell death clearly evident in the lineages themselves (Fig. 9a,f,i; S12). In total, the tHMM model was effective in identifying subsets of cells with divergent phenotypic responses to drug treatment alongside the relationships between cells in the population.

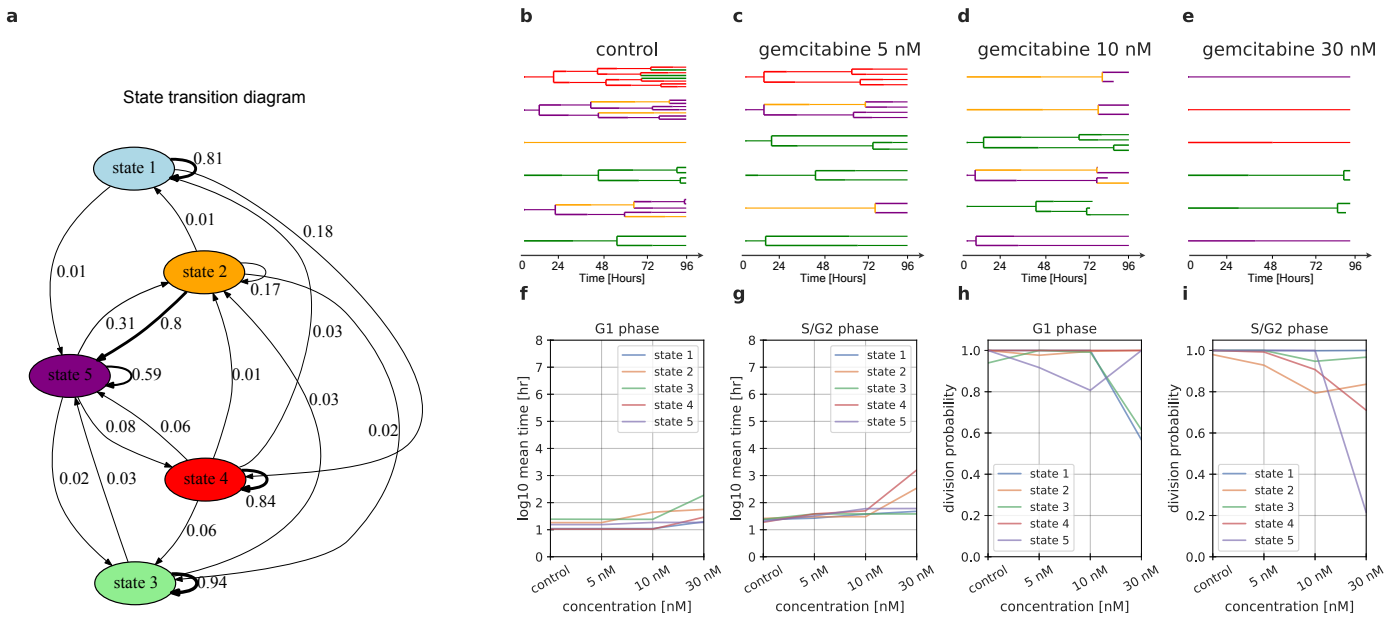

Figure 9: State-specific emissions of the gemcitabine-treated data. (a) State transition graph showing the probability of state transitions among the three predicted states. (b-e) Lineage trees after fitting the model and state assignment (control, 5 nM, 10 nM, and 30 nM). (f-g) The log10 fit mean time of G1 and S/G2 phase durations for different concentrations. (h-i) Probability of dividing versus dying in the G1 and S/G2 phase for different concentrations.

Discussion

Intratumoral heterogeneity in cancer cells enables them to adapt in response to therapy. Even in the absence of genetic mutations, other heritable variation serves as a substrate for selection.[44] In this paper, we introduced a tree-based hidden Markov model that classifies single cells from a heterogeneous population based solely on their phenotypic traits and relationships. Benchmarking the model showed that it can provide accurate results using feasible experimental designs. Of particular importance, the tHMM showed good sensitivity to subpopulations at lower frequencies (Figs. S6, S7). Also, comparing the model to more standard clustering, the tHMM showed that relationship information helps to more accurately identify cell states (Fig. 5). Two critical advancements in the current work are a modular interface for using tHMM models with various phenotypes (Fig. 2), and properly handling time censorship as a consequence of experiments with finite duration (Fig. 3).

We used single cell lineage tracking data of AU565 cancer cells treated with lapatinib and gemcitabine as a demonstration of the model. G1 and S/G2 cell cycle phase durations and cell fate measurements were used as relevant cell phenotypes to quantify the anti-cancer effects of these drugs. We were able to identify 6 and 5 distinct subpopulations within the lapatinib and gemcitabine-treated data, respectively (Fig. 7). The phenotypic features of each state were quantified in parallel (Figs. 8, 9). Lapatinib is known to inhibit cell proliferation by inhibiting Akt/mTOR pathway activity, which is a key regulator of G1 phase progression.[45] Similarly, our analysis in the lapatinib-treated population indicated that cells, regardless of their state, experienced a prolonged G1 phase, but varied in their susceptibility. We observed that those states with a higher susceptibility for cell death had lesser effects on the duration of each cell cycle phase (e.g., state 3 versus 4). In gemcitabine treated cells, we observed that most predicted states in gemcitabine-treated cells were highly heritable, and the

phenotypic effect of drug treatment was more varied. Drug effects were minimal up to 10 nM and then effects were most pronounced in inducing cell death. For instance, cells in state 5 have only a 20% chance of survival in S/G2 phase, and two states (1 and 3) had about a 60% chance of surviving G1 phase.

Modeling advancements will further improve on our approach. Cells may express a continuum of, rather than discrete, phenotypic states.[39] If this is the case, a continuous latent variable model would lead to a refined view of the population-level heterogeneity. A discrete model like the one used here should, however, still provide an accurate estimate by breaking up the continuous state-space. Continuous latent variable models also have additional challenges in implementation and interpretation.[46,47] Careful handling of each states' phenotypic distributions might also improve the model's accuracy and power to identify distinct states. For example, the eventual fate of cells and their cell cycle durations are likely correlated which could be handled through a multivariate distribution accounting for this covariance.[48] This becomes even more important with the incorporation of other phenotypic information such as migration, cell shape, or other features, all of which are likely to be correlated to some extent.

Experimental advancements will improve the utility and accuracy of single cell analysis using lineage information. Currently our experimental data is limited to 96 hours, covering up to roughly five generations of cells. However, traits such as resistance may develop over more generations and longer timescales.[13,49–51] Longer data collection becomes challenging due to factors such as phototoxicity and cell stress.[52] Improved imaging modalities and experimental platforms might allow for longer tracking experiments, with reduced phototoxicity, in more physiologically representative environments such as engineered 3D extracellular matrix.[53,54] Currently, the model is agnostic as to whether the heterogeneity it identifies is pre-existing or drug-induced. Collecting data in which cells are tracked before and after drug treatment would enable some linkage between the pre- and post-exposure phenotypes of cells.[55] Finally, while we identify these states as subpopulations of cells that are phenotypically distinct, we currently cannot comment on the molecular factors leading to these phenotypes. However, we expect that single cell molecular analyses such as transcriptional profiling, tied with molecular barcodes, should allow us to align molecular and phenotypic states in the same populations of cells.[56]

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

Materials and Methods

Data and Software Availability

All analysis and data were implemented in Python v3.9 and can be found at https://github.com/meyer-

lab/tHMM.

Experimental cell lineage data

Stable cell line creation, drug treatments, and cell tracking were performed as previously described.[57] Briefly, AU565 cells were co-transfected with a transposase plasmid (Addgene #34879) and a donor plasmid that drove expression of a nuclear localized mCherry, puromycin resistance, and a fragment of HDHB fused to the clover fluorescent protein, which was used to track cell cycle progression. Cells stably expressing the nuclear and cell cycle reporter were selected for 7 days with 0.75 μg/ml puromycin. To track drug responses AU565 reporter cells were plated into 24-well plates with fluorobrite media containing 10% FBS glutamine, and penicillin-streptomycin. Then 24 hours later fresh media containing escalating doses of lapatinib and gemcitabine was added. After drug addition, plates were placed in the IncuCyte S3 and four image locations per treatment were imaged every 30 minutes for 96 hours. At 48 hours fresh media and drugs were added. Cell lineages from the IncuCyte images were manually tracked in FIJI[58] in order to record cell division, death, and the transition from G1 to S/G2 phase. Three biological replicates were collected and combined in the final data set.

Lineage tree-based hidden Markov model

The core assumption of a Markov chain is that the next state and current observations are only dependent on the current state. Proof of many of the expressions below can be found in Durand et al.[33] The tHMM is composed of several intermediate steps to carry out predictions and estimations.

Basic model structure

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

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

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

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

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

$$EL(n, k) = P(x_n = x | z_n = k)$$

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

Since the hidden states are unobserved, we need an expectation-maximization (EM) algorithm, for HMMs called the Baum-Welch algorithm, to find the states and their specifications. The EM algorithm consists of two steps: (1) the expectation step (E-step) and (2) maximization step (M-step.) In the E-step, using the whole lineage tree the probability of a cell and its parent being in specific states are calculated, such that for every cell and every state we have $P(z_n = k \,|X_n)$ and

$P(z_n = k,\ z_{n+1} = l\ |X_n)$. The E-step is calculated by the upward and downward recursion algorithms.

In the M-step, the distribution parameters of each state, the initial (π) probabilities, and the transition probability (T) matrices are estimated, given the state assignments of each cell. During fitting we switch between the E-step and the M-step and calculate the likelihood. If the likelihood stops improving below some threshold we take that to indicate that convergence has been reached. The following explains each step in detail.

E-step

Upward recursion

An *upward-downward* algorithm for calculating the probabilities in hidden Markov chains (HMCs) was previously proposed by Erphaim and Merhav[59] which suffered from underflow. This problem was originally solved by Levinson[60] for HMCs, where they adopted a heuristic based scaling, and then was improved by Devijver[61] where they introduced smooth probabilities. Durand et al.,[33] however, revised this approach for hidden Markov trees to avoid underflow when calculating $P(Z|X)$ probability matrices. To explain we need the following definitions:

- $p(n)$ is noted as the parent cell of the cell n, and $c(n)$ is noted as children of cell n.
- $\bar{X}$ is the observation of the whole tree and $\bar{X}_a$ is a subtree of $\bar{X}$ which is rooted at cell a.
- $\bar{Z}$ is the complete hidden state tree.
- $\bar{X}_{a/b}$ is the subtree rooted at a except for the subtree rooted at cell b, if $\bar{X}_b$ is a subtree of $\bar{X}_a$.

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

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

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

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

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

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

For non-leaf cells the values are given by:

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

where we calculate the non-leaf NF using:

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

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

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

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

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

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

$$logP(\bar{X} = \bar{x}) = \sum_n logNF(n)$$

This quantity acts as the convergence measure of the EM algorithm.

Downward recursion

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

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

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

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

Viterbi algorithm

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

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

$$\delta(n,k) = \max_{\bar{z}_{c(n)}}\{P(\bar{X}_n = \bar{x}_n, \bar{Z}_{c(n)} = \bar{z}_{c(n)} | z_n = j)\}$$

and the links between parent-daughter cells as:

$$\delta_{p(n),n}(k) = \max_{\bar{z}_n}\{P(\bar{X}_n = \bar{x}_n, \bar{Z}_n = \bar{z}_n | z_{p(n)} = j)\} = \max_j\{\delta(n,j)T_{k,j}\}$$

We initialize from the leaf cells as:

$$\delta(n,k) = P(X_n = x_n | z_n = k) = EL(n,k)$$

and for non-leaf cells use:

$$\delta(n,k) = \Big[\prod_{v \in c(n)} \delta_{n,v}(k)\Big] \times EL(n,k)$$

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

$$Z^* = \max_k\{\delta(1,k)\pi_k\}$$

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

M-step

In the M-step, we find the maximum likelihood of the hidden Markov model distribution parameters. We estimate the initial probabilities, the transition probability matrix, and the parameters of the observation distributions. The maximum likelihood estimation of the initial probabilities can be found from each state's representation in the root cells:

$$\pi_k^* = \gamma_1(k)$$

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

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

Estimating emissions distribution parameters

In the current study, we used two emissions distributions; first, a Bernoulli distribution for the probability of each cell fate, either at the end of each phase or at the end of cell's lifetime; second, a Gamma distribution for the durations of each cell cycle phase or overall cell lifetime. To estimate the distribution parameters after finding the cell state assignments, we calculated their maximum likelihood estimation weighted by their proportional assignment to that state. For estimating the

Bernoulli distribution parameter for cell fate we simply found the state assignment-weighted sample mean of the observations. For the Gamma distribution parameters we used linear constraint, trust-region method fitting as implemented in scipy.optimize as the "trust-constr" method.[62] The likelihood was calculated for measurements weighted by their proportional assignment to that state and based upon whether they were censored. Where data was completely uncensored, we used a simplified iterative fitting scheme.[63]

Synthetic lineage data

We generated synthetic lineage trees with K discrete states and N total number of cells for benchmarking. Lineages were composed of two primary data structures, the state and emissions trees. The state tree was randomly seeded with a root cell determined by the starting probabilities, then expanded by randomly sampling transitions based on the transition probability matrix. After creating the state tree with the desired number of cells, the emission tree is built upon it. Emissions were randomly sampled from the distributions for each cell's state. Finally, the effects of the emissions were applied to the tree when necessary. If any cells died, their progeny were marked as unobserved by making their emissions equal to NaN. If applicable, the effects of finite-duration experiments were also applied. Cells existing outside of the experiment duration were marked as unobserved, and those crossing the bounds of an experiment were marked as censored with duration clipped by the experiment.

Model evaluation

In this part we briefly mention the measures used to quantify the model goodness-of-fit. To find the most likely number of states corresponding to the observations, the Bayesian Information Criterion (BIC) was used.[64] Our model estimates a k element initial probability vector, a $k \times k$ transition matrix, and a $k \times m$ matrix of state-wise parameters where k is the number of states and m is the number of parameters associated with observation distributions. For the phase-specific observation distributions we have a total of 6 parameters including 2 Bernoulli parameters and 2 pairs of shape and scale parameters for Gamma distribution. Since the row-sums for transition and initial probability matrices must be 1, these values are not entirely independent. From distribution analysis of the phase lengths, we realized the shape parameter of the Gamma distribution remains fairly constant over different conditions, while the scale parameter changes. Therefore, the shape parameter was shared between the populations treated with 4 different concentrations of the same compound. Each of these conditions has 2.25 free parameters (1 Bernoulli parameter, 1 scale parameter, and 1/4 shape parameter). The number of free parameters for each state when using phase-specific observations is 4.5. Altogether, the model's degrees of freedom when using phase-specific measurements are $k \times (k-1) + (k-1) + k \times 4.5$.

The Wasserstein or Kantorovich–Rubinstein metric is a measure of distance between two distributions. This metric was used to determine the difference between state emissions.[65] An analytical solution, the absolute value of the difference in distribution means, was used for the Gamma distribution.

Model benchmarking

We used emission distributions to represent the phenotypic characteristics of the cells within the lineages. To create our synthetic data we considered two possible options as our set of observations throughout an experiment. In one case, we modeled cell fate and cell lifetime, and in the second, phase-specific fate and duration. In both, we used a Bernoulli distribution for the fate outcomes and a Gamma distribution for time durations. Following is the collection of distribution parameters for the synthetic data.

Phase non-specific observations

The parameters are reflective of the cell phenotypes we observed with 5 nM lapatinib treatment. Figures S4, S5, S8, S9, and S10 are based on these parameters.

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

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

Phase-specific observations

The data in Figures 3, 4, S1, S2, and S3 were created based on the following parameters. These parameters are based on estimations from AU565 cells treated with 5 nM lapatinib.

In this table, "G1 bern" and "S/G2 bern" are the cell division probabilities at the end of G1 and S/G2 phase, respectively. The "G1 shape" and "G1 scale" are the Gamma distribution parameters of G1 phase lengths. "S/G2 shape" and "S/G2 scale" are the Gamma distribution parameters of S/G2 phase lengths.

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

Distant emissions

To create synthetic data with subpopulations of varying dissimilarity (Fig. 5), we use the phase-specific parameters set except that the values for the G1 phase Gamma shape parameter for state 1 is varied between [4, 12]. This results in an increase in the Wasserstein distance between the two cell states, allowing us to measure state assignment accuracy for different dissimilarity amounts between the two states. Likewise, for Figures S6 and S7, we simulated the overall cell lifetime duration and varied the Gamma distribution scale parameter from 1 to 8 for state 1.

Emissions for BIC figures

Figure 6 uses a unique set of parameters for the emissions matrix to simulate varying state numbers in the BIC calculation.

State	G1 bern	S/G2 bern	G1 shape	G1 scale	S/G2 shape	S/G2 scale
State 1	0.99	0.9	10	2	10	2
State 2	0.9	0.9	20	3	20	3
State 3	0.85	0.9	30	4	30	4
State 4	0.8	0.9	40	4	40	5

Acknowledgements

This work was supported by the Jayne Koskinas Ted Giovanis Foundation and by NIH U01-CA215709 to A.S.M. The authors thank Scott Taylor for his critical feedback that helped to improve the manuscript. The authors thank Ali Farhat, Adam Weiner, and Nikan Namiri for early exploratory work. **Competing financial interests:** The authors declare no competing financial interests.

Author contributions statement

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

Supplementary Figures

Performance on uncensored single lineages of increasing size

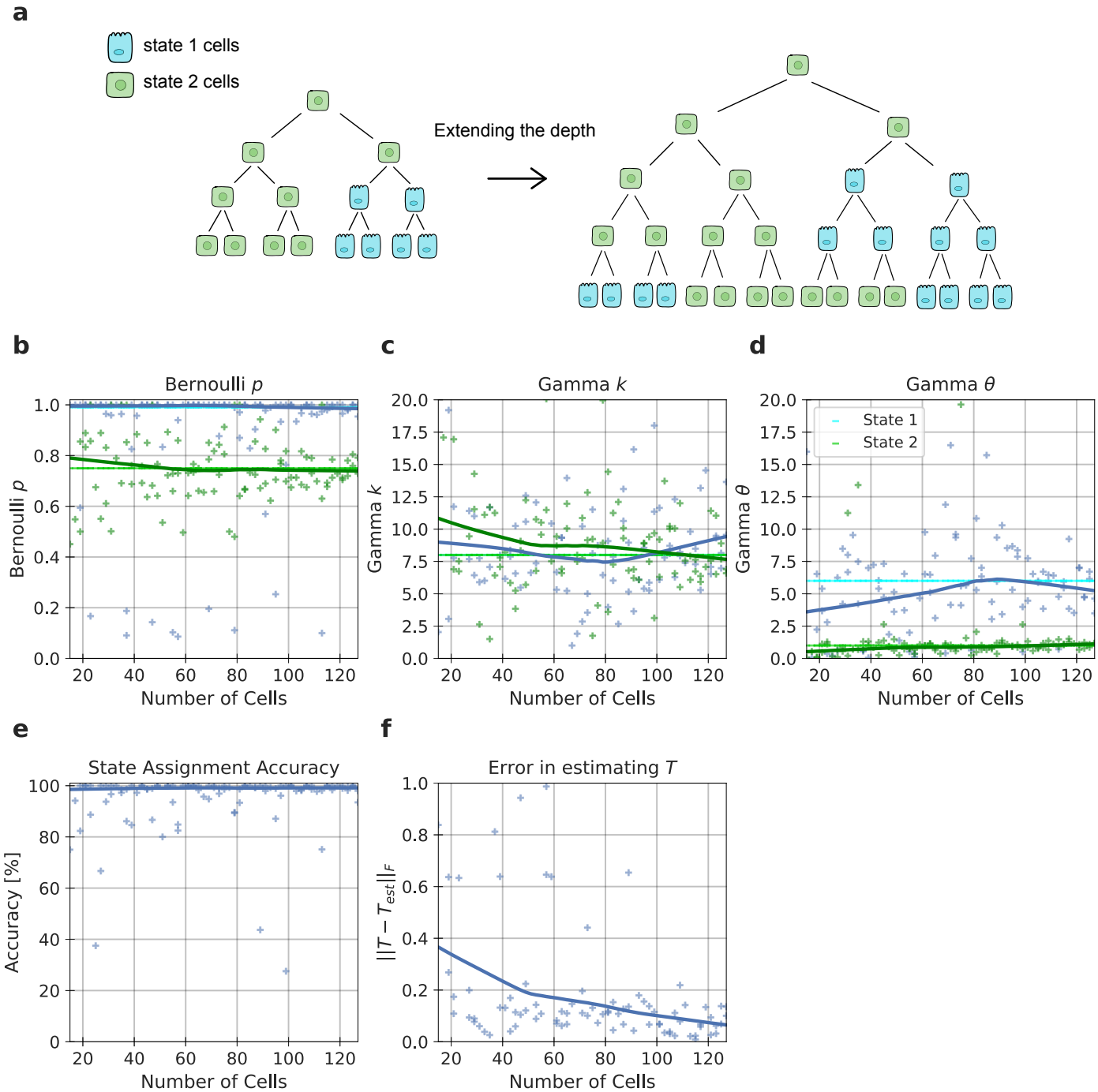

Figure S1: Performance on uncensored single lineages of increasing size. (a) Visual representation of increasing the lineage size with two states. (b) The Bernoulli parameter for states 1 and 2 as the number of cells increase. (c) The shape parameter, k, and (d) scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. (e) The state assignment accuracy as the number of cells increase. (f) The error in the estimate of the transition probability matrix, T as the number of cells increase.

Performance on uncensored lineages of increasing number

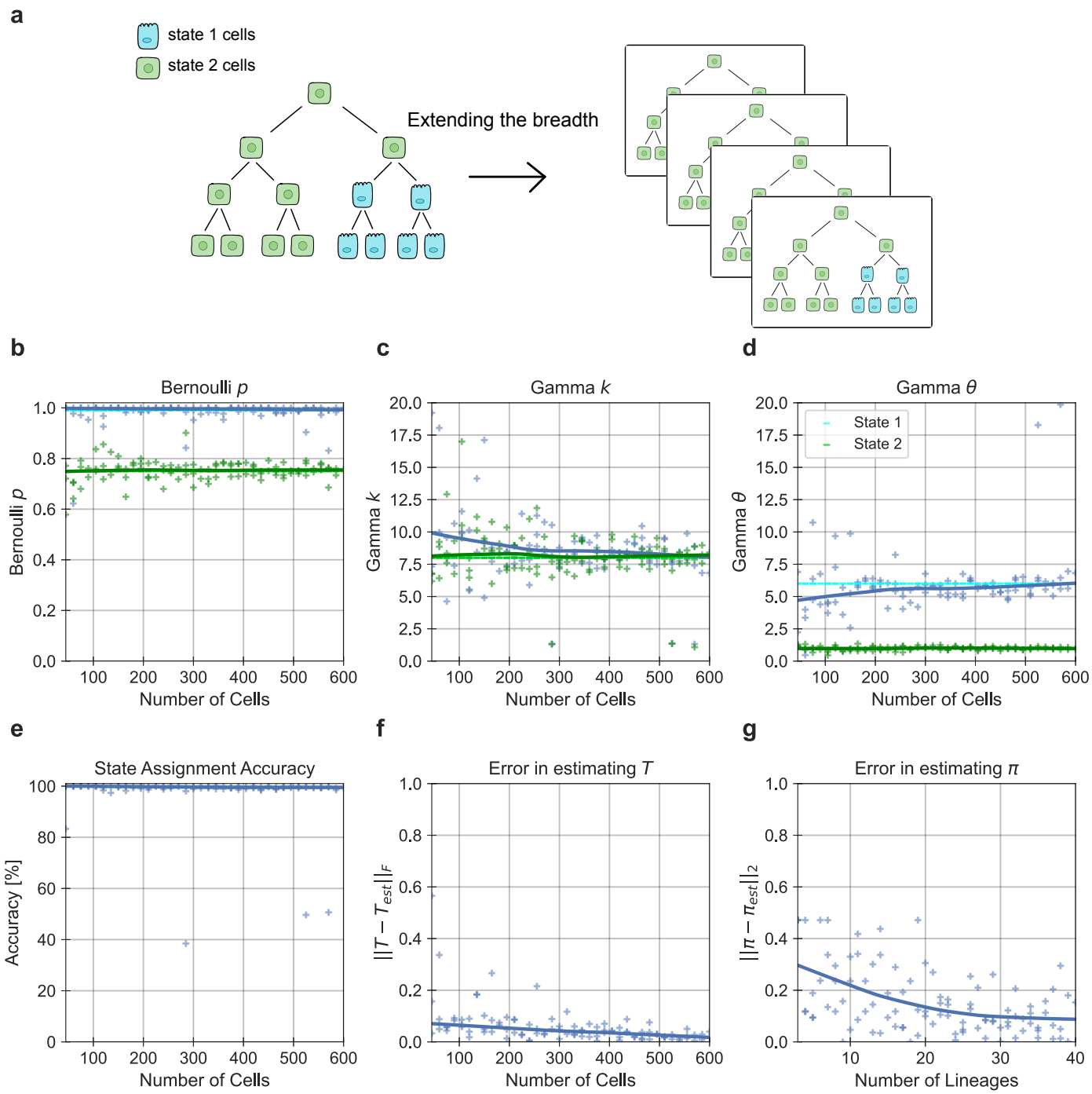

Figure S2: Performance on uncensored lineages of increasing number. (a) Visualization of increasing the number of fully-observed lineages. (b) The Bernoulli parameter for states 1 and 2 as the number of cells increase. (c) The shape parameter, k, and (d) scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. (e) The state assignment accuracy as the number of cells increase. (f) The error in the estimate of the transition probability matrix, T, as the number of cells increase. (g) The errors in the estimate of the initial probability matrix, π, as the number of lineages increase.

Performance on censored lineages of increasing number

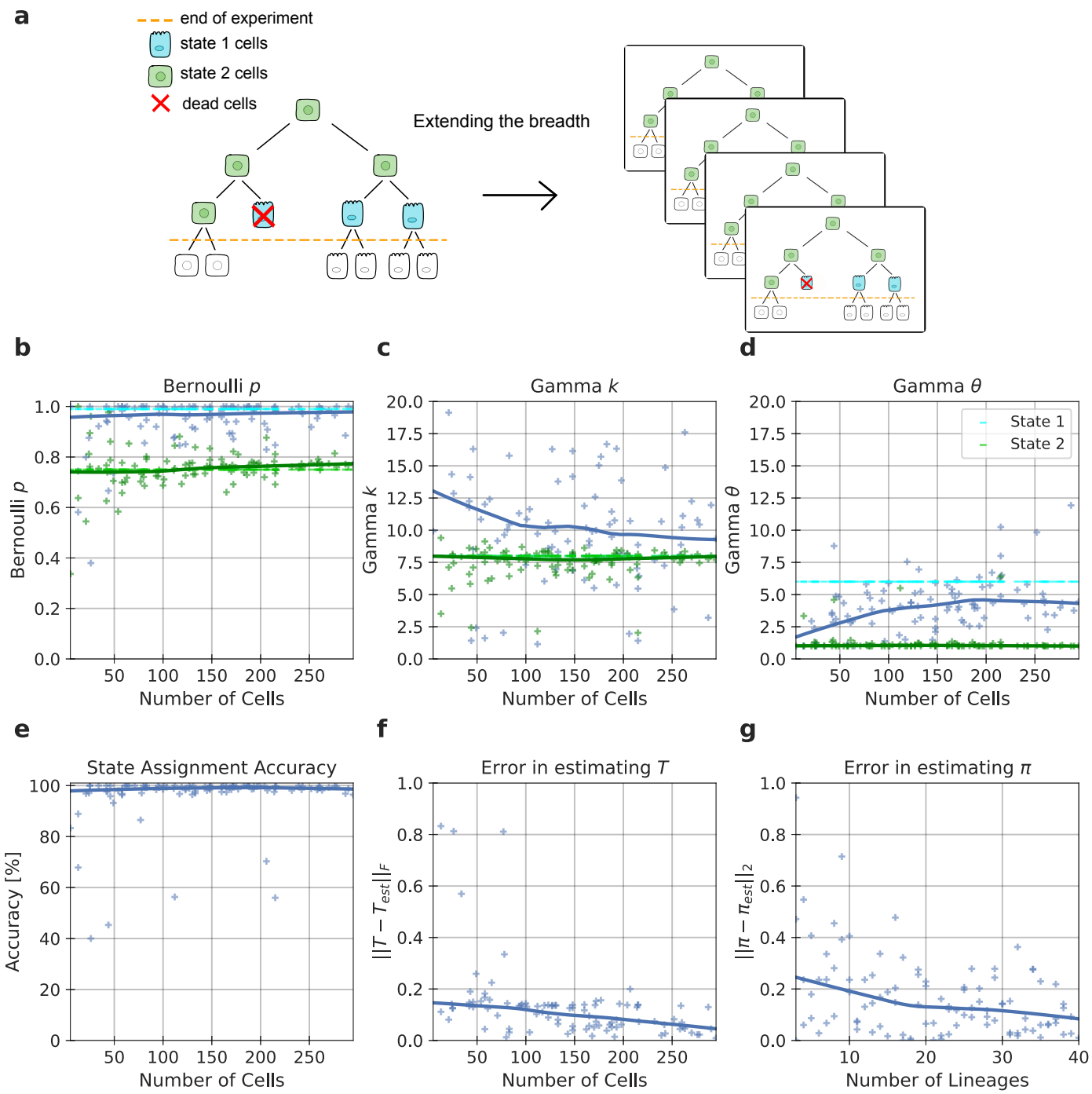

Figure S3: Performance on censored lineages of increasing number. (a) Visualization of the number of censored lineages increasing. (b) The Bernoulli parameter for states 1 and 2 as the number of cells increase. (c) The shape parameter, k, and (d) scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the number of cells increase. (e) State assignment accuracy as the number of cells increase. (f) The error in the estimate of the transition probability matrix, T, as the number of cells increase. (g) The errors in the estimate of the initial probability matrix, π, as the number of lineages increase.

Change in model performance when varying presence of a state for an uncensored lineage

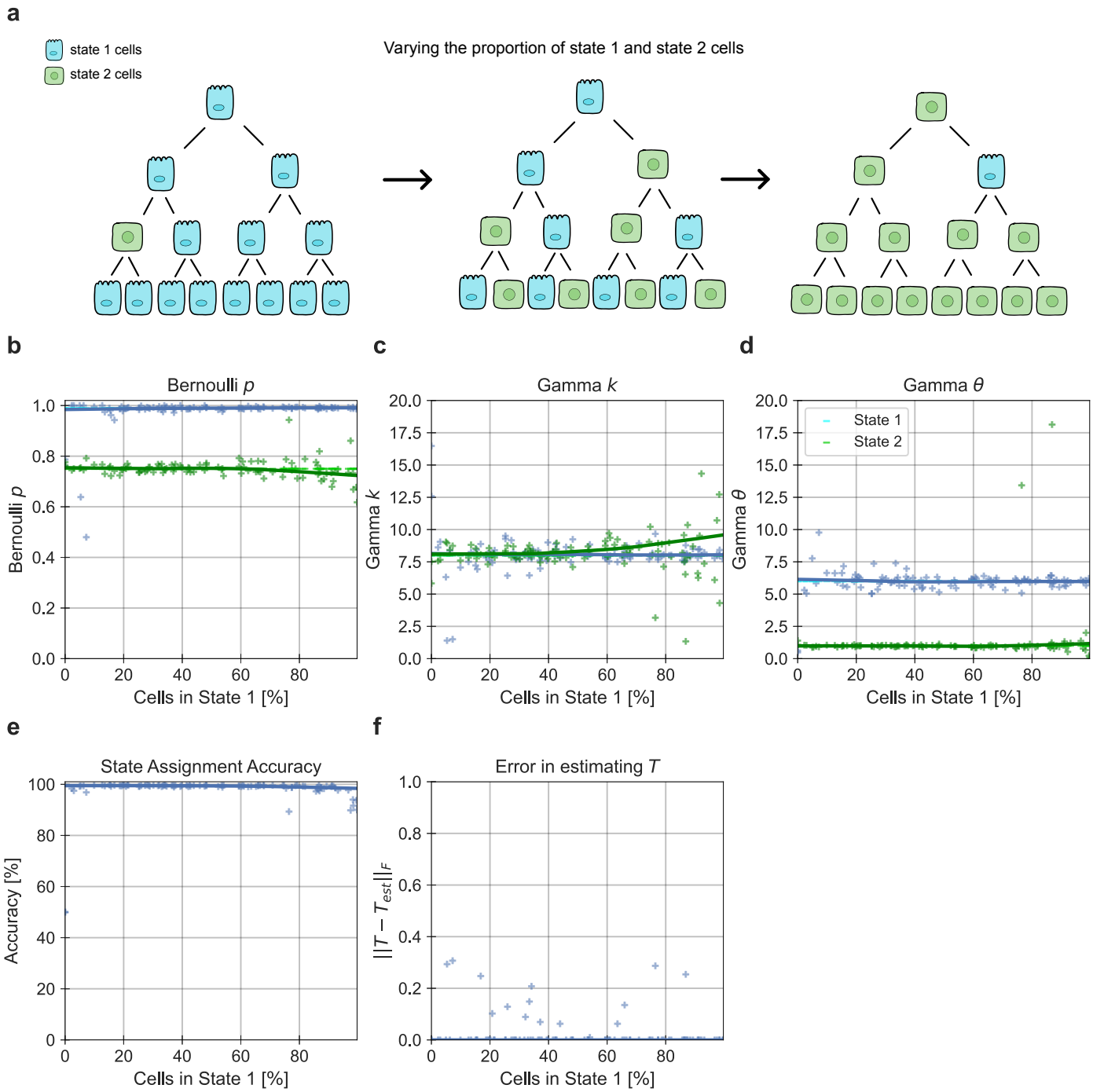

Figure S4: Model performance relative to the presence of each state for an uncensored lineage. (a) Visualization of the distribution of cells in the lineage transitioning between state 1 and state 2. (b) The Bernoulli parameter for states 1 and 2 as the proportion of cells in state 1 increase. (c) The shape parameter, k, and (d) scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the proportion of cells in state 1 increase. (e) The state assignment accuracy as the proportion of cells in state 1 increase. (f) The errors in the estimate of the transition probability matrix, T, as proportion of cells in state 1 increase.

Change in model performance when varying the presence of a state for a censored lineage

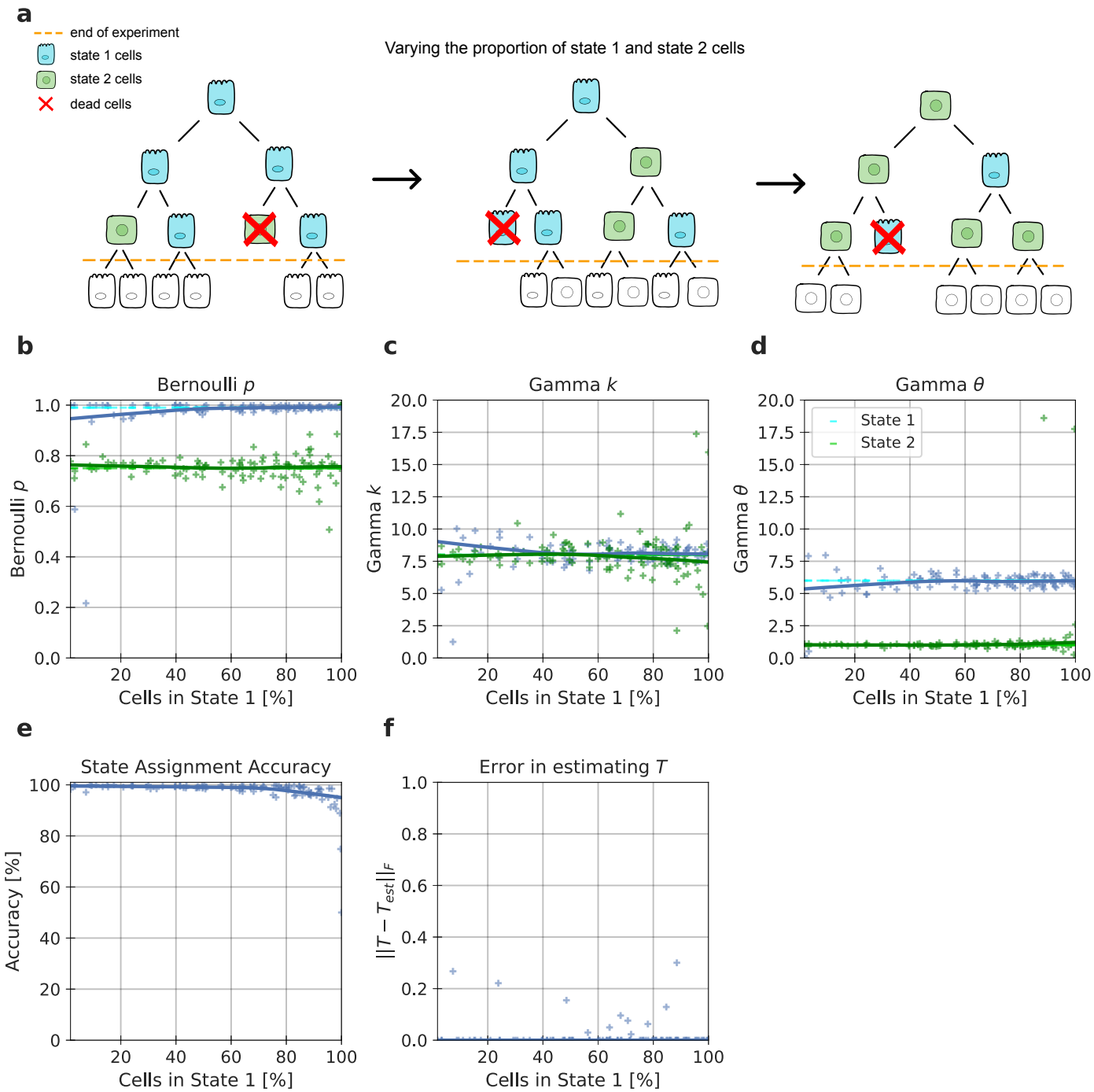

Figure S5: Change in model performance when varying the presence of a state for a censored lineage. (a) Visualization of the proportion of cells in a censored lineage transitioning between state 1 and 2. (b) The Bernoulli parameter for states 1 and 2 as the proportion of cells in state 1 increase. (c)-(d) The shape parameter, k, and scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime for states 1 and 2 as the proportion of cells in state 1 increase. (e) The state assignment accuracy as the proportion of cells in state 1 increase. (f) The error in the estimate of the transition probability matrix, T, as the proportion of cells in state 1 increase.

Change in model performance when varying state distribution similarity for an uncensored population of lineages

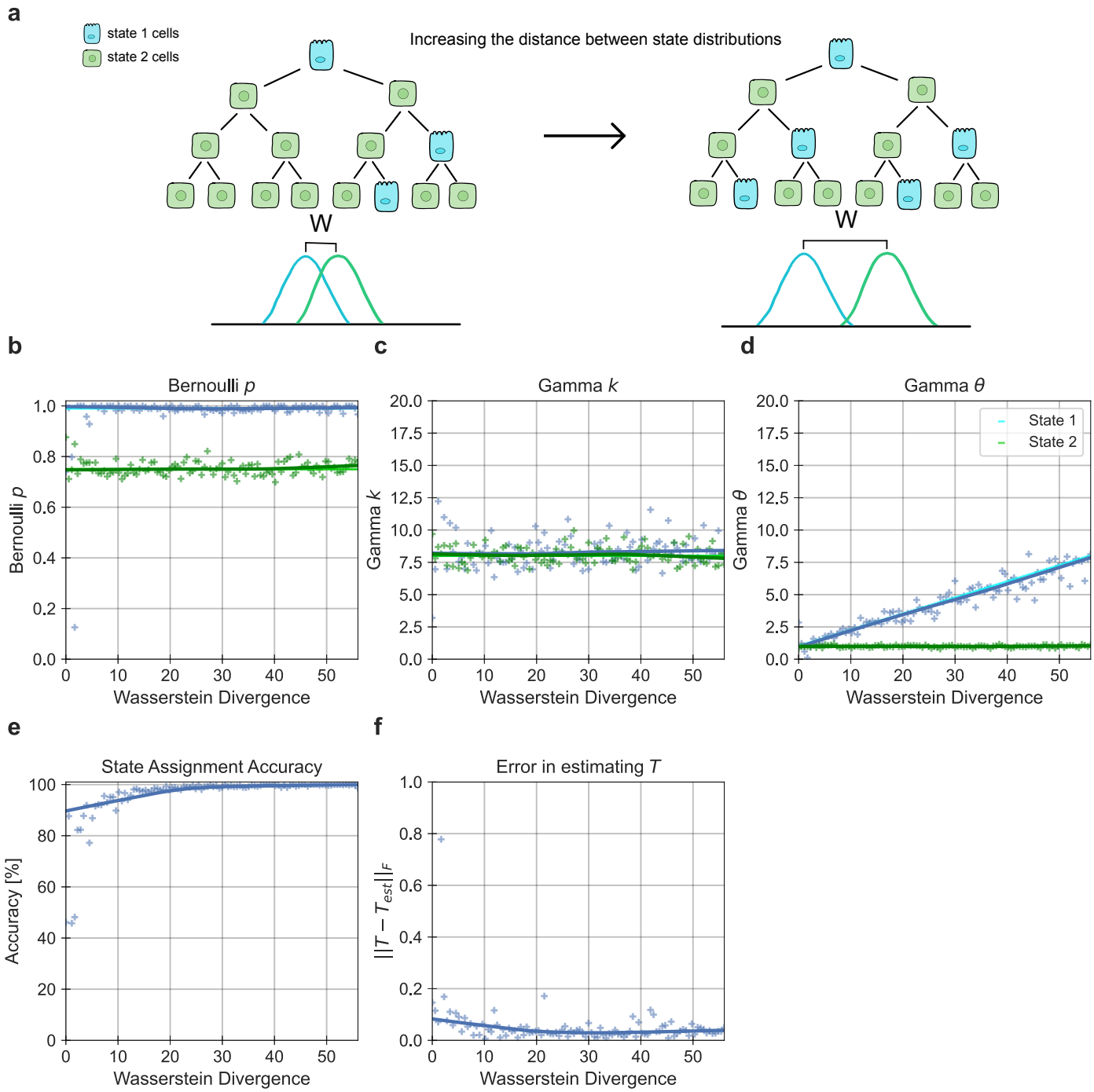

Figure S6: Change in model performance when varying state distribution similarity for an uncensored population of lineages. (a) Visualization of the Wasserstein divergence increasing as the state distribution in the lineage varies. (b) The cell fate Bernoulli parameter compared to the true value. (c) The shape parameter, k, (d) the scale parameter, θ, of the Gamma distribution corresponding to the cell lifetime compared to the true values. (e) The state assignment accuracy as the Wasserstein divergence increases. (f) The errors in the estimate of the transition probability matrix, T, as the Wasserstein divergence increases.

Change in model performance when varying state distribution similarity for a censored population of lineages

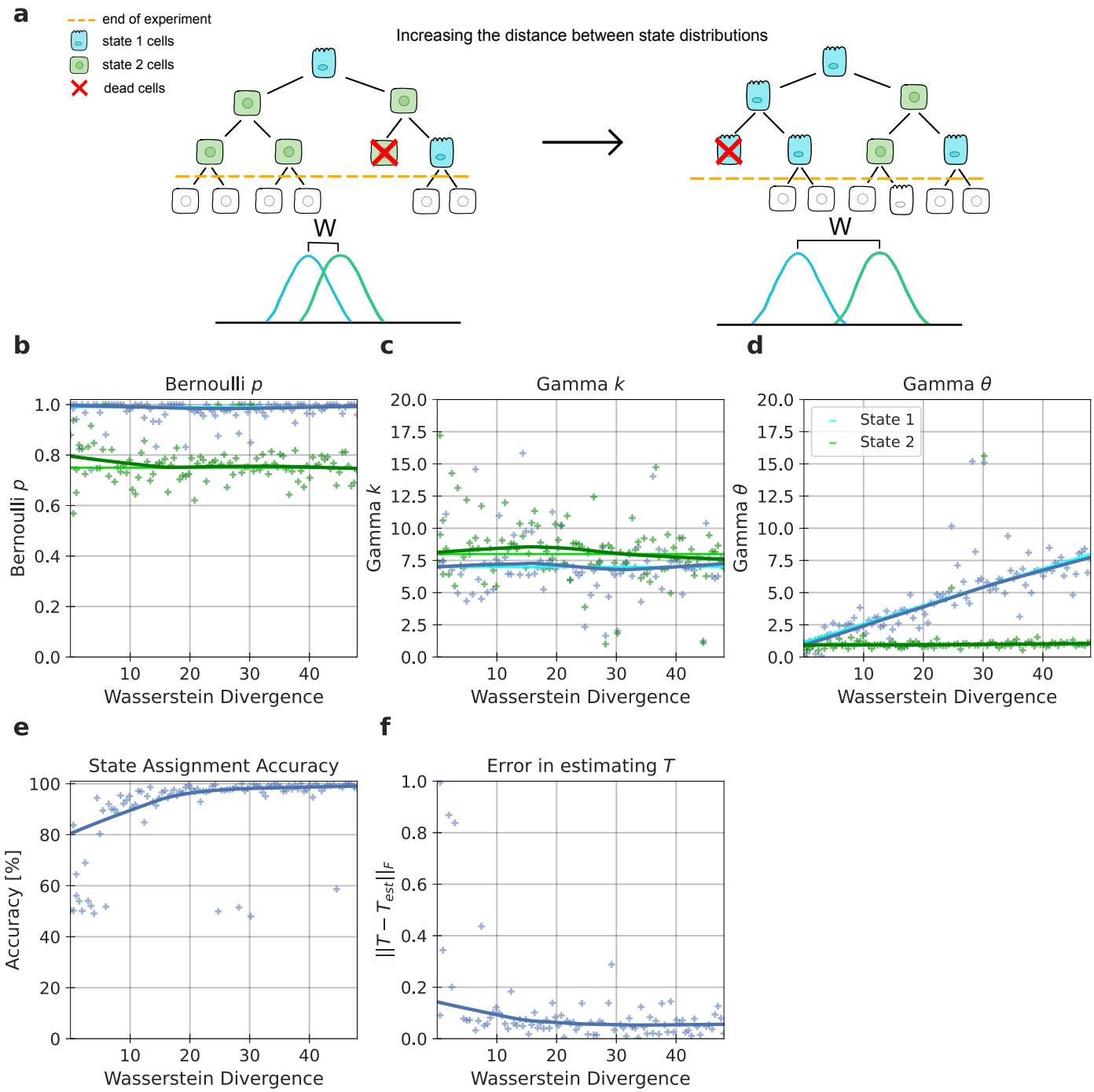

Figure S7: Change in model performance when varying state distribution similarity for a censored population of lineages. (a) Visualization of the Wasserstein divergence increasing as the state distribution in the censored lineage varies. (b) The Bernoulli, (c) Gamma shape, and (d) Gamma scale parameters for states 1 and 2 as the Wasserstein divergence increases. (e) The state assignment accuracy as the Wasserstein divergence increases. (f) The error in the estimate of the transition probability matrix, T, as the Wasserstein divergence increases.

Performance of increasing cell numbers in an uncensored single lineage given phase specific observations

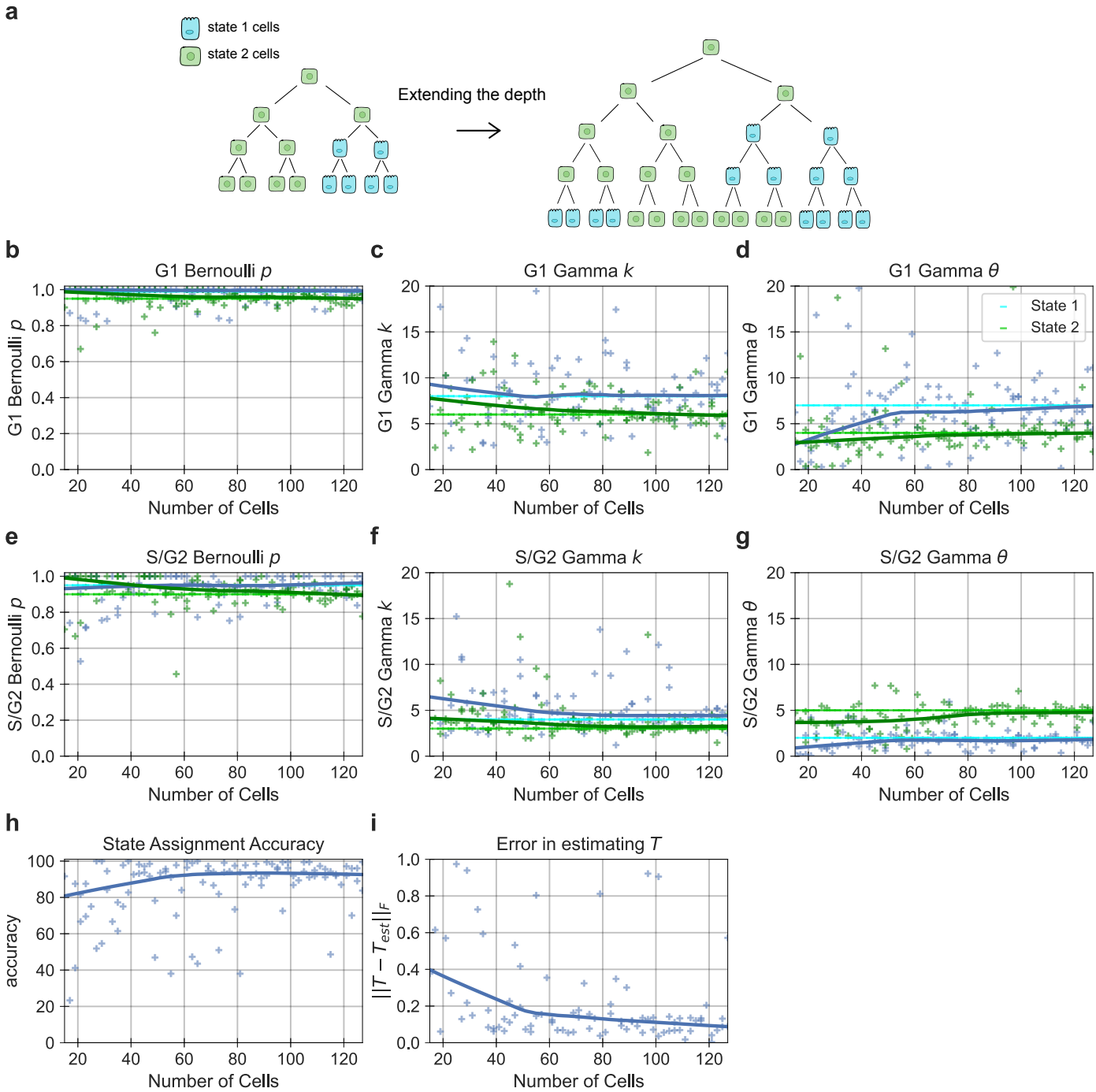

Figure S8: Performance of increasing cell numbers in an uncensored single lineage. (a) Visualization of a single lineage with increasing cell number. (b) The G1 phase Bernoulli, (c) Gamma shape, and (d) Gamma scale parameters for states 1 and 2 as the number of cells increase. (e) The S/G2 phase Bernoulli, (f) Gamma shape, and (g) Gamma scale parameter for states 1 and 2 as the number of cells increase. (h) The state assignment accuracy as the number of cells increase. (i) The errors in the estimate of the transition probability matrix, T, as the number of cells increase.

Performance of increasing lineage numbers in an uncensored population given phase specific observations

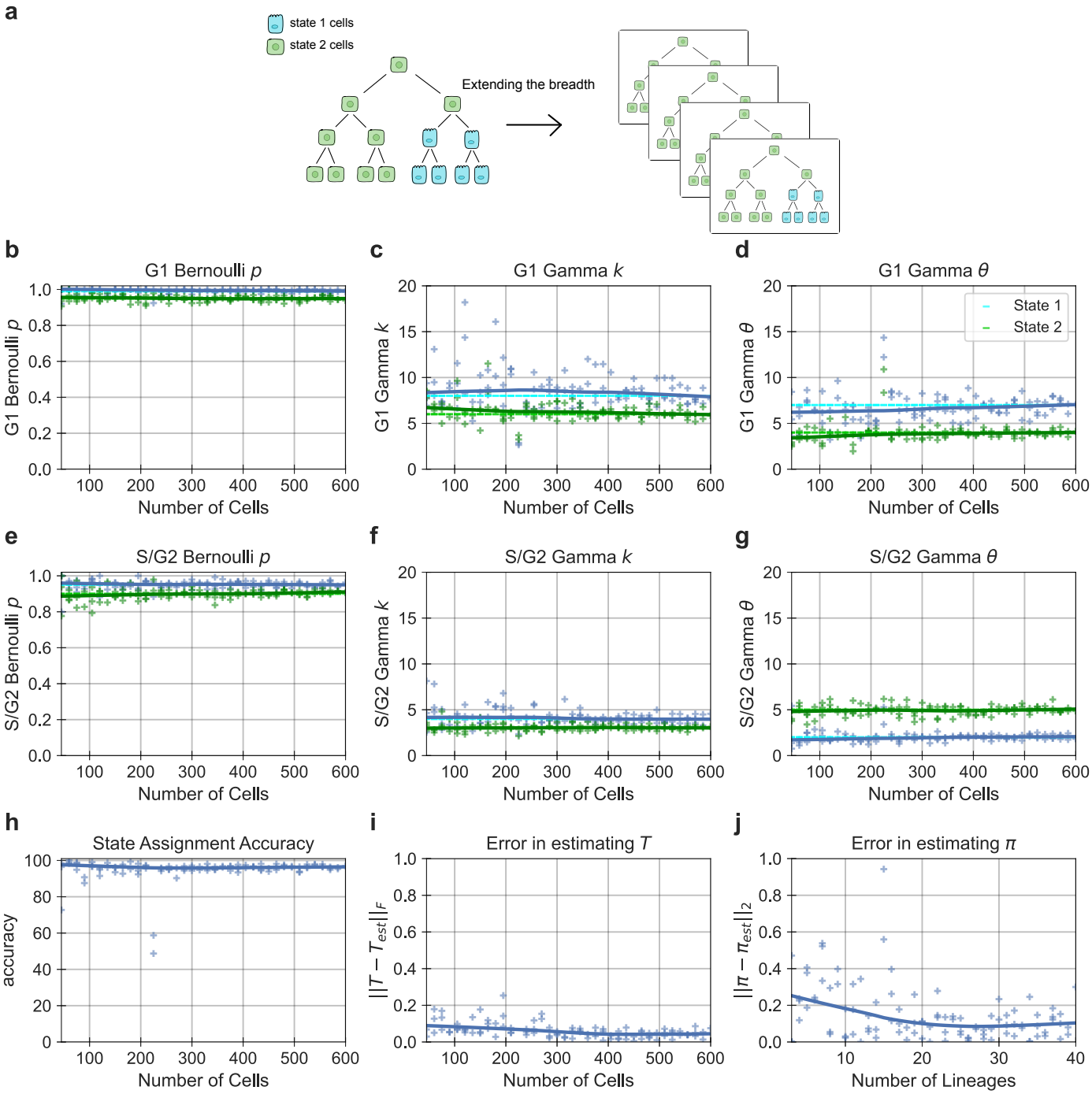

Figure S9: Performance of increasing lineage numbers in an uncensored population. (a) Visualization of the number of uncensored lineages within a population increasing. (b) The G1 phase Bernoulli, (c) Gamma shape, and (d) Gamma scale parameters for states 1 and 2 as the number of cells increase. (e) The S/G2 phase Bernoulli, (f) Gamma shape, and (g) Gamma scale parameters for states 1 and 2 as the number of cells increase. (h) The state assignment accuracy as the number of cells increase. (i) The error in the estimate of the transition probability matrix, T, as the number of cells increase. (j) The errors in the estimate of the initial probability matrix, π, as the number of lineages increase.

Performance of increasing lineage numbers in a censored population given phase specific observations

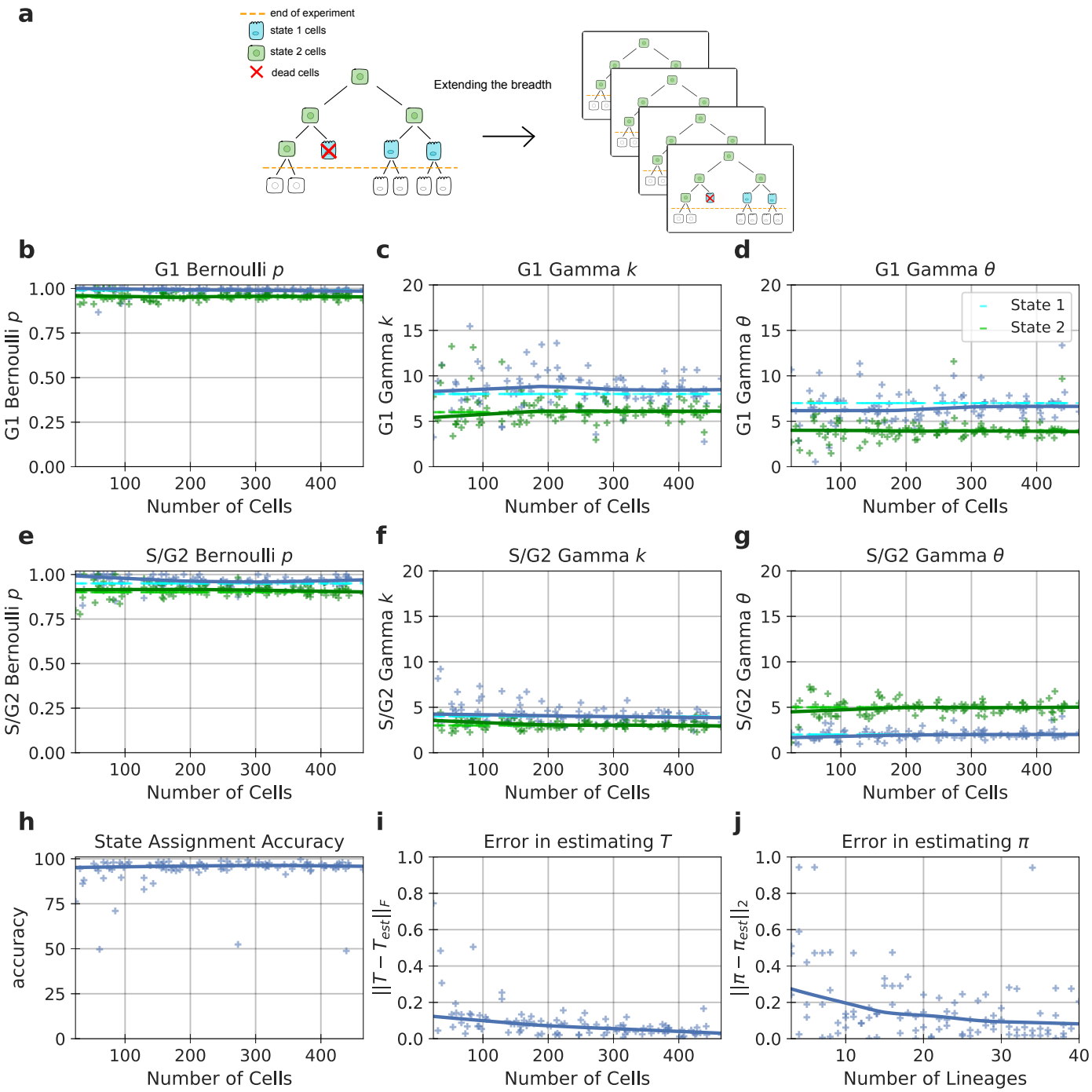

Figure S10: Performance of increasing lineage numbers in a censored population. (a) Visualization of the number of censored lineages within a population increasing. (b) The G1 phase Bernoulli, (c) Gamma shape, and (d) Gamma scale parameters for states 1 and 2 as the number of cells increase. (e) The S/G2 phase Bernoulli, (f) Gamma shape, and (g) Gamma scale parameter for states 1 and 2 as the number of cells increase. (h) The state assignment accuracy as the number of cells increase. (i) The error in the estimate of the transition probability matrix, T, as the number of cells increase. (j) The error in the estimate of the initial probability matrix, π, as the number of lineages increase.

Single cell state assignments for lapatinib-treated lineages

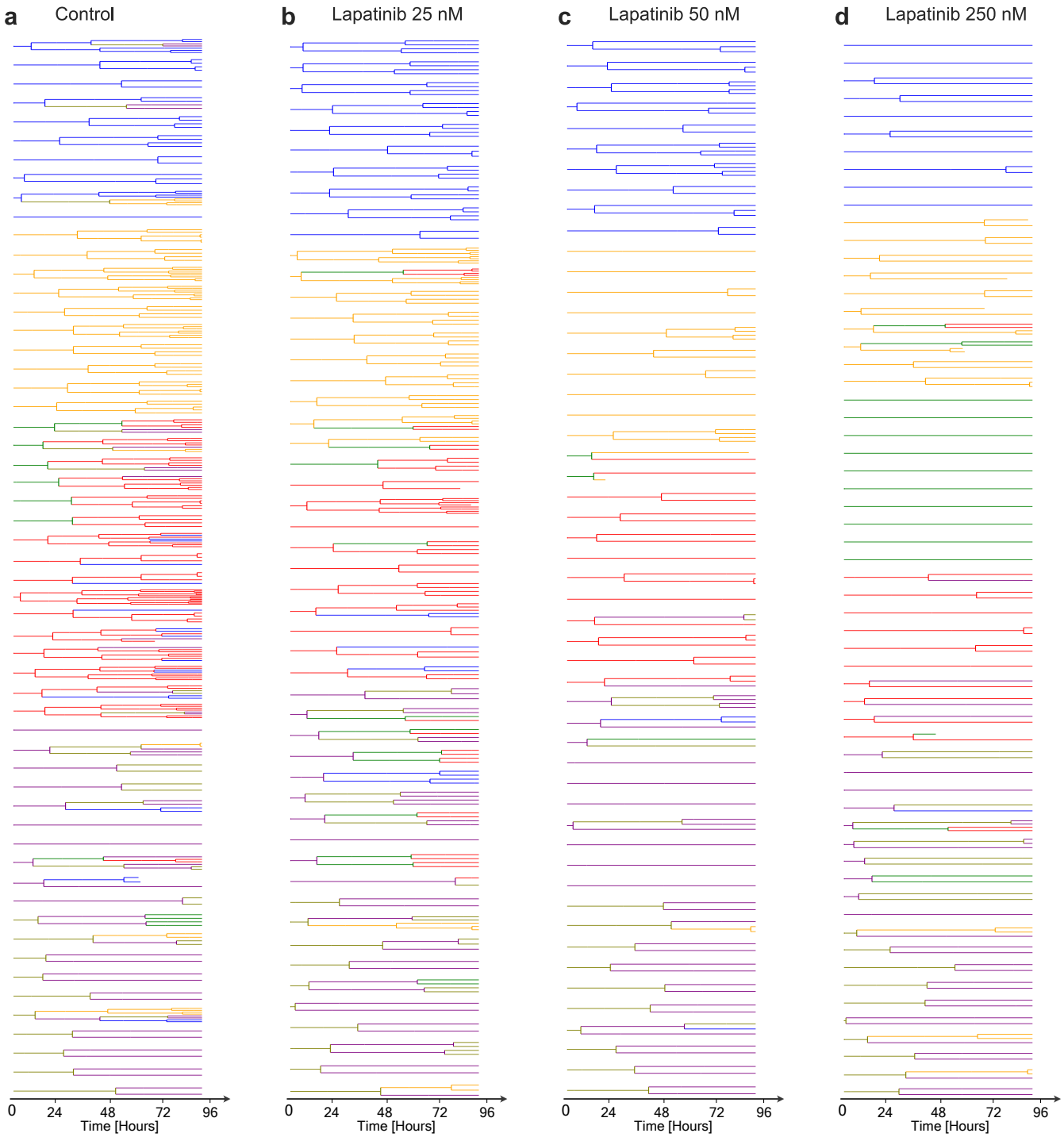

Figure S11: The single cell data after fitting and state assignment (a) Control, (b) 25 nM, (c) 50 nM, and (d) 250 nM lapatinib treatment. Each line represents a cell and the length of the line represents the cell's lifetime. G1 and S/G2 phase durations are depicted by thick and thin lines, respectively. Termination and branching indicate cell death and division, respectively. Different colors show the state of each cell. Blue: state 1, orange: state 2, green: state 3, red:state 4, purple: state 5, and olive: state6.

Single cell state assignments for gemcitabine-treated lineages

Figure S12: The single cell data after fitting and state assignment (a) Control, (b) 5 nM, (c) 10 nM, and (d) 30 nM gemcitabine treatment. Each line is a cell and the length of the line represents the cell's lifetime. G1 and S/G2 phase durations are depicted by thick and thin lines, respectively. Termination and branching indicate cell death and division, respectively. Different colors show the state of each cell. Blue: state 1, orange: state 2, green: state 3, red:state 4, and purple: state 5.

References

1. Di Maio, M. *et al.* Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. *Lancet Oncol* **6**, 669–77 (2005).

2. De Roock, W. *et al.* Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol* **11**, 753–62 (2010).

3. Sharma, S. V. *et al.* A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).

4. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–32 (2009).

5. Sigal, A. *et al.* Variability and memory of protein levels in human cells. *Nature* **444**, 643–646 (2006).

6. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics* **17**, 284–299 (2016).

7. Falkenberg, K. J. & Johnstone, R. W. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat Rev Drug Discov* **13**, 673–91 (2014).

8. Inde, Z. & Dixon, S. J. The impact of non-genetic heterogeneity on cancer cell death. *Crit Rev Biochem Mol Biol* **53**, 99–114 (2017).

9. Pisco, A. O. & Huang, S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: “What does not kill me strengthens me”. *British Journal of Cancer* **112**, 1725–1732 (2015).

10. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

11. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–7 (2012).

12. Gupta, P. B. *et al.* Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633–44 (2011).

13. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).

14. Gett, A. V., Sallusto, F., Lanzavecchia, A. & Geginat, J. T cell fitness determined by signal strength. *Nature Immunology* **4**, 355–360 (2003).

15. Arai, T. *et al.* Tumor doubling time and prognosis in lung cancer patients: evaluation from chest

films and clinical follow-up study. Japanese Lung Cancer Screening Research Group. *Jpn J Clin Oncol* **24**, 199–204 (1994).

16. Bourhis, J. *et al.* Potential doubling time and clinical outcome in head and neck squamous cell carcinoma treated with 70 GY in 7 weeks. *Int J Radiat Oncol Biol Phys* **35**, 471–6 (1996).

17. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–7 (2010).

18. Huang, D. *et al.* High-Speed Live-Cell Interferometry: A New Method for Quantifying Tumor Drug Resistance and Heterogeneity. *Anal Chem* **90**, 3299–3306 (2018).

19. Tyson, D. R., Garbett, S. P., Frick, P. L. & Quaranta, V. Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. *Nat Methods* **9**, 923–8 (2012).

20. O'Connor, J. P. B. *et al.* Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res* **21**, 249–57 (2014).

21. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).

22. Brumpton, B. *et al.* Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications* **11**, 3519 (2020).

23. Concannon, P. *et al.* Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes* **58**, 1018–22 (2009).

24. Concannon, P., Rich, S. S. & Nepom, G. T. Genetics of type 1A diabetes. *N Engl J Med* **360**, 1646–54 (2009).

25. Chen, K. *et al.* Phenotypically supervised single-cell sequencing parses within-cell-type heterogeneity. *iScience* **24**, 101991 (2021).

26. Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).

27. Kuchen, E. E., Becker, N., Claudino, N. & Höfer, T. Long-range memory of growth and cycle progression correlates cell cycles in lineage trees. *Cold Spring Harbor Laboratory* (2018). doi:10.1101/373258

28. Mitchell, S., Roy, K., Zangle, T. A. & Hoffmann, A. Nongenetic origins of cell-to-cell variability in B lymphocyte proliferation. *Proceedings of the National Academy of Sciences* **115**, E2888–E2897 (2018).

29. Choo, K. H., Tong, J. C. & Zhang, L. Recent applications of Hidden Markov Models in computational biology. *Genomics Proteomics Bioinformatics* **2**, 84–96 (2004).

30. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286

31. Yanagawa, M. *et al.* Single-molecule diffusion-based estimation of ligand effects on G protein–coupled receptors. *Science Signaling* **11**, eaao1917 (2018).

32. Crouse, M. S., Nowak, R. D. & Baraniuk, R. G. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46**, 886–902 (1998).

33. Durand, J.-B., Goncalves, P. & Guedon, Y. Computational Methods for Hidden Markov Tree Models —An Application to Wavelet Trees. *IEEE Transactions on Signal Processing* **52**, 2551–2560 (2004).

34. Choi, H. & Baraniuk, R. G. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans Image Process* **10**, 1309–21 (2001).

35. Bykova, N. A., Favorov, A. V. & Mironov, A. A. Hidden Markov models for evolution and comparative genomics analysis. *PLoS One* **8**, e65012 (2013).

36. Olariu, V. *et al.* Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics* **25**, 2824–30 (2009).

37. Nakashima, S., Sughiyama, Y. & Kobayashi, T. J. Lineage EM Algorithm for Inferring Latent States from Cellular Lineage Trees. *Cold Spring Harbor Laboratory* (2019). doi:10.1101/488981

38. Lund, B., Kristjansen, P. E. & Hansen, H. H. Clinical and preclinical activity of 2′,2′-difluorodeoxycytidine (gemcitabine). *Cancer Treat Rev* **19**, 45–55 (1993).

39. Kuchen, E. E., Becker, N. B., Claudino, N. & Höfer, T. Hidden long-range memories of growth and cycle speed correlate cell cycles in lineage trees. *eLife* **9**, e51002 (2020).

40. Bolstad, B. M. Comparing some iterative methods of parameter estimation for censored gamma data. (1998).

41. Lee, J. A. *et al.* Microbial phenotypic heterogeneity in response to a metabolic toxin: Continuous, dynamically shifting distribution of formaldehyde tolerance in Methylobacterium extorquens populations. *PLOS Genetics* **15**, e1008458 (2019).

42. van Boxtel, C., van Heerden, J. H., Nordholt, N., Schmidt, P. & Bruggeman, F. J. Taking chances and making mistakes: non-genetic phenotypic heterogeneity and its consequences for surviving in dynamic environments. *Journal of The Royal Society Interface* **14**, 20170141 (2017).

43. Tang, L., Wang, Y., Strom, A., Gustafsson, J.-Å. & Guan, X. Lapatinib induces p27 ^{Kip1} -dependent G_1 arrest through both transcriptional and post-translational mechanisms. *Cell Cycle* **12**, 2665–2674 (2014).

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

45. Zhu, X. *et al.* Autophagy stimulates apoptosis in HER2-overexpressing breast cancers treated by lapatinib. *Journal of Cellular Biochemistry* **114**, 2643–2653 (2013).

46. Lartillot, N. A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data. *Bioinformatics* **30**, 488–496 (2014).

47. Ding, J. *et al.* Reconstructing differentiation networks and their regulation from time series single-cell expression data. *Genome Research* **28**, 383–395 (2018).

48. Tentner, A. R. *et al.* Combined experimental and computational analysis of DNA damage signaling reveals context-dependent roles for Erk in apoptosis and G1/S arrest after genotoxic stress. *Mol Syst Biol* **8**, 568 (2012).

49. Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med* **22**, 262–9 (2016).

50. Bhang, H.-e. C. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Medicine* **21**, 440–448 (2015).

51. Barney, L. E. *et al.* Tumor cell–organized fibronectin maintenance of a dormant breast cancer population. *Science Advances* **6**, eaaz4157 (2020).

52. Burke, R. T. & Orth, J. D. Through the Looking Glass: Time-lapse Microscopy and Longitudinal Tracking of Single Cells to Study Anti-cancer Therapeutics. *Journal of Visualized Experiments* 53994 (2016). doi:10.3791/53994

53. Han, K. *et al.* CRISPR screens in cancer spheroids identify 3D growth-specific vulnerabilities. *Nature* **580**, 136–141 (2020).

54. Schwartz, A. D. *et al.* A biomaterial screening approach reveals microenvironmental mechanisms of drug resistance. *Integrative Biology* **9**, 912–924 (2017).

55. Emert, B. L. *et al.* Variability within rare cell states enables multiple paths toward drug resistance. *Nature Biotechnology* (2021). doi:10.1038/s41587-021-00837-3

56. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18**, 35–45 (2017).

57. Gross, S. M., Sanchez-Aguila, C., Zhan, P. J. & Heiser, L. M. Single cell tracing reveals heterogeneous drug-, dose-, and time-dependent effects on cancer cell fates. *Cold Spring Harbor Laboratory* (2020). doi:10.1101/2020.07.24.219907

58. Meijering, E., Dzyubachyk, O. & Smal, I. Methods for Cell and Particle Tracking. in *Methods in Enzymology* (Elsevier BV, 2012). doi:10.1016/b978-0-12-391857-4.00009-4

59. Ephraim, Y. & Merhav, N. Hidden Markov processes. *IEEE Transactions on Information Theory* **48**, 1518–1569 (2002).

60. Levinson, S. E., Rabiner, L. R. & Sondhi, M. M. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technical Journal* **62**, 1035–1074 (1983).

61. Devijver, P. A. *Baum's forward-backward algorithm revisited*. *Pattern Recognition Letters* **3**, (Elsevier, 1986).

62. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

63. Minka, T. P. *Estimating a Gamma distribution*. (Unpublished, 2002).

64. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics* **4**, 199–203 (2011).

65. Vallender, S. S. Calculation of the Wasserstein Distance Between Probability Distributions on the Line. *Theory of Probability & Its Applications* **18**, 784–786 (1974).