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

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

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

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

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

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

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

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

Abstract

Cell plasticity operates alongside other sources of cell-to-cell heterogeneity, such as genetic mutations and variation in signaling, together preventing 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 lineage tree-based adaptation of a hidden Markov model that employs single cell lineages as input to learn the characteristic patterns of single cell phenotypic heterogeneity and state transitions in an unsupervised fashion. To benchmark our model, we paired cell fate with either cell lifetimes or individual cell cycle phase lengths on synthetic data and demonstrated that the model successfully classifies cells within experimentally tractable dataset sizes. As an application, we analyzed experimental measurements of same measurements in cancer and non-cancer cell populations under various treatments. We find that in each case multiple phenotypically distinct states exist, with significant heterogeneity and unique drug responses. In total, this framework allows for the flexible classification of single cell heterogeneity across lineages.

Introduction

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

Large-scale profiling studies can find molecular features that associate with drug response using population-level samples[10,11]. These associations, while valuable, can miss the contribution of cell-to-cell heterogeneity, and especially stochastic changes in individual cell states that 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 role of individual cells in the overall population response[20]. Though population heterogeneity is usually defined through molecular measurements, studies that have explicitly linked molecular and phenotypic variation have been able to identify mechanisms that underly cell-to-cell variation that would otherwise remain hidden[21], and studies starting with phenotypic analysis have generally found that phenotypic variability arises from a small number of molecular factors leading to the phenotypic variation[4,22,23].

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

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[31]. These models derive their relative simplicity by assuming a Markov process, meaning that the current behavior of a system can be assumed to be independent of its earlier history should its current state be known. This assumption naturally applies in many contexts. In the case of cells, this assumption aptly captures cell inheritance because daughter cells inherit both molecular signals and their environment from their predecessor. Indeed, several recent examples of cell-to-cell inheritance mechanisms can be represented as a Markov process through linear chains or cycles of states[12,22,23]. HMMs have been adapted to lineage trees (tHMMs) so that each measurement across the tree can similarly provide accumulating evidence for a prediction. Just like with time-series data, these models can provide very accurate predictions despite noisy measurements and limited information by recognizing the co-dependence between related measurements[32,33]. tHMMs have been used in a multitude of applications, from image classification to comparative genomics[34,35]. 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 of these models still depends on more easily usable implementations, examples of successful tHMM-based discoveries, and standards for experimental application.

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

Results

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

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

their relatives in each lineage. These observations demonstrate some of the unique advantages of collecting lineage-based measurements.

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

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

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

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

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

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 the

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

To correct for this effect in our model, we marked cells that encountered the start or end bounds of the experiment. When estimating the properties of these cells' lifetime we instead used a censored estimator or the survival function of the distribution[39]. Using synthetic data, we verified that this correction resulted in accurate phenotype estimations (Figs. 3d, S3, S10). Thus, accounting for cells that outlive the bounds of the experiment through a censored estimator removes the contribution of this experimental confounder.

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. Cells in state 0 and 1 are shown in green and blue. **(c)** State assignment accuracy with censored lineages using an uncorrected model. **(d)** State assignment accuracy with censored lineages using the corrected model. Each scatter point represents the state assignment accuracy of the model when fit to a lineage with the indicated number of cells. The solid lines show the Lowess trendline of the individual run accuracies. 100 trials are plotted.

Synthetic lineage benchmarks show a tHMM can accurately infer population behavior

To evaluate how accurately a tHMM model could infer the behavior of multi-state cell populations, we used synthetic populations of cells in a wide variety of configurations, such as various populations sizes, numbers of states, and abundance of the states. In each case, we determined that the tHMM model could accurately infer the hidden states and parameters of a population given at least 100 cells. This synthetic data included uncensored (Figs. S1, S2, S8, S9, Table 1-2) or censored (Figs. 4, S10, S3, S15; Table 1–3) situations. Synthetic data were created by lengthening the simulated experiment time, in effect creating "deeper" lineages, or by increasing the number of initial cells to have a greater number of lineages, increasing the experiment's "breadth". In addition to varying the number of cells in a population, we benchmarked populations with varied cell state percentages (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, and generally provided accurate results with datasets consisting of at least 10 lineages, 100 cells overall, and 10 cells from each state.

More specifically, one of the benchmarking studies we performed was with data matching our measurements of AU565, where G1 and S/G2 phase durations represented by a gamma distribution, and their corresponding cell fate represented by a Bernoulli distribution, were quantified (Fig. 4). The choice of the gamma distribution for cell cycle phase was inspired by a previous study[40] and verified by evaluating a variety of distributions; the gamma distribution fit the cell lifetime data best. Although the tHMM model was fit with no information about the true

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

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

Lineage information improves cell state identification with heritable phenotypes

Cells of even very distinct molecular states can have partly overlapping phenotypes due to non-heritable variation. Therefore, we sought to evaluate how different two states need to be for us to accurately identify them as distinct (Fig. 5a). We varied the G1 phase duration of two states from identical to very distinct (Fig. 5b) and quantified the state assignment accuracy of our model (Fig. 5c). While the phenotypic observation of a given state had to be different for our model to accurately assign cells, even moderately overlapping phenotypes (Wasserstein distance of ~20) could be distinguished by using the lineage relationships of cells. As a baseline comparison, we analytically identified the optimal classifier in the absence of lineage information (see Methods). The tHMM consistently outperformed this approach (Fig. 5c). The model performance in censored and uncensored populations were similar (Fig. S6, S7). This shows that lineage relationships can be used to identify cell states more accurately with partially overlapping phenotypes.

Figure 5: Model performance versus the difference between states. (a) Cartoon of how two states can vary in their phenotypic similarity, in a synthetic population of two states. On the top, cells might be virtually indistinguishable (here based on shape). On the bottom, they might be so different that looking at one cell is sufficient to identify its state. **(b)** The distribution of G1 duration is varied in state 1 (blue) while the other state is kept constant. **(c)** State assignment accuracy versus the Wasserstein distance between state phenotypes. Each point represents the accuracy of state assignment for a lineage created by a set of parameters that yield the shown

Wasserstein distance between the two state distributions. 100 trials are plotted. Either the tHMM model (blue) or an optimal classifier without lineage information (orange) was used. The solid lines show a Lowess trendline of the model accuracy.

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 included approximately 250 to 650 cells with known cell phase fate and phase lengths (Table 3). The inferred number of cell states was consistently 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 synthetic uncensored lineages with 1–4 states. **(e-h)** Model BIC for synthetic censored lineages with 1–4 states. BIC values are normalized such that the optimum is equal to 0. The minimum BIC value corresponds to the predicted number of states in each repetition. 5 trials plotted.

tHMM infers multiple distinct subpopulations in experimental drug response data

As an application of our model, we used phenotypic measurements from two cell lines. With the first, AU565, we measured of the G1 and S/G2 phase durations and terminal cell fates of cells in a control condition and when treated with 3 concentrations of gemcitabine or lapatinib. For the second, MCF10A, we measured the overall cell lifetimes and terminal fates of cells treated with PBS or single concentrations of the growth factors EGF, HGF, or OSM. Cells were imaged every 30 minutes and then tracked over time to assemble lineage relationships. The lapatinib and gemcitabine treated AU565 populations (including control) contained a total of 5290 and 4537 cells, respectively. The MCF10A population contained 1306 cells. Lineages included 1–5 generations of cells. The model was then fit to each experiment's data across all conditions, enforcing that the initial and transition probabilities are shared across concentrations but allowing the phenotype distributions to vary. We enforced a unidirectional phenotypic shift with drug concentration in AU565 cells, reflecting the expectation of a dose-response effect on cell phenotype within each state. The cell fate parameters were estimated without constraints. We assumed the number of states is shared across drug concentrations in AU565 cells and across growth factor treatments in MCF10A cells. To determine the number of cell states, we compared models of 1–7 states using the BIC, where the lowest BIC value across numbers of states indicates the most optimal model correcting for complexity (Fig. 7a–c). The data for each compound indicated the presence of multiple inherited states.

To verify the model's predictive ability, we additionally implemented a cross-validation scheme for the lineage data. Briefly, roughly 20% of the cells were chosen at random and then masked from the fitting process. The model parameters were estimated using only the unmasked cells, though all cells received state assignments through use of their relatives. At the end, the log-likelihood of the masked cells' observations were evaluated using the fit model. We tested this cross-validation approach by creating synthetic cell populations of 2–5 true states with conditions matching the experimental data. For each scenario, we were able to identify the correct number of states based on which gave the highest log-likelihood (Fig. S16a–d, Table 3). Cross-validating the experimental data again confirmed the 4 and 5 phenotypic states within the lapatinib and gemcitabine data, respectively (Fig. S16e,f). It also directly demonstrated that the inclusion of multiple states enables the tHMM model to predict unseen data, and that this prediction is dependent on inheritance; a no-inheritance model in which all transitions were equally likely performed relatively poorly.

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

Lapatinib response is defined by both stable and cyclical states

We fit the lapatinib-treated data to the model with 4 states based on our BIC-based model selection, confirmed by cross-validation (Fig. 7a, S16f). Fitting revealed states of widely varying persistence over generations, from less than a 0.01 probability of remaining in state 2 to a 0.94 probability of remaining in states 1 and 3 (Fig. 8a). Interestingly, states 2 and 4 formed a cycle wherein the most probable transition was between these two states (Fig. 8a, S11).

Examining the phenotypes of each state revealed distinct drug responses. Lapatinib is an EGFR/HER2 inhibitor that induces cell cycle arrest in G1 phase[43]. Every state displayed a dose-dependent increase in G1 phase lifetime with lapatinib treatment, and G1 effects were more pronounced as compared to those involving S/G2 (Figs. 8b–i, S11). While the probability of survival at the end of the cell cycle phase decreased at higher concentrations, very few cell death events were observed (Figs. 8h–i, S11). Consequently, the chances of cell death likely have high uncertainty at higher concentrations of lapatinib. States 2 and 4 were highly arrested in both G1 and S/G2 phase; in contrast, states 1 and 3 experienced little arrest in G1 and no arrest in S/G2 phase (Fig. 8f–g). Thus, cell states seemed to be primarily distinguished based on the degree of lapatinib response. The cycle between states 2 and 4 seems to reflect the observation that cells more highly arrested in G1 than G2/S give rise to cells that spend longer in G2/S than G1, and vice versa (Fig. 8, S11).

Figure 8: Lapatinib response is defined by phenotypically distinct stable and cyclical states. **(a)** State transition graph showing the probability of state transitions among the predicted states. Transitions with less than a 0.03 probability have been removed. **(b–e)** A sample of lineage trees after fitting the model and state assignment (control, 25 nM, 50 nM, and 250 nM). **(f–g)** The log10 of fit mean time of G1 and S/G2 phase durations for different concentrations. **(h–i)** The

Bernoulli parameter, indicating the probability of G1-to-S phase transition versus cell death **(h)**, and the probability of division versus cell death **(i)** for each concentration.

Gemcitabine-treated populations are clustered into phase-specific responses

Gemcitabine, as mentioned previously, is a chemotherapy agent that induces cell cycle arrest and apoptosis in S phase by disrupting DNA repair. The AU565 cells were treated with 5, 10, and 30 nM of gemcitabine; model selection, confirmed by cross-validation, inferred 5 states in the population (Figs. 7b, S16a/g). Examining the 5-state fit revealed relatively stable states 1, 3, and 4 (Fig. 9a–e, S12). States 2 and 5 formed a cycle with high rates of interconversion between one another.

The phenotypic effects of drug treatment were evident in both cell cycle phases, though the most affected phase differed between states (Fig. 9f–i). State 5 showed S/G2-specific arrest and always resulted in cell death at the highest concentration (Fig. 9g, i, S12). Meanwhile, cells in state 4 grew almost normally, with some cell death in G1 at the highest concentration (Fig. 9f–i). At the highest concentration, state 3 represents the cells arrested at S/G2 that have not divided even once, and state 5 is the representative of almost all of those undergoing cell death at S/G2 (Fig. 8i, S12).

Lastly, we wished to explore whether the phenotypic heterogeneity we observed was limited to cancer cells or cytotoxic drug treatment. To determine this, we tracked non-tumorigenic MCF10A breast cells. These cells are normally grown in the presence of epidermal growth factor (EGF); we compared this condition to growth factor withdrawal (PBS) or rescue with hepatocyte growth factor (HGF) or oncostatin M (OSM)[44]. Each growth factor consistently promoted proliferation on a population level compared to the PBS control, though with considerable inter- and intra-lineage variation (Fig. S13). BIC-based model selection inferred the presence of 3 distinct states (Fig. 7c). Inspecting the model revealed generally more dynamic transitions between states as compared to the AU565 experiments (Fig. S14). Due to the lack of growth factors, most cells arrested in the PBS condition; the lack of cells completing their lifetime is the reason for the division probability being 0.5 (Fig. S14g). State 1 was distinct in being relatively less responsive to HGF and OSM treatments (Fig. S14a/f), while state 1 displayed higher rates of cell death overall (Fig. S14a/g). In total, the tHMM model was effective in identifying subsets of cells with divergent phenotypic responses to drug treatment alongside the relationships between cells in the population.

Figure 9: State-specific emissions of the gemcitabine-treated data. (a) State transition graph showing the probability of state transitions among the predicted states. The transitions with less than a 0.03 probability have been removed. **(b–e)** A sample of lineage trees after fitting the model and state assignment (control, 5 nM, 10 nM, and 30 nM). **(f–g)** The log10 fit mean time of G1 and S/G2 phase durations for different concentrations. **(h–i)** The Bernoulli parameter, indicating the probability of G1-to-S phase transition versus cell death **(h)**, and the probability of division versus cell death **(i)** for each concentration.

Discussion

Heterogeneity and plasticity in cancer cells enables them to adapt in response to therapy. Even in the absence of genetic mutations, other heritable variation serves as a substrate for selection[45,46].

375 In this paper, we introduced a tree-based hidden Markov model that clusters single cells from a
376 heterogeneous population based solely on their phenotypic traits and relationships. Model
377 benchmarking showed that it can provide accurate results using feasible experimental designs.
378 Of particular importance, the tHMM showed good sensitivity to subpopulations at lower
379 frequencies (Figs. S6, S7). Comparing the model to more standard clustering, the tHMM showed
380 that lineage information helps to identify cell states more accurately (Fig. 5). Using cross-
381 validation, we were able to show that accounting for cell inheritance allowed the model to
382 accurately predict unseen observations (Fig. S16). Several critical advancements in the current
383 work are a modular interface for using tHMM models with various phenotypes (Fig. 2), proper
384 censorship handling (Fig. 3), strategies for model evaluation (Figs. 7 & S16), and demonstrating
385 that such a model can be applied to study cancer heterogeneity at baseline and in response to
386 perturbation.

387 We used single cell lineage tracking data of AU565 cancer cells treated with lapatinib and
388 gemcitabine as a demonstration of the model. G1 and S/G2 cell cycle phase durations and cell
389 fate measurements were used as relevant cell phenotypes to quantify the anti-cancer effects of
390 these drugs. We were able to identify 4 and 5 distinct subpopulations within the lapatinib and
391 gemcitabine-treated data, respectively (Fig. 7). The phenotypic features of each state were
392 quantified in parallel (Figs. 8, 9). Lapatinib is known to inhibit cell proliferation by inhibiting
393 Akt/mTOR pathway activity, which is a key regulator of G1 phase progression[47]. Similarly, our
394 analysis in the lapatinib-treated population indicated that cells, regardless of their state,
395 experienced a prolonged G1 phase, but individual states varied in their susceptibility. In
396 gemcitabine treated cells, we observed that most states were highly heritable, with more varied
397 phenotypic effects. This included cells that became arrested in S/G2 and underwent apoptosis
398 (state 5), cells that were selectively arrested in G1 (state 1), and cells that hardly responded to
399 drug treatment at all (state 4; Fig. 9). While gemcitabine canonically works by inducing cell
400 arrest in G2/S, previous work has characterized its effects on G1 phase by separating the effects
401 on both cell cycle phases[49]. They similarly identified that G1 arrest was associated with cell
402 death, which is also evident in cells of state 3 where G1 arrest is seen alongside cell death in both
403 phases (Fig. 9f–i). Our results would further suggest that those cells with G1 effects are
404 molecularly and heritably distinct from those that are arrested in S/G2. MCF10A cells with
405 growth-factor-induced proliferation showed a very distinct pattern of variation, suggesting that
406 the phenotypic cell states identified by the model reflect a confluence of cell features and
407 treatment conditions (Fig. S13 & S14).

408 We present several lines of evidence supporting the accuracy of the model and existence of
409 heritable cell states. First, across a diverse array of benchmarking experiments, we show that the
410 model can derive accurate conclusions from synthetic data with properties like those we find in
411 the experimental measurements (e.g., Fig. 4). Through an informatic model selection scheme, we
412 find statistical evidence for the existence of multiple states (Figs. 6 & 7). Examining these cell
413 states, we find patterns consistent with the biological mechanisms of the compounds we used to
414 alter cell proliferation (Figs. 8 & 9). Reassuringly, we were able to confirm that the abundance of
415 cell states was consistent across experimental replicates, ruling out day-to-day variation between
416 experiments. Finally, and most critically, we showed that the model could more effectively
417 predict the behavior of masked cells with the inclusion of multiple cell states, and that this
418 prediction is dependent on allowing inheritance between cell generations (Fig. S16). While we
419 have considered the use of experimental control conditions, it is important to keep in mind that

the variation observed here arises both through external perturbation and natural variation within the population. Consequently, we have not been able to identify a context in which one might expect to *not* observe such variation. Also, while experiments in which distinct cell lines are mixed can help to validate methods in which cell relationships are inferred, such as pseudotime methods[50], the cell relationships are not modeled here because they are explicitly known through the measured lineage relationships. Ultimately, experiments uncovering molecular markers and mechanisms of these cell states will provide the best independent validation for their biological significance.

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

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

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

In total, the pipeline developed here provides a unique approach for understanding the structure of dynamic, heterogeneous tumor populations. By capturing the dynamics of state transitions, it

links single cell phenotypes to overall population behavior. Incorporating molecular measurements, and a broader set of drug interventions, will then also help to identify means of modulating state and overall population behavior. Ultimately, we expect this integrative view will help to identify treatments alone and in combination that allow for population-level control by affecting the growth of and transitions between individual cell subpopulations.

Materials and Methods

Experimental cell lineage data

Stable cell line creation, drug treatments, and tracking of AU565 and MCF10A cells were performed as previously described in Gross et al[61] and Gross et al[44], respectively. Briefly, AU565 cells were co-transfected with a transposase plasmid (Addgene #34879) and a donor plasmid that drove expression of a nuclear localized mCherry, puromycin resistance, and a fragment of HDHB fused to the clover fluorescent protein, which was used to track progression through the cell cycle[62]. Cells stably expressing the nuclear and cell cycle reporter were selected for 7 days with 0.75 µg/ml puromycin. The phase of the cells is determined based on whether the amount of fluorescence is greater within nucleus or the cytoplasm[62]. As a result, the reporter signal is invariant to changes in exposure and background. To track drug responses AU565 reporter cells were plated into 24-well plates with fluorobrite media containing 10% FBS, glutamine, and penicillin-streptomycin. 24 hours later fresh media containing escalating doses of lapatinib and gemcitabine was added. MCF10A cells were cultured in growth media (DMEM/F12, 5% horse serum, 20 ng/ml cholera toxin, 10 µg/ml insulin, and 1% Pen/Strep), grown to 50–80% confluency, and detached with 0.05% trypsin-EDTA. 7 hours after seeding 75000 cells, they were washed with PBS and the experiment media (DMEM/F12, 5% horse serum, 0.5 µg/ml hydrocortisone, 100 ng/ml cholera toxin, and 1% Pen/Strep) was added to the 8 well-plates which was followed by 18 hours of incubation. Afterward, cells were treated with growth factors 10 ng/ml EGF, 40 ng/ml HGF, and 10 ng/ml OSM in fresh experiment media. After drug addition, plates were placed in the IncuCyte S3 and four image locations per treatment were imaged every 30 minutes. AU565 were imaged for 96 hours and MCF10A cells for 48 hours. After half the experiment times, fresh media and drugs/growth factors were added. Cell lineages from the IncuCyte images were manually tracked in FIJI[63] to record cell division, death, and the transition from G1 to S/G2 phase (in AU565). AU565 cells are non-motile and fewer than 4% of cells were within one cell length of the image boundary, ensuring minimal sampling bias from the microscopy field of view. Three biological replicates were collected and combined in the final data set. To verify that results did not reflect batch effects, we checked that state assignments were not enriched or depleted within a replicate.

Lineage tree-based hidden Markov model

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

Basic model structure

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

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

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

$$T_{i,j} = T(z_i \rightarrow z_j) = P(z_j \mid z_i), \qquad i, j \in \{1, \dots, K\} \tag{2}$$

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

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

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

Assigning cell states (expectation step)

Upward recursion

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

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

For the state prediction we start by calculating the marginal state distribution (MSD) matrix. MSD is an $N \times K$ matrix that for each cell is marginalizing the transition probability over all

possible current states by traversing from root to leaf cells:

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

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

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

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

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

For non-leaf cells the values are given by:

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

where we calculate the non-leaf NF using:

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

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

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

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

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

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

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

Downward recursion

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

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

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

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

Viterbi algorithm

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

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

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

and the links between parent-daughter cells as:

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

We initialize from the leaf cells as:

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

and for non-leaf cells use:

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

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

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

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

Fitting the cell phenotypes (maximization step)

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

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

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

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

where

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

Estimating emissions distribution parameters

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

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

Baum-Welch

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

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

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

Model evaluation

To find the most likely number of states corresponding to the observations, the Bayesian Information Criterion (BIC) was used[68]. The BIC requires the number of cells and the degrees of freedom, which we calculate using the number of independent parameters. Our model estimates a k element initial probability vector, a $k \times k$ transition matrix, and a $k \times m$ matrix of state-wise parameters where k is the number of states and m is the number of parameters associated with

observation distributions. For the phase-specific observation distributions we have a total of 6 parameters including 2 Bernoulli parameters and 2 pairs of shape and scale parameters for gamma distribution. Since the row-sums for transition and initial probability matrices must be 1, these values are not entirely independent. From distribution analysis of the phase lengths, we realized the shape parameter of the gamma distribution remains constant over different conditions, while the scale parameter changes. Therefore, the shape parameter was shared between the populations treated with 4 different concentrations of the same compound. Each condition therefore introduced 2 free parameters (1 Bernoulli parameter and 1 scale parameter). For the MCF10A experiments, terminal fates and cell cycle durations were also assumed to be Bernoulli- and gamma-distributed, respectively. The shape of cell lifetime was similarly shared among the four conditions (PBS, EGF, HGF, and OSM).

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

Model benchmarking

We used emission distributions to represent the phenotypic characteristics of the cells within the lineages. To create our synthetic data, we considered two possible options as our set of observations throughout an experiment. In one case, we modeled the overall cell fate and cell lifetime; in the second, we modeled the phase-specific fate and duration. In both, we used a Bernoulli distribution for the fate outcomes and a gamma distribution for durations. The state assignment accuracy was calculated using the Rand Index[70]. The difference between true and estimated probability matrices was assessed using the Frobenius norm, or the sum of each element squared.

Synthetic lineage data generation

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

Time censorship

Our phenotypic measurements include the cell fate (progression or cell death) and duration. These measurements are made for each cell cycle phase (G1 or S/G2) in the case of AU565 cells and the entire lifetime for MCF10A cells. These measurements can contain incomplete

information due to the bounds of an experiment. For instance, it is unknown when initial cells present at the start of the experiment began their cell cycle. The same is true of the cells present at the end of the experiment because we do not observe their end. Hence, a cell's lifetime and/or fate may be partially observed. To ensure our synthetic data is a close reflection of experimental data, we incorporated this effect in our synthetic data. Cells with lifetimes that extend beyond the end of the experiment were marked as censored for the lifetime estimation.

Cell overall lifetime observations

The parameters are reflective of the cell phenotypes we observed with 5 nM lapatinib treatment. Figures S1-S5 are based on these parameters. Each figure includes 100 trials.

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

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

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

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

In the following 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.

Table 1. State distribution parameters.

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

Cell cycle phase-specific observations

The synthetic data used in Figures 3, 4, S8–S10 were created based on the following parameters. These parameters are based on estimations from AU565 cells treated with 5 nM lapatinib. Each figure includes 100 trials.

In the following 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.

Table 2. State distribution parameters.

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

To benchmark the model with 5 states, we simulated 25–500 lineages, each with up to 31 cells, to create a population with 5 states. Like with the experimental data, we assumed the experiment ends after 96 hours and censored the cells' observations accordingly. The model parameters, including the distribution parameters, transition probabilities, and initial probabilities, are listed below. The analysis results are shown in Figure S14.

Table 3. State distribution parameters.

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

$$\boldsymbol{T} = \begin{bmatrix} \mathbf{0.6} & \mathbf{0.1} & \mathbf{0.1} & \mathbf{0.1} & \mathbf{0.1} \\ \mathbf{0.05} & \mathbf{0.8} & \mathbf{0.05} & \mathbf{0.05} & \mathbf{0.05} \\ \mathbf{0.01} & \mathbf{0.1} & \mathbf{0.7} & \mathbf{0.09} & \mathbf{0.1} \\ \mathbf{0.1} & \mathbf{0.1} & \mathbf{0.05} & \mathbf{0.7} & \mathbf{0.05} \\ \mathbf{0.1} & \mathbf{0.1} & \mathbf{0.05} & \mathbf{0.05} & \mathbf{0.7} \end{bmatrix} \tag{22}$$

$$\boldsymbol{\pi} = \begin{bmatrix} \mathbf{0.13} \\ \mathbf{0.33} \\ \mathbf{0.16} \\ \mathbf{0.18} \\ \mathbf{0.18} \end{bmatrix} \tag{23}$$

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

Varying emission differences

To create synthetic data with subpopulations of varying dissimilarity (Fig. 5), we use the phase-specific parameters, with the values for the G1 phase gamma scale parameter for state 1 varying over [4, 20]. This results in an increase in the Wasserstein distance between the two cell states, allowing us to measure state assignment accuracy for different dissimilarity amounts between the two states. Likewise, for Figures S6 and S7, we simulated the overall cell lifetime and varied the gamma distribution scale parameter from 1 to 8 for state 1.

Optimal baseline classifier

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

$$p(x \mid G(k_1, \theta_1)) = p(x \mid G(k_2, \theta_2)) \tag{24}$$

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

$$x = \frac{k \ \ln \frac{\theta_2}{\theta_1}}{\frac{1}{\theta_1} - \frac{1}{\theta_2}} \tag{25}$$

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

Cross-validation

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

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

Lowess trendline

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

Code availability

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

Data availability

The experimental lineage data for AU565 and MCF10A cell lines can be found at https://github.com/meyer-lab/tHMM

The synthetic data from which we plotted Fig. 3c–d, Fig. 4b–g, Fig. 5c, Fig. 6, and Fig. 7 uses the code in the file named after the corresponding figure number. Data used in Figs. 7a–b, 8, and

9 uses the AU565 cell line experimental lineage data, and figures 7c and 10 use the MCF10A lineage data.

The cell lines used in this study (AU565, MCF10A) can be made available upon request.

Statistics and Reproducibility

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

Acknowledgements

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

References

1. Di Maio, M. *et al.* Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. *Lancet Oncol.* **6**, 669–677 (2005).

2. De Roock, W. *et al.* Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol.* **11**, 753–762 (2010).

3. Sharma, S. V. *et al.* A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* **141**, 69–80 (2010).

4. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).

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

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

7. Falkenberg, K. J. & Johnstone, R. W. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat. Rev. Drug Discov.* **13**, 673–691 (2014).

8. Inde, Z. & Dixon, S. J. The impact of non-genetic heterogeneity on cancer cell death. *Crit.*

Rev. Biochem. Mol. Biol. **53**, 99–114 (2018).

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

10. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nat. 2019 5697757* **569**, 503–508 (2019).

11. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nat. 2012 4837391* **483**, 603–607 (2012).

12. Gupta, P. B. *et al.* Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* **146**, 633–644 (2011).

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

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

15. Arai, T. *et al.* Tumor Doubling Time and Prognosis in Lung Cancer Patients: Evaluation from Chest Films and Clinical Follow-up Study. *Jpn. J. Clin. Oncol.* **24**, 199–204 (1994).

16. Bourhis, J. *et al.* Potential doubling time and clinical outcome in head and neck squamous cell carcinoma treated with 70 GY in 7 weeks. *Int. J. Radiat. Oncol.* **35**, 471–476 (1996).

17. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).

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

19. Tyson, D. R., Garbett, S. P., Frick, P. L. & Quaranta, V. Fractional proliferation: A method to deconvolve cell population dynamics from single-cell data. *Nat. Methods* (2012) doi:10.1038/nmeth.2138.

20. O'Connor, J. P. B. *et al.* Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin. Cancer Res.* **21**, 249–257 (2015).

21. Chen, K. *et al.* Phenotypically supervised single-cell sequencing parses within-cell-type heterogeneity. *iScience* **24**, 101991 (2020).

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

23. Mitchell, S., Roy, K., Zangle, T. A. & Hoffmann, A. Nongenetic origins of cell-to-cell variability in B lymphocyte proliferation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2888–E2897 (2018).

24. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of

genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).

25. Brumpton, B. *et al.* Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. (2019) doi:10.1101/602516.

26. Concannon, P. *et al.* Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes* **58**, 1018–1022 (2009).

27. Concannon, P., Rich, S. S. & Nepom, G. T. Genetics of Type 1A Diabetes. *N. Engl. J. Med.* **360**, 1646–1654 (2009).

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

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

30. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).

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

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

33. Durand, J. B., Gonçalvès, P. & Guédon, Y. Computational methods for hidden Markov tree models-An application to wavelet trees. *IEEE Trans. Signal Process.* **52**, 2551–2560 (2004).

34. Choi, H. & Baraniuk, R. G. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans. Image Process.* **10**, 1309–1321 (2001).

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

36. Olariu, V. *et al.* Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics* **25**, 2824–2830 (2009).

37. Nakashima, S., Sughiyama, Y. & Kobayashi, T. J. Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics* **36**, 2829–2838 (2020).

38. Lund, B., Kristjansen, P. E. G. & Hansen, H. H. Clinical and preclinical activity of 2′,2′-difluorodeoxycytidine (gemcitabine). *Cancer Treat. Rev.* (1993) doi:10.1016/0305-7372(93)90026-N.

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

40. Chao, H. X. *et al.* Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Mol. Syst. Biol.* **15**, e8604–e8604 (2019).

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 Genet.* **15**, e1008458–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. *J. R. Soc. Interface* **14**, 20170141 (2017).

43. Tang, L., Wang, Y., Strom, A., Gustafsson, J. A. & Guan, X. Lapatinib induces p27Kip1-dependent G1 arrest through both transcriptional and post-translational mechanisms. *Cell Cycle* (2013) doi:10.4161/cc.25728.

44. Gross, S. M. *et al.* A LINCS microenvironment perturbation resource for integrative assessment of ligand-mediated molecular and phenotypic responses. (2021) doi:10.1101/2021.08.06.455429.

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

46. Fan, Y. & Meyer, T. Molecular control of cell density-mediated exit to quiescence. *Cell Rep.* **36**, 109436 (2021).

47. Zhu, X. *et al.* Autophagy stimulates apoptosis in HER2-overexpressing breast cancers treated by lapatinib. *J. Cell. Biochem.* **114**, 2643–2653 (2013).

48. Richards, R. *et al.* Drug antagonism and single-agent dominance result from differences in death kinetics. *Nat. Chem. Biol.* **16**, 791–800 (2020).

49. Johnson, T. I. *et al.* Quantifying cell cycle-dependent drug sensitivities in cancer using a high throughput synchronisation and screening approach. *EBioMedicine* **68**, 103396 (2021).

50. Campbell, K. R. & Yau, C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat. Commun. 2018 91* **9**, 1–12 (2018).

51. Lartillot, N. A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data. *Bioinformatics* **30**, 488–496 (2013).

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

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

54. Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* **22**, 262–269 (2016).

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

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

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

58. Schwartz, A. D. *et al.* A biomaterial screening approach reveals microenvironmental mechanisms of drug resistance. *Integr. Biol. (Camb).* **9**, 912–924 (2017).

59. Emert, B. L. *et al.* Variability within rare cell states enables multiple paths toward drug resistance. *Nat. Biotechnol.* **39**, 865–876 (2021).

60. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2017).

61. Gross, S. M. *et al.* Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *bioRxiv* 2020.07.24.219907 (2021) doi:10.1101/2020.07.24.219907.

62. SL, S. *et al.* The proliferation-quiescence decision is controlled by a bifurcation in CDK2 activity at mitotic exit. *Cell* **155**, 369 (2013).

63. Meijering, E., Dzyubachyk, O. & Smal, I. Methods for cell and particle tracking. in *Methods in Enzymology* (2012). doi:10.1016/B978-0-12-391857-4.00009-4.

64. Ephraim, Y. & Merhav, N. Hidden Markov processes. *IEEE Trans. Inf. Theory* **48**, 1518–1569 (2002).

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

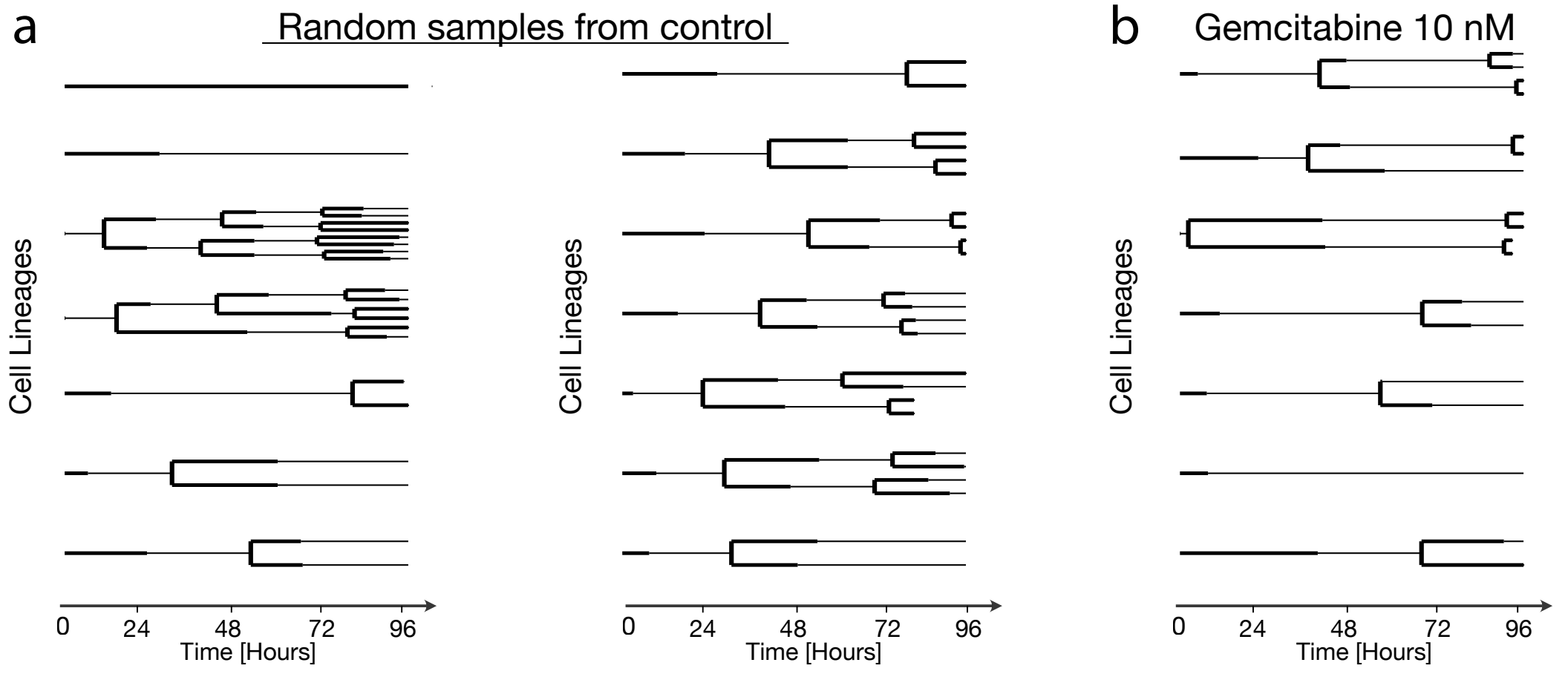
66. Devijver, P. A. Baum's forward-backward algorithm revisited. *Pattern Recognit. Lett.* **3**, 369–373 (1985).

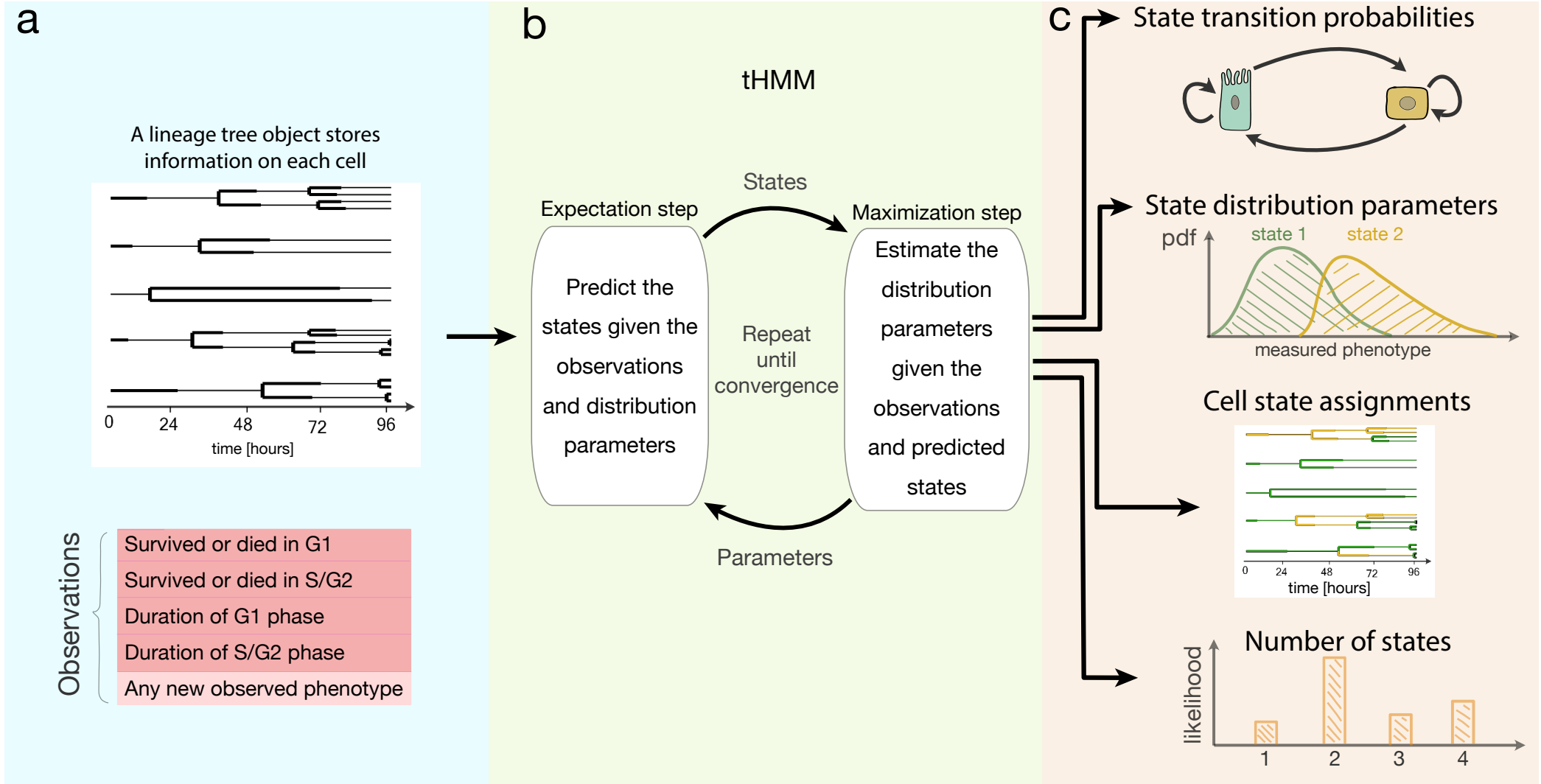
67. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods 2020 173* **17**, 261–272 (2020).

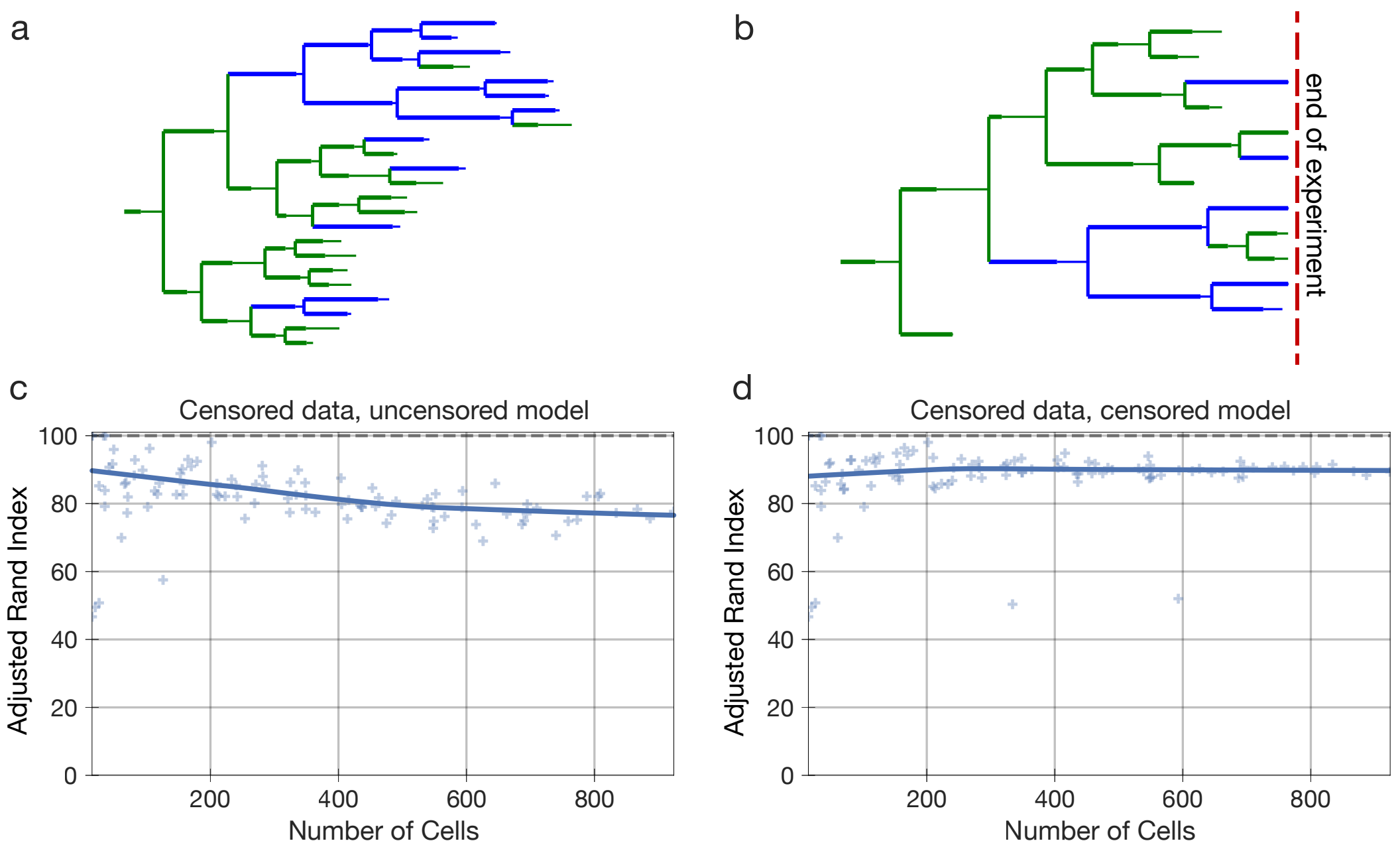
68. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2011).

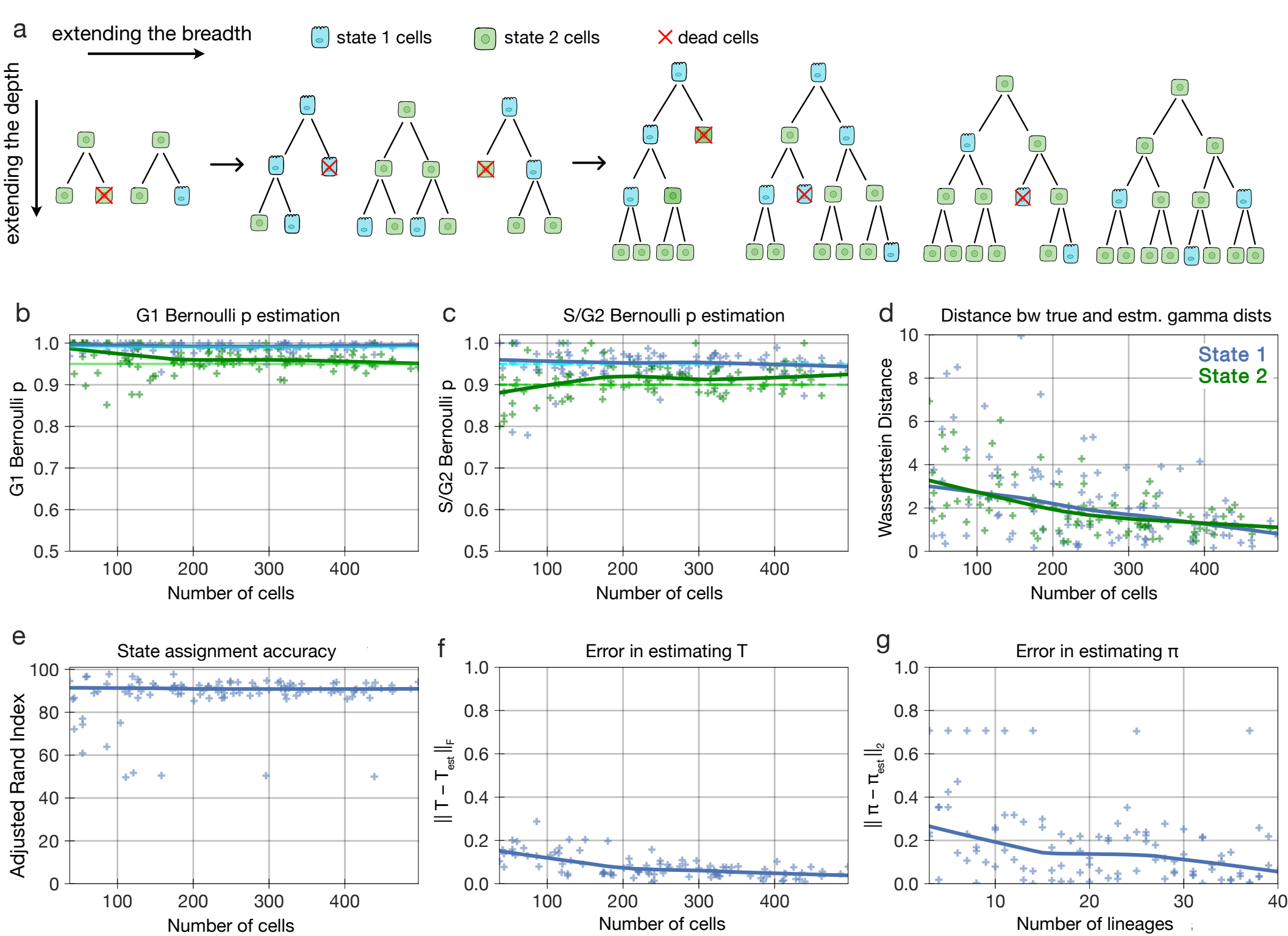
69. Vallender, S. S. Calculation of the Wasserstein Distance Between Probability Distributions on the Line. *Theory Probab. & Its Appl.* **18**, 784–786 (1974).

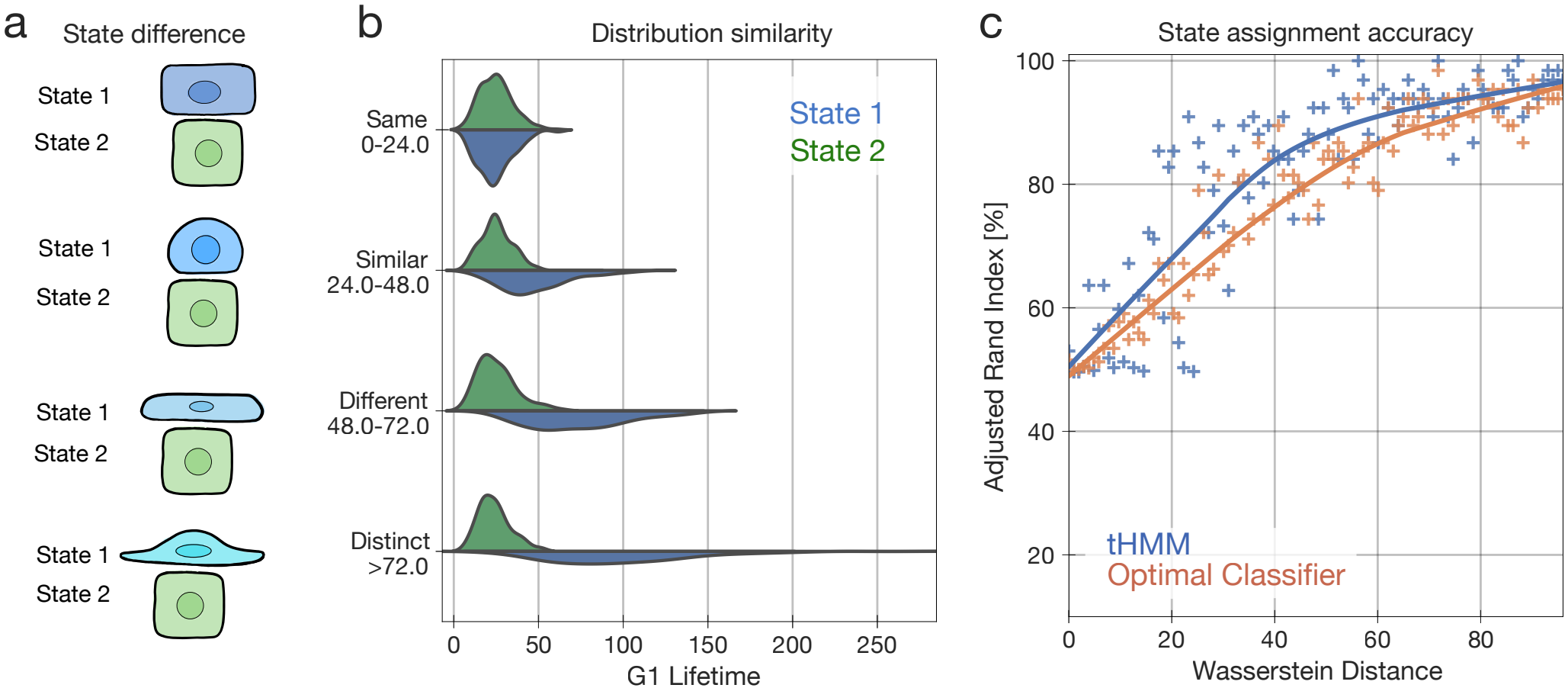
70. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

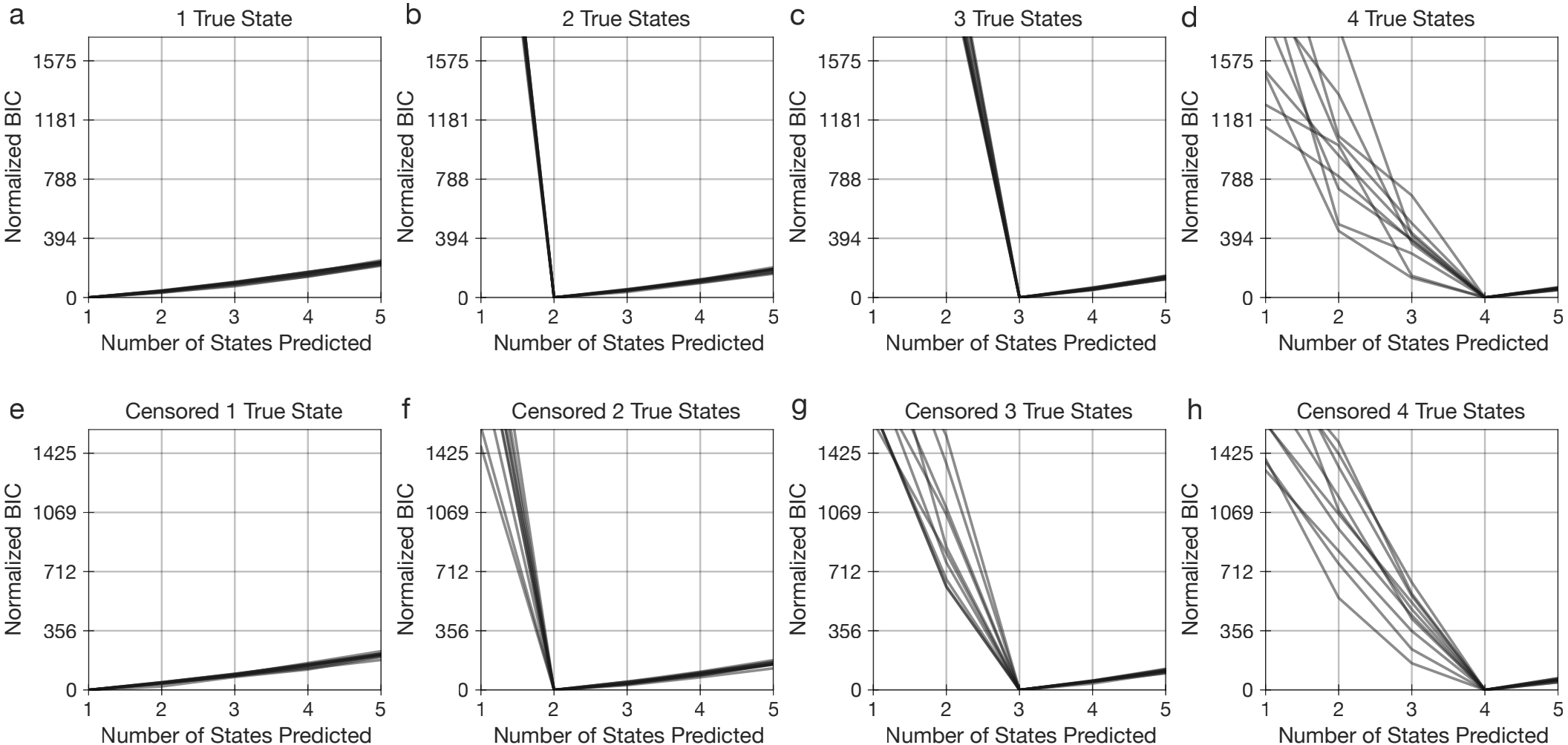

a
Random samples from control
Cell Lineages
0
24
48
72
96
Time [Hours]
Cell Lineages
0
24
48
72
96
Time [Hours]
b
Gemcitabine 10 nM
Cell Lineages
0
24
48
72
96
Time [Hours]

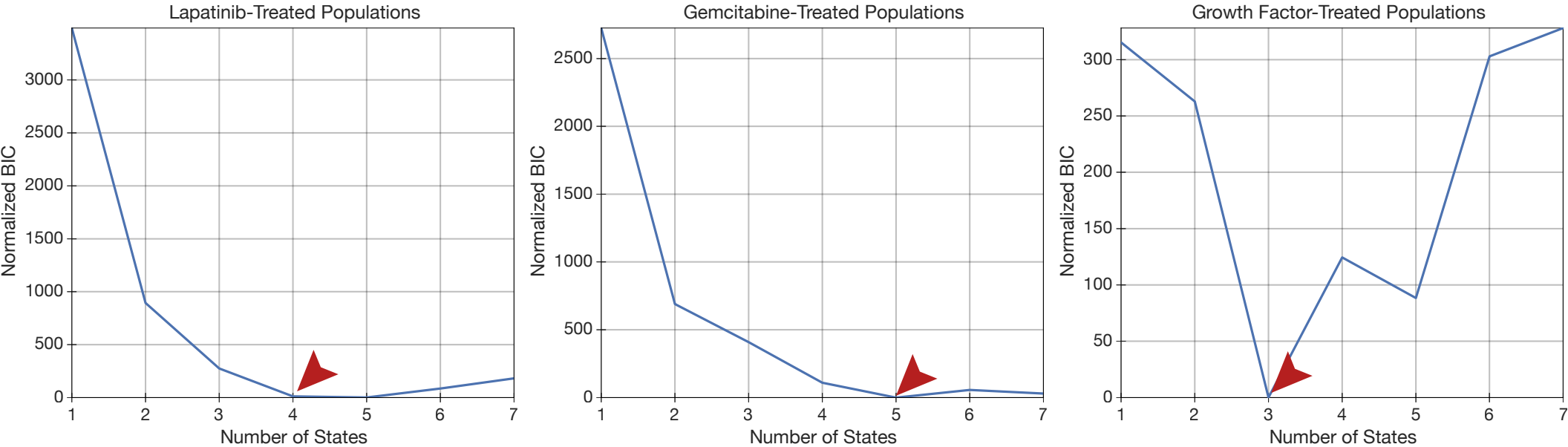

a
A lineage tree object stores information on each cell
0
24
48
72
96
time [hours]
Observations
Survived or died in G1
Survived or died in S/G2
Duration of G1 phase
Duration of S/G2 phase
Any new observed phenotype
b
tHMM
Expectation step
Predict the states given the observations and distribution parameters
States
Repeat until convergence
Parameters
Maximization step
Estimate the distribution parameters given the observations and predicted states
c
State transition probabilities
State distribution parameters
pdf
state 1
state 2
measured phenotype
Cell state assignments
0
24
48
72
96
time [hours]
Number of states
likelihood
1
2
3
4

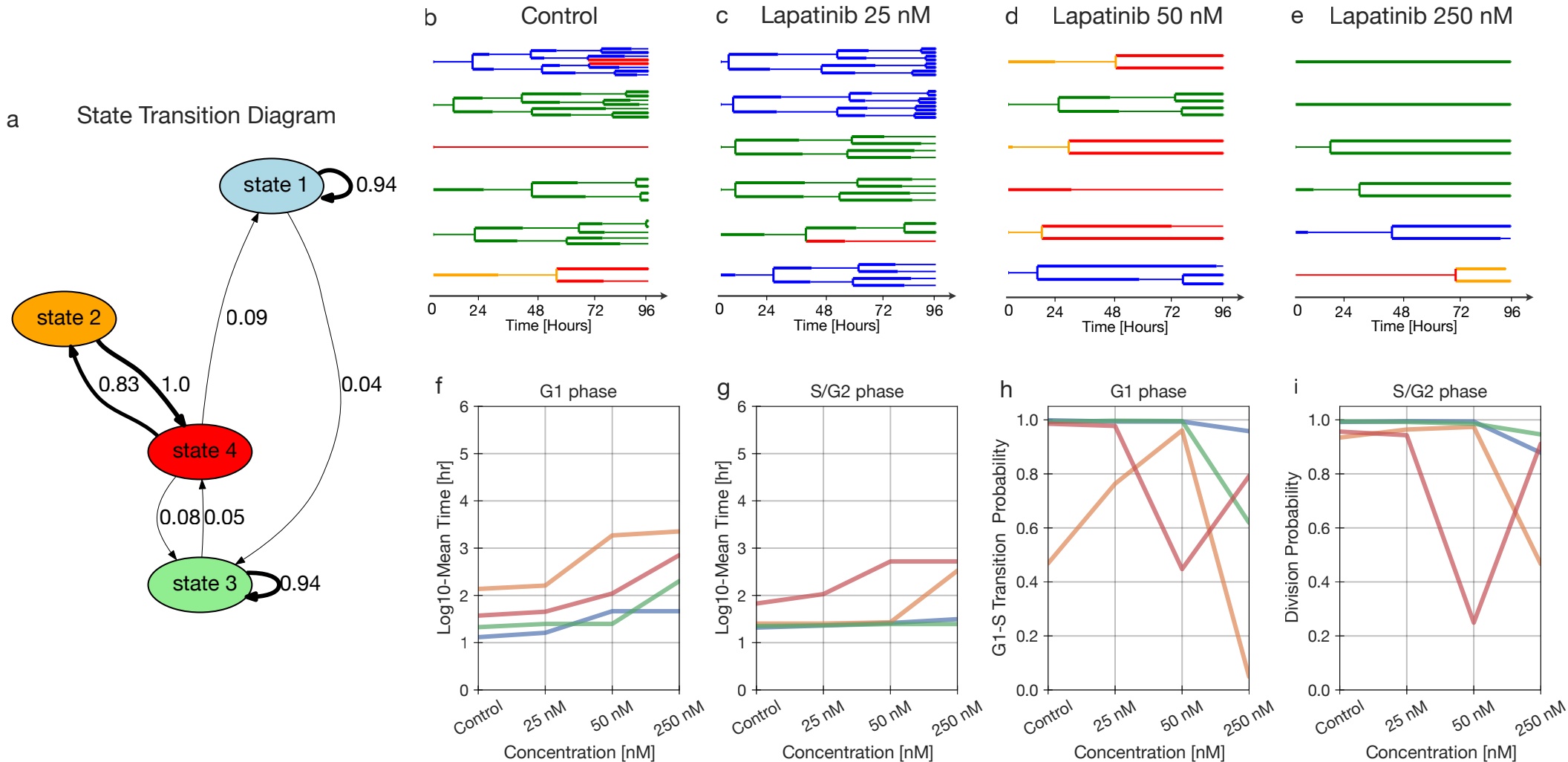

a
b
end of experiment
c
Censored data, uncensored model
Adjusted Rand Index
Number of Cells
d
Censored data, censored model
Adjusted Rand Index
Number of Cells

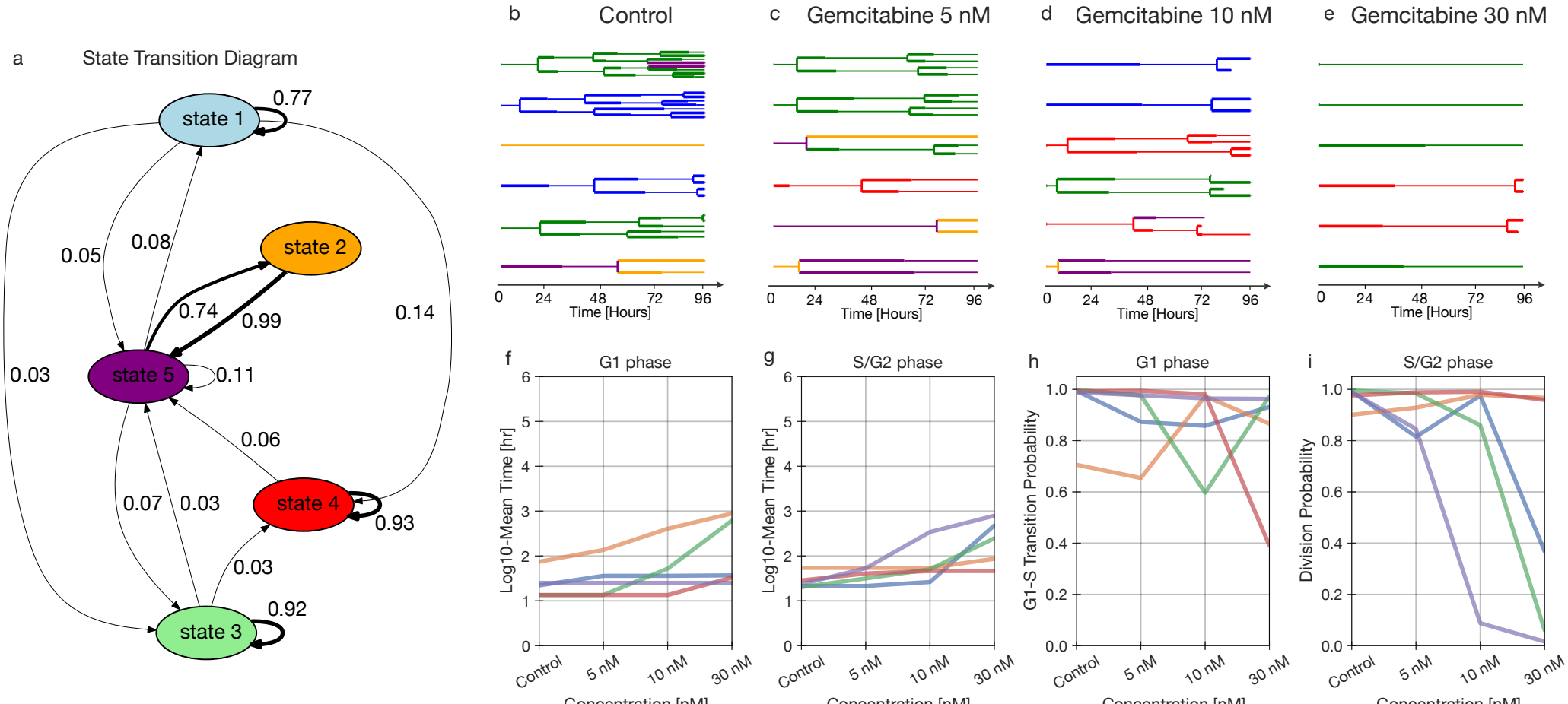

a
extending the breadth
extending the depth
state 1 cells
state 2 cells
dead cells
b
G1 Bernoulli p estimation
G1 Bernoulli p
Number of cells
c
S/G2 Bernoulli p estimation
S/G2 Bernoulli p
Number of cells
d
Distance bw true and estm. gamma dists
Wassertstein Distance
State 1
State 2
Number of cells
e
State assignment accuracy
Adjusted Rand Index
Number of cells
f
Error in estimating T
$\| T - T_{est} \|_F$
Number of cells
g
Error in estimating π
$\| \pi - \pi_{est} \|_2$
Number of lineages


a
State difference
State 1
State 2
State 1
State 2
State 1
State 2
State 1
State 2
b
Distribution similarity
Same
0-24.0
Similar
24.0-48.0
Different
48.0-72.0
Distinct
>72.0
State 1
State 2
0
50
100
150
200
250
G1 Lifetime
c
State assignment accuracy
Adjusted Rand Index [%]
100
80
60
40
20
tHMM
Optimal Classifier
0
20
40
60
80
Wasserstein Distance


a
1 True State
b
2 True States
c
3 True States
d
4 True States
e
Censored 1 True State
f
Censored 2 True States
g
Censored 3 True States
h
Censored 4 True States
Normalized BIC
Number of States Predicted
1575
1181
788
394
0
1425
1069
712
356
1
2
3
4
5


Lapatinib-Treated Populations
Gemcitabine-Treated Populations
Growth Factor-Treated Populations
Normalized BIC
Number of States
0
500
1000
1500
2000
2500
3000
50
100
150
200
250
300
1
2
3
4
5
6
7


a
State Transition Diagram
state 1
0.94
state 2
0.09
0.83
1.0
0.04
state 4
0.08
0.05
state 3
0.94
b
Control
c
Lapatinib 25 nM
d
Lapatinib 50 nM
e
Lapatinib 250 nM
0
24
48
72
96
Time [Hours]
f
G1 phase
Log10-Mean Time [hr]
g
S/G2 phase
Log10-Mean Time [hr]
h
G1 phase
G1-S Transition Probability
i
S/G2 phase
Division Probability
Control
25 nM
50 nM
250 nM
Concentration [nM]


a
State Transition Diagram
state 1
state 2
state 3
state 4
state 5
0.77
0.05
0.08
0.74
0.99
0.14
0.03
0.11
0.06
0.07
0.03
0.93
0.03
0.92
b
Control
c
Gemcitabine 5 nM
d
Gemcitabine 10 nM
e
Gemcitabine 30 nM
0
24
48
72
96
Time [Hours]
f
G1 phase
g
S/G2 phase
h
G1 phase
i
S/G2 phase
Log10-Mean Time [hr]
G1-S Transition Probability
Division Probability
Control
5 nM
10 nM
30 nM
Concentration [nM]