Lineage EM Algorithm for Inferring Latent States from Cellular Lineage Trees

So Nakashima[a,1], **Yuki Sughiyama**[b], **and Tetsuya J. Kobayashi**[a,b,c,1]

[a]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-8654, Japan; [b]Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku 153-8505, Tokyo, Japan; [c]PRESTO, Japan Science and Technology Agency (JST), 4-1-8 Honcho Kawaguchi, Saitama 332-0012, Japan

This manuscript was compiled on November 29, 2019

Phenotypic variability in a population of cells can work as the bet-hedging of the cells under an unpredictably changing environment, the typical example of which is the bacterial persistence. To understand the strategy to control such phenomena, it is indispensable to identify the phenotype of each cell and its inheritance. Although recent advancements in microfluidic technology offer us useful lineage data, they are insufficient to directly identify the phenotypes of the cells. An alternative approach is to infer the phenotype from the lineage data by latent-variable estimation. To this end, however, we must resolve the bias problem in the inference from lineage called survivorship bias. In this work, we clarify how the survivor bias distorts statistical estimations. We then propose a latent-variable estimation algorithm without the survivorship bias from lineage trees based on an expectation-maximization (EM) algorithm, which we call Lineage EM algorithm (LEM). LEM provides a statistical method to identify the traits of the cells applicable to various kinds of lineage data.

1. Introduction

A population of genetically identical cells is phenotypically heterogeneous and the heterogeneity is partially inherited over generations (1–3). Such phenotypic variety leads to behavioral individuality of each cell, which, in turn, generates complicated population phenomena. One example is bacterial persistence in which a fraction of cells in a population survives when the population experiences an antibiotic exposure even though the other fraction dies out (4–8). Persistence is also recently recognized relevant to the drug-resistance of cancers (9, 10). While the survivors, which are also called persisters, were originally conjectured as dormant and thereby drug-insensitive cells in a population, recent bioimaging analysis revealed that persistence is a more intricate phenomenon, which involves resistant but still growing cells (5, 6). Because drug-resistance is tightly related to the manner how the drug is incorporated into a cell and interferes with the self-replicating process, quantitative analysis of persistence requires a characterization of growth states of individual cells and their competition in a population. More generally, the heterogeneities in the growth speed and the death rate as well as their inheritance from a mother to daughter cells constitute Darwinian natural selection among the cells. Such natural selection at the cellular level is also highly relevant to drug-resistances of pathogens and cancers, immunological memories, cell competitions in tissues, and induction of iPS cells (9–13). Moreover, the population dynamics under selection is integratively determined by the growth states of cells, their statistical property, and their inheritance dynamics over generations. Therefore, identification of growth states of cells from data is crucial for predicting and controlling those selection-driven phenomena (14, 15).

To determine the growth states and their dynamics, we can take advantage of recent bioimaging and microfluidic technology, which provide abundant but *incomplete* data of the population growth of the cells (16–18). Bioimaging enables us to track growing and replicating cells over time and microfluidic devises such as dynamics cytometer substantially boosts our capability of tracking cells beyond hundreds of generations by keeping growing cells in a fixed-sized chamber (19, 20) (Fig. 1 (a)). The tracked cells in the chamber reconstitute trees of lineages, which contain more detailed information on the mother-daughter relationship of the cells and on the actual competition among the cells in the population (Fig. 1 (b)). From the division patterns in such trees, we may infer cellular growth states as well as their statistical properties and inheritance dynamics.

However, the inference is accompanied by two difficulties. First, the inference from a tree should be conducted by appropriately handling the branching relationship among the cells in the tree. This problem has been studied by using the kin-correlation (21, 22), an algebraic invariance of the lineage tree (23, 24), clustering algorithms (25, 26), Monte-Carlo algorithms (27), and model selection (28). See also Supplementary Information (Section S1). While this problem seems to be addressed by combining these existing estimation techniques in the machine learning, this naïve anticipations is hampered by the second difficulty. In the cellular lineage tree, each edge has a different length that reflects the actual division time of the cell. Therefore, clades of cells replicating faster by chance are represented more than the others in the tree, which inevitably

SN, YS, and TJK designed and performed research; SN and TJK analyzed data; SN, YS, and TJK wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. E-mail: so_nakashima@mist.i.u-tokyo.ac.jp; tetsuya@mail.crmind.net

introduces bias in the data sample. This is a generalization of the so-called survivorship bias in statistics such that fast-growing clades are overrepresented in a population whereas the slow ones are underrepresented (18, 19, 29, 30) (Fig. 1 (b)).

To illustrate the overrepresentation, we show a simulated lineage tree of cells with two states. The mean division times of the states are the same but cells in the blue state has greater variability in the division time (Fig. 2 (a)). The transition probability between them is symmetric. In Fig. 2 (b), each edge in a tree represents the duration between consecutive divisions of a cell and a binary branch is the generation of two daughter cells from a mother cell by its division. As shown in Fig. 2 (b), the cells in the blue state appear more in the tree even though their mean division time is the same as that of the red state. Thereby, statistical estimators can suffer from sampling bias. Previous works circumvent this problem by pruning a lineage tree so that each leaf cell has the same number of branching points along the lineage up to the root cell. However, as shown in Fig. 2 (c), such pruning inevitably loses the information about growth statistics of individual cells, which are manifested by the pattern of division times over lineages.

The situation can be even worse when we work on real lineage trees obtained by experiments. First, the growth states of the cells are not provided as observations but should be inferred from the pattern of divisions of the cells. Second, as exemplified by the lineage tree of *E.coli* obtained by dynamics cytometer (Fig. 2 (d)), a real tree has frequent terminations of edges because tracking of the corresponding cells is hampered by various reasons. In the case of dynamics cytometer, a termination occurs because a cell is carried away from the chamber by the constant flow of medium on the two ends (Fig. 1 (a)). However, by the virtue of this carry-away, the population size is kept almost constant and we can track a fraction of cells over a hundred generations. As shown in Fig. 2 (d), the pruning substantially reduces the number of cells (samples) that we can use for the subsequent inference of cellular states and their division statistics. If, by contrast, we try to use all the cells in the tree for inference without pruning, we have to appropriately correct the survivorship bias in estimating characteristics of growth states.

Even if we use different culturing devices and cell types, we still face technological limitation upon the number of tracked cells in an exponentially growing population. Therefore, similar problems should arise when the long-term lineage tracking devices of bacteria are extended to other cells, e.g., cancer cells, immune cells, and iPS cells.

The survivorship bias in an exponentially growing system with states has been analyzed only from recently (31–34). Thus, direct inference of cellular growth states without pruning is still immature and we have to establish a method that can estimate and characterize growth states of cells that properly correct the bias. Furthermore, such method contributes not only to accurate inference but also to estimation of the selection pressure as the strength of the bias (20, 35).

To address these problems, we first clarify how the survivorship bias distorts statistical estimations depending on the way to collect a sample of cells from the tree (Section 3). By deriving explicit relations between unbiased and biased estimates, we establish a correction method of the survivorship bias under the condition that the states of the cells are known. Then, we propose an estimation algorithm of the latent states from lineages based on an expectation-maximization (EM) algorithm, which we call **Lineage EM algorithm (LEM)** (Section 4). Our LEM works under a wide variety of the circumstances without regard to the shape of the lineages nor to the properties of the models in contrast to the previous researches assuming them. LEM is therefore applicable to the lineage data obtained both by the dynamics cytometer and by the mother machine (17) (See also Section 3). We verify the effectiveness of LEM by using synthetic data (Section 5). Finally, we apply LEM to a lineage tree of *E. coli*, and identify a latent three-dimensional continuous state, one component of which encodes the information on the inheritance of the states and the replicative capabilities over generations.

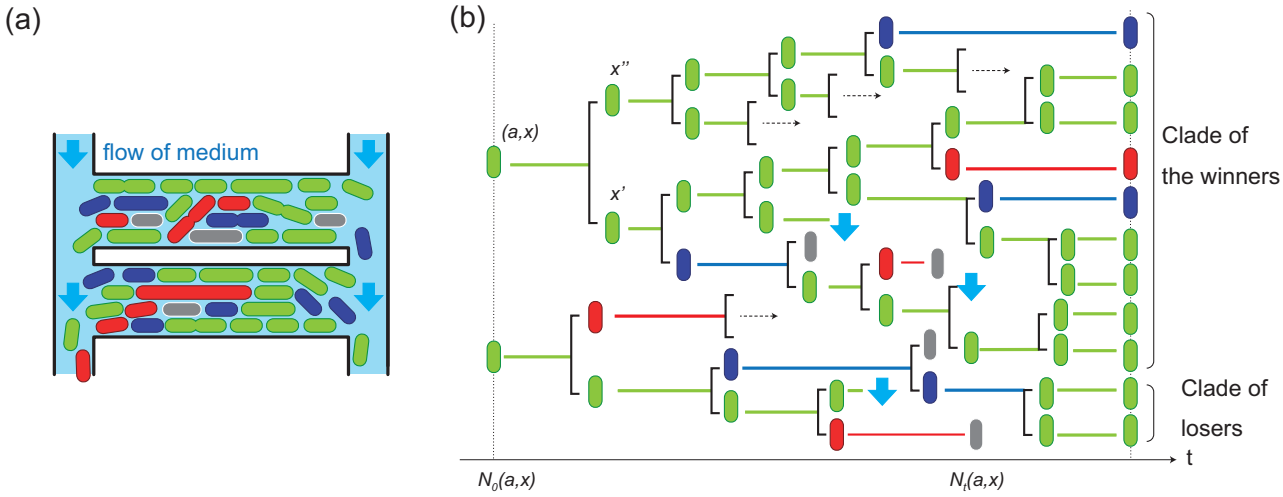

Fig. 1. (a) The outline of the dynamics cytometer (18). (b) A schematic representation of a lineage tree obtained by the dynamics cytometer, and an illustration of the survivorship bias. The lineage tree is composed of two kinds of information: mother-daughter relationship of the cells and the division times how long each cell took until divides. In this panel, the green state is assumed to divide faster than the blue and red ones. Thereby, the cells with the green state are overrepresented more in the clade of the winner than in that of the losers.

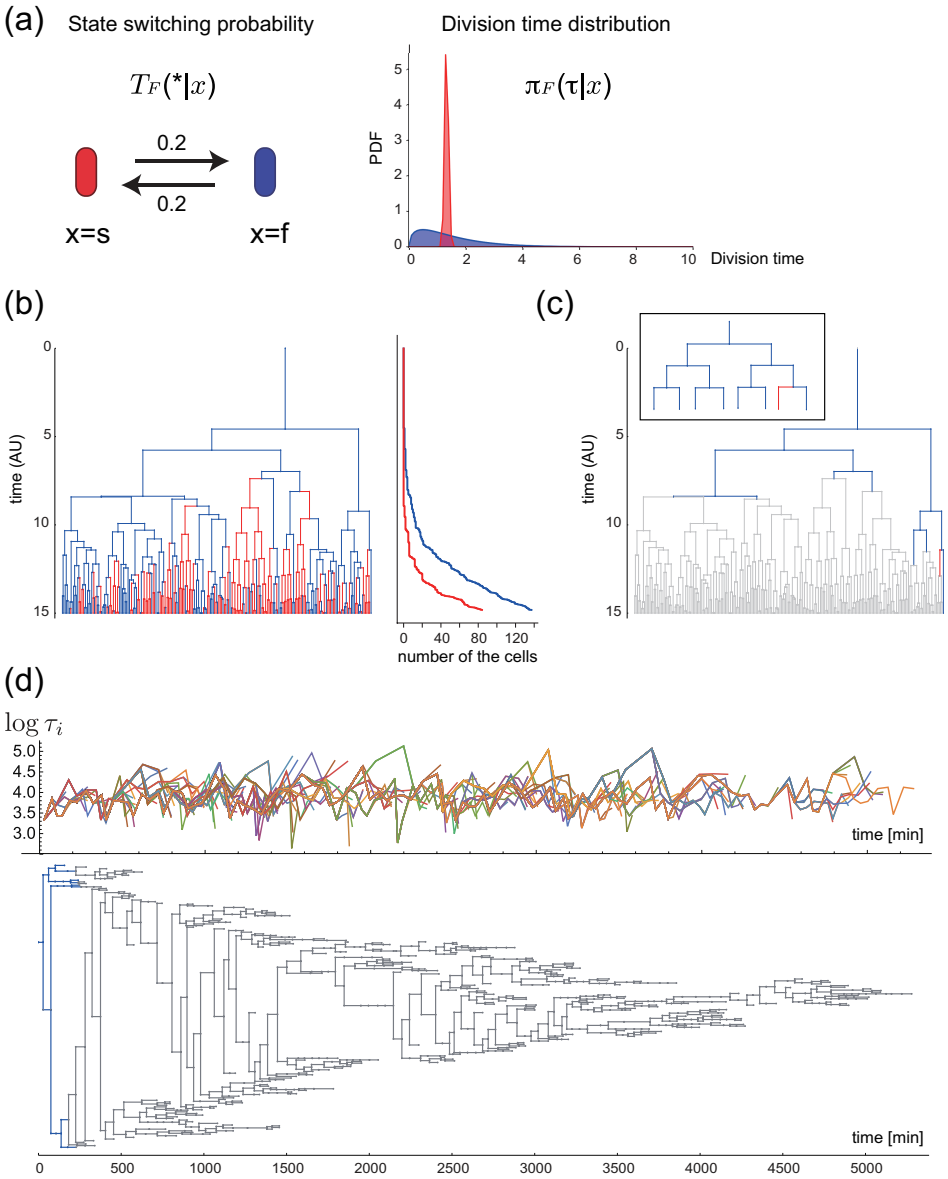

Fig. 2. (a) A two-state model to generate a synthetic lineage tree of cells. Red and blue states have a symmetric transition probability $\mathbb{T}_{\mathrm{F}}(*|\boldsymbol{x})$ (left panel). Both states have the same mean division time but the blue state has greater variability in the division time (right panel). (b) A synthetic lineage tree obtained by simulating the model in (a). See also Supplementary Information (Section S12). The vertical axis shows time (arbitrary unit). Each edge in the lineage tree represents the time duration between the birth and the division (or death) of the corresponding cells. The horizontal edges show mother-daughter relationships. The right panel shows the cumulative number of the cells in the blue or the red state in the lineage tree over time. The cells in the blue state are much more than those in the red state even though their mean division times are the same. (c) A pruned lineage tree of (b) to avoid the survivorship bias, which is shown in color. The cells represented by gray are discarded after pruning. Inset shows the same pruned tree with fixed edge length. (d) A lineage tree of *E. coli* (F3 rpsL-gfp strain) cells grown with M9 minimum medium supplemented with 0.2% glucose at 37°. The upper panel in (d) is the time series plot of the division times. The lower panel represents the lineage tree of E. coli. The data is adopted from (18) and re-plotted. The sub-tree after pruning is also shown in blue. The gray cells are discarded after pruning.

2. Statistical modeling of state-switching and division

In this paper, we use a variant of the branching process as a model of a proliferating population with state switching (Fig. 3). We consider the symmetric division upon which a mother cell always turns into two daughters. Each cell is supposed to have its state $\boldsymbol{x} \in \Omega$ where Ω is either discrete or continuous. For simplicity, we assume that, upon the division of the mother cell, each daughter cell switches its state stochastically. However, we note that all the results in this paper is applicable to the cases where the number of the daughters is not necessarily two. The state-switching of a daughter cell is assumed to be dependent on the state of the mother but independent of the state switching of its sister cell. Then, the probability to change the state from $\boldsymbol{x}$ to $\boldsymbol{x}'$ is given by a transition matrix $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$, where $\sum_{\boldsymbol{x}'} \mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x}) = 1$ (Fig. 3). For notational simplicity, we use $\sum_{\boldsymbol{x}' \in \Omega}$ instead of $\int_{\boldsymbol{x}' \in \Omega} d\boldsymbol{x}'$ even for Ω being a continuous state space. The division time τ, the duration time between consecutive divisions, is dependent on the state $\boldsymbol{x}$ of the cell, the probability distribution of which is denoted by $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ (Fig. 3). We note that $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ and $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ define a multi-type age-dependent branching process (Fig. 1 (b)) (36), whereas they also constitute

a continuous semi-Markov process if one of the daughter cells is ignored (Fig. 3) (37). See also Supplementary Information (Section S2 and S3). In general, the state $\boldsymbol{x}$ of a cell should be characterized as a point or a trajectory in the high dimensional state space consisting of the abundance of intracellular metabolites and molecules in the cell. However, the state to be inferred in this work can be its low dimensional projection being relevant for the division time, because any two states, $\boldsymbol{x}$ and $\boldsymbol{x}'$, that give the same division statistics as $\pi_{\mathrm{F}}(\tau|\boldsymbol{x}) = \pi_{\mathrm{F}}(\tau|\boldsymbol{x}')$ cannot be distinguished by the inference only from the data of τ. Note that, in principle, we can employ other quantities for the inference if provided.

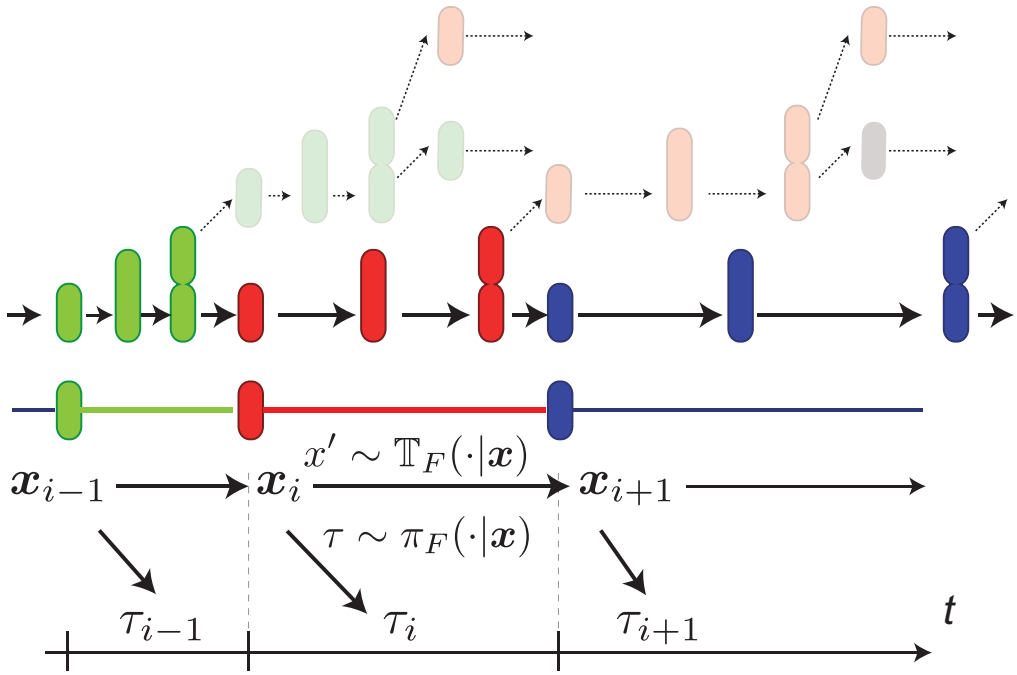

Fig. 3. Schematic diagram of stochastic divisions and associated state transitions along a cellular lineage. The colors of cells represent their states. Transparent cells indicate daughter cells that are not in the lineage we are focusing on. All daughter cells are assumed to follow the same statistical laws as those in the focused lineage.

3. Correction of the survivorship bias in estimation of state-switching and division dynamics

If the states of cells are known or experimentally observed, the division time statistics and the state-switching probability, $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$, may be empirically estimated by the histogram of τ of the cells with $\boldsymbol{x}$ and by counting the number of the state-switching from $\boldsymbol{x}$ to $\boldsymbol{x}'$ in a given data set, respectively:

$$\pi^{\mathcal{D}}_{\mathrm{emp}}(\tau|\boldsymbol{x}) := \frac{1}{|\mathcal{D}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{D}_{\boldsymbol{x}}}\delta(\tau-\tau_i), \qquad [1]$$

$$\mathbb{T}^{\mathcal{D}}_{\mathrm{emp}}(\boldsymbol{x}'|\boldsymbol{x}) := \frac{\text{The number of the transitions from } \boldsymbol{x} \text{ to } \boldsymbol{x}'}{\text{The number of the transitions from } \boldsymbol{x}}, \qquad [2]$$

where the symbol $|A|$ denotes the cardinality of a finite sample point set A, and τ_i is the division time of the cell i. $\mathcal{D}$ is the set of all cells used for the estimation, i.e., a data sample, and $\mathcal{D}_{\boldsymbol{x}} \subset \mathcal{D}$ is the subset of the cells with the state $\boldsymbol{x}$. $\mathbb{T}_{\mathrm{emp}}$ and π_{emp} may converge for a sufficient large number of cells in $\mathcal{D}$. However, the converged distributions are dependent on the way how the cells in $\mathcal{D}$ were sampled (Fig. 4), and can be substantially biased thereby.

A. Chronological sampling and forward process. Tracking a dividing single cell under a constant condition is the most straight forward way to obtain a data sample of the state-switching events and the division times. The popular measurement system is the mother machine with which we can trace a cell located at the bottom of a chamber (17). Because the cell to be observed is determined at the beginning of an experiment and its lineage is traced chronologically by ignoring one of the sibling cells at each division, the state-switching and the division dynamics obtained in this way is characterized by the semi-Markov stochastic process with π_{F} and $\mathbb{T}_{\mathrm{F}}$ (Fig. 3). See also Supplementary Information (Section S2 and S3). Thereby, π_{emp} and $\mathbb{T}_{\mathrm{emp}}$ converge to π_{F} and $\mathbb{T}_{\mathrm{F}}$, respectively, for a large sample size. We specifically call this type of sampling the chronological sampling and the dynamics generated by π_{F} and $\mathbb{T}_{\mathrm{F}}$ the forward process (18, 20, 37). We can also effectively obtain a chronologically sampled lineage from the tree by using the weighting technique proposed in (35).

Even with its straight forward interpretation, the chronological sampling has some drawbacks in terms of the estimation. First, the observation should be terminated by the death of the tracked cell (Fig. 4 (a)), which limits the size of the data sample and the length of the lineages especially when the cells are cultured in a harsh condition. Second, the tracked cells may

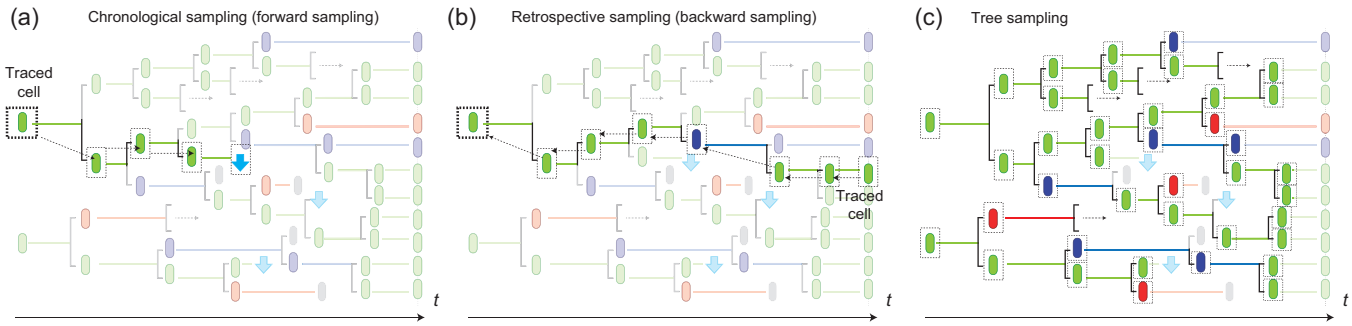

Fig. 4. The chronological (a), retrospective (b), and tree (c) samplings are illustrated by using lineage trees. The cells sampled from the tree in each sampling are designated by dashed squares. (a) In the chronological sampling, we choose a cell to trace at the beginning of an experiment. Typically, the cell at the bottom of a chamber is chosen in the case of the mother machine. Then, the cell is traced chronologically until we can no longer trace it by either cell death or other reasons. We can effectively obtain the chronological sample of a lineage from a lineage tree by tracing the root cell and by randomly choosing one of the two daughter cells upon division. (b) In the retrospective sampling, we choose a cell to trace randomly at the end of an experiment. Then, we trace the cell retrospectively to its ancestor cells. Because we choose the cell from the survived population, the retrospective lineage cannot be terminated either by cell death or by out of the observation frame. The length of the retrospective lineage is therefore the same as that of the experiment. (c) In the tree sampling, we sample all the cells but the leaves, the division times of which were not observed, e.g, by the termination of experiment or by the cell death.

be exposed to a disturbed environment, because the bottom of the chamber is far from the flowing fresh medium. Finally, the chronological sampling does not directly observe the selection process induced by the different replication speeds of the cells in a population.

B. Retrospective sampling and retrospective process. These problems can be resolved by using the retrospective sampling of a cell lineage from a proliferating population observed by the dynamics cytometer (Fig. 4 (b)) (18). In the dynamics cytometer, a population of cells is cultured in a more spacious chamber that can accommodates hundreds of the cells, and a cellular lineage tree can be reconstituted from the observed movie. By sampling a cell from the survived cells in the tree, we can always obtain a cell lineage with the same length of the experiment so long as the cell population rather than a cell does not extinct (18, 19). However, the cells in this retrospective lineage are subject to the survivorship bias, because the lineage is sampled from survived cells. Thereby, π_{emp} and $\mathbb{T}_{\mathrm{emp}}$ converge to π_{B} and $\mathbb{T}_{\mathrm{B}}$, which are different from those of the forward process, π_{F} and $\mathbb{T}_{\mathrm{F}}$.

In order to correct the survivorship bias, in this work, we have proved that $\pi_{\mathrm{B}}(\tau|\boldsymbol{x})$ is exponentially biased from $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ as

$$\pi_{\mathrm{B}}(\tau|\boldsymbol{x}) = \frac{2\pi_{\mathrm{F}}(\tau|\boldsymbol{x})e^{-\lambda\tau}}{Z(\boldsymbol{x})}, \qquad [3]$$

where λ is the population growth rate of the cells, and $Z(\boldsymbol{x})$ is a normalization factor (37). See Supplementary Information for the proof (Section S4). This is an extension of Wakamoto *et al.* 2011 (19) in which the states of the cells were not considered. We also have derived that $\mathbb{T}_{\mathrm{B}}(\boldsymbol{x}'|\boldsymbol{x})$ is biased from $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ as

$$\mathbb{T}_{\mathrm{B}}(\boldsymbol{x}'|\boldsymbol{x}) = \frac{u(\boldsymbol{x}')\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})Z(\boldsymbol{x})}{u(\boldsymbol{x})}, \qquad [4]$$

where u is the left eigenvector associated with the largest eigenvalue of the matrix

$$M(\boldsymbol{x}'|\boldsymbol{x}) := \mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})Z(\boldsymbol{x}). \qquad [5]$$

See (37) for the derivation. This is also an extension of our previous work (38) in which the division time was not considered. $\mathbb{T}_{\mathrm{B}}$ and π_{B} together define a semi-Markov process of $\boldsymbol{x}$, which, by construction, asymptotically generates the retrospective cell lineage. Thus, we call this process the retrospective process. See Supplementary Information (Section S4) for the details about the retrospective process.

Eq. [3] shows that the correction of the bias in π_{B} requires the population growth rate λ, which is easily estimated in the dynamics cytometer experiment. From lineage trees obtained by the dynamics cytometer experiment, we can estimate λ as the ratio of the number of the flown cells from the chamber per unit time to that of the whole cells in the chamber. See also Supplementary Information (Section S7). On the other hand, Eq. [4] indicates that the correction of $\mathbb{T}_{\mathrm{B}}$ necessitates u, which can be neither directly observed nor easily estimated. This fact limits the use of the retrospective sampling for estimating the cellular state $\boldsymbol{x}$ and its dynamics. In addition to this limitation, another problem shared by both chronological and retrospective samplings is that only a lineage of the tracked cell is used for the estimation, which requires quite a long-term tracking to obtain a sufficiently large number of sample points, i.e., the cell divisions and the state-switching events. In the case of the dynamics cytometer, especially, it seems a huge waste of the data points to abandon the information of the cells being in the tree but out of the tracked lineage.

C. Tree sampling: estimation from the whole cells in the lineage tree. These problems can be resolved by the tree sampling in which we use all the cells but the leaves in the lineage tree for estimation (Fig. 4 (c)). Here, the leaves correspond to the cells in the tree, the division times of which were not observed, e.g., by the termination of the experiment or flown out from the chamber. Yet to be clarified is the bias in the estimation introduced by using the sample obtained in this way. By employing the many-to-one formulae of the branching process (29, 31), we have proven in this work that π_{emp} converges to π_{B}, whereas $\mathbb{T}_{\mathrm{emp}}$ does to $\mathbb{T}_{\mathrm{F}}$. See Supplementary information for the proof (Section S5 and S6). The converged distributions of the chronological, retrospective, and tree sampling are summarized in Tab. 1.

Owing to the direct convergence of $\mathbb{T}_{\mathrm{emp}}$ to $\mathbb{T}_{\mathrm{F}}$ in this tree sampling, we can circumvent the difficulty of reconstructing $\mathbb{T}_{\mathrm{F}}$ from $\mathbb{T}_{\mathrm{B}}$, while enjoying the large number of the sample points in the tree. Thus, the tree sampling is more efficient than the other samplings. The correction of the bias requires the population growth rate λ. We can obtain λ in two ways. The first way is to estimate λ from the dynamics cytometer experiment as we mentioned before. The second way is to calculate λ analytically from the estimated $\mathbb{T}_{\mathrm{F}}$ and π_{B}. See also Supplementary Information (Section S7).

Table 1. Comparison of the converged distributions obtained by the chronological, the retrospective, and the tree samplings.

	chronological	retrospective	tree
Division time	π_{F}	π_{B}	π_{B}
State switching	$\mathbb{T}_{\mathrm{F}}$	$\mathbb{T}_{\mathrm{B}}$	$\mathbb{T}_{\mathrm{F}}$

4. Estimation of latent states from a lineage tree

In the preceding section, we have clarified the converged distributions for different samplings under the assumption that the states of the cells as well as the division times are experimentally observed. However, the information of the states of the cells may not always be accessible. Even when we observe the expression of several genes over lineages, such genes may not be sufficiently relevant to the determination of the division times, because the division time is generally a consequence of the complicated interactions of intracellular genetic and metabolic networks. Moreover, even if we could observe the high dimensional state over a lineage, we would have to make it interpretable by finding the low dimensional relevant representation of the states to the division times; which generally requires a huge computational cost.

Such problems can be handled by inferring the effective states of the cells based only on the division time observations. By extending the EM algorithm for the hidden Markov models (39) to a branching tree with hidden states, we construct an algorithm, **Lineage EM algorithm (LEM)**, for estimating the latent states of the cells in a lineage tree and their dynamic laws. To this end, we introduce the following parametric models with discrete or continuous state-spaces, which enable us to employ well-established statistical methods, e.g. maximum likelihood estimation (MLE), for the estimation.

A. A parametric discrete state-space model. For a discrete state-space model, we assume that π_{F} belongs to an exponential family (39). The exponential family includes a broad range of probability distributions such as the gamma-distribution and the log-normal distribution, which have been commonly used for fitting the division time distributions of microbes (19, 40). By assuming a parametric model, the estimation of $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ is reduced to that of the parameter set of the model. The gamma distribution is a common choice of the parametric model of the division time distribution:

$$\mathbb{P}_G(\tau;\boldsymbol{\theta}) = \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}, \qquad [6]$$

where $\Gamma(a)$ is the gamma function and $\boldsymbol{\theta} := (a, b)$, a and b of which are the shape and rate parameters, respectively.

Then, the division time distribution for the forward process $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ is represented by a $\boldsymbol{x}$-dependent parameter set $\boldsymbol{\theta}^F_{\boldsymbol{x}} = (a_{\boldsymbol{x}}, b_{\boldsymbol{x}})$ as

$$\pi_{\mathrm{F}}(\tau|\boldsymbol{x}) = \mathbb{P}_G(\tau|\boldsymbol{\theta}^F_{\boldsymbol{x}}). \qquad [7]$$

When $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ is a gamma distribution, so is π_{B} with a different parameter set, $\boldsymbol{\theta}^B_{\boldsymbol{x}}$, as

$$\pi_{\mathrm{B}}(\tau|\boldsymbol{x}) = \mathbb{P}_G(\tau|\boldsymbol{\theta}^B_{\boldsymbol{x}}). \qquad [8]$$

Thereby, we can covert $\boldsymbol{\theta}^B_{\boldsymbol{x}}$ to $\boldsymbol{\theta}^F_{\boldsymbol{x}}$ via Eq. [3] after estimating $\boldsymbol{\theta}^B_{\boldsymbol{x}}$. On the other hand, the state-switching can be straight-forwardly represented by the components of the matrix, $\mathbb{T}_{\mathrm{F}}$.

B. A parametric continuous state-space model. Suppose that the continuous state space Ω is k-dimensional Euclidean as $\Omega \subseteq \mathbb{R}^k$. Because the estimation by considering all the possible dynamics in a continuous state-space is unfeasible, we here adopt a linear diffusion dynamics for the state-switching, $\mathbb{T}_{\mathrm{F}}$, which is characterized by a $k \times k$ matrix $\boldsymbol{A}$ as

$$\boldsymbol{x}' = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{w}, \qquad [9]$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the states of a mother and its daughter cell, respectively. $\boldsymbol{w}$ is a multidimensional Gaussian random variable with a mean vector $\boldsymbol{0}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{w}}$.

The retrospective distribution of the division time, π_{B}, is also assumed to follow a log-normal distribution:

$$\log \tau = \boldsymbol{C}\,\boldsymbol{x} + \boldsymbol{v}, \qquad [10]$$

where $\boldsymbol{C}$ is a $1 \times k$ matrix and $\boldsymbol{v}$ is a Gaussian random variable with mean 0 and variance $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. In this model, the estimation problem is reduced to estimating parameters $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{\Sigma}_{\boldsymbol{w}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, simultaneously. We remark that the expectation of $\log \tau$ is non-zero due to the non-zero expectation of the initial state $\boldsymbol{x}_{\mathrm{root}}$ at the root of the lineage. This setting can be interpreted as a linear approximation of a general continuous-state model of $\boldsymbol{x}$.

C. Lineage EM algorithm. To obtain LEM, we extend the Baum-Welch algorithm (BW algorithm) to the estimation of $\boldsymbol{\theta}^B_{\boldsymbol{x}}$ and $\mathbb{T}_{\mathrm{F}}$ from a lineage tree. LEM algorithm iterates two steps, the E-step and the M-step, and updates the parameters until convergence. Let $\Theta^{(n)}$ denote the estimate of the parameters $(\mathbb{T}_{\mathrm{F}}, \{\boldsymbol{\theta}^B_{\boldsymbol{x}}\})$ after the nth iteration. In the E-step, we compute the posterior probabilities of the states for all the pairs of the mother and daughter cells, $\xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')$, conditioned on the currently estimated parameters $\Theta^{(n)}$ and observation. $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')$ are the states of the cell i and its one of the daughter labeled as the cell j, respectively. $\gamma_i(\boldsymbol{x})$ is the posterior probability of the state of the cell i, which is obtained by marginalization as $\gamma_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}'} \xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')$. $\xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')$ and $\gamma_i(\boldsymbol{x})$ are computed via the belief propagation (39). The belief propagation recursively computes the posterior distributions efficiently for a graphical model without loops. LEM belongs to this class, because a tree is loopless. See Supplementary Information for the detail (Section S8 and S10). For the continuous state-space model, we can employ the well-established estimation technique of the Kalman filter (39). In the M-step, the parameters $\Theta^{(n)} = (\mathbb{T}_{\mathrm{F}}, \{\boldsymbol{\theta}^B_{\boldsymbol{x}}\})$ is updated so that $\pi_{\mathrm{B}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ are fitted to the following modification of the empirical distributions, respectively (1):

$$\pi^{\mathrm{BW}}_{\mathrm{emp}}(\tau|\boldsymbol{x}) := \frac{1}{\sum_{i \in \mathcal{T}_{\boldsymbol{x}}} \gamma_i(\boldsymbol{x})} \sum_{i \in \mathcal{T}_{\boldsymbol{x}}} \gamma_i(\boldsymbol{x}) \delta(\tau - \tau_i), \qquad [11]$$

$$\mathbb{T}^{\mathrm{BW}}_{\mathrm{emp}}(\boldsymbol{x}'|\boldsymbol{x}) := \frac{\sum_{i,j} \xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')}{\sum_{i,j,\boldsymbol{x}'} \xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')}, \qquad [12]$$

where $\mathcal{T}_{\boldsymbol{x}}$ is the set of all non-leaf cells with state $\boldsymbol{x}$ in the lineage tree, and (i, j) in the second equation runs over all the mother-daughter pairs. These are empirical distributions weighted by the posterior distributions $\gamma_i(\boldsymbol{x})$ and $\xi_{i,j}(\boldsymbol{x}, \boldsymbol{x}')$. For the details on the fitting process by MLE, see Supplementary Information (Section S9). It is known that each update always increases the likelihood (39). In the continuous case, we update $\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Sigma}_{\boldsymbol{w}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in the same way, that is, update the parameters so that $\pi_{\mathrm{B}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ are fitted to $\pi^{\mathrm{BW}}_{\mathrm{emp}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}^{\mathrm{BW}}_{\mathrm{emp}}(\boldsymbol{x}'|\boldsymbol{x})$, respectively. We remark that the computational time of LEM grows linearly with respect to the number of the cells in the lineage (39).

5. Validation of LEM

We check the validity of our LEM by the estimation from synthetic data. We also give a validation by the estimation from real lineage trees of *E. coli* and its biological consequences in the Supplementary Information S15.

A. Validation of LEM with synthetic data sets. We tested the validity of LEM by numerical experiments of the discrete-state model. We consider the situation that each cell has two states: a fast-growing ($\boldsymbol{x} = f$) and slow-growing ($\boldsymbol{x} = s$) states as depicted in Fig. 5 (a) and obtained a synthetic lineage tree as shown in Fig. 5 (b). By applying LEM to the lineage tree in Fig. 5 (b), we could recover the states of the cells from the tree as in Fig. 5 (c) without using any state information of the cells. The states are reliably inferred from the tree containing an experimentally reasonable number of cells, e.g., 500 cells. See Supplementary Information for the details (Section S12). The states of the leaf cells cannot be inferred in Fig. 5 (c), because the division times of the leaf cells were not observed. If the state information of the leaves is supplemented for the inference, the accuracy of the estimation is further improved as in Fig. 5 (d). Such information on the states of the leaves may be obtained by conducting single-cell staining or scFISH (41), or scRNA sequencing at the end of the experiment, as assumed in the previous attempts of the state inference from lineage trees (21, 22). The convergence of the log-likelihoods was also checked for both situations (Figs. 5 (e) and (f)). We note that negligible decrease of the likelihood occurs in Fig. 5 (f) due to numerical errors. We have further compared the empirical and estimated retrospective distributions of the division times (Figs. 5 (g) and (h)) to verify good coincidences between the empirical and the estimated distributions. Finally, we estimated $\mathbb{T}_{\mathrm{F}}$ and π_{F} of the model in Fig. 5 (a) for 1000 times to evaluate the accuracy of our estimation. See Supplementary Information (Section S12). We observed that the estimation is consistent with the true parameter in total by virtue of the correction of the survivorship bias.

We also tested the validity of LEM by using the model shown in Fig. 6 (a), in which the survivorship bias is larger. This is the same model as in Fig. 2 (a). By applying LEM to the lineage tree in Fig. 6 (b), we could recover the states of the cells from the tree as in Fig. 6 (c) without using any state information of the cells. We estimated $\mathbb{T}_{\mathrm{F}}$ and π_{F} of the model in Fig. 6 (a) for 100 times to evaluate the accuracy of our estimation. See Supplementary Information (Section S12). The survivorship bias is corrected by our method (Fig. 6 (d)). In conclusion, our LEM works even for the lineage trees with large survivorship bias.

We note that LEM terminates within a few seconds in this numerical experiments. See Supplementary information (Section S11) for the environment where LEM was executed.

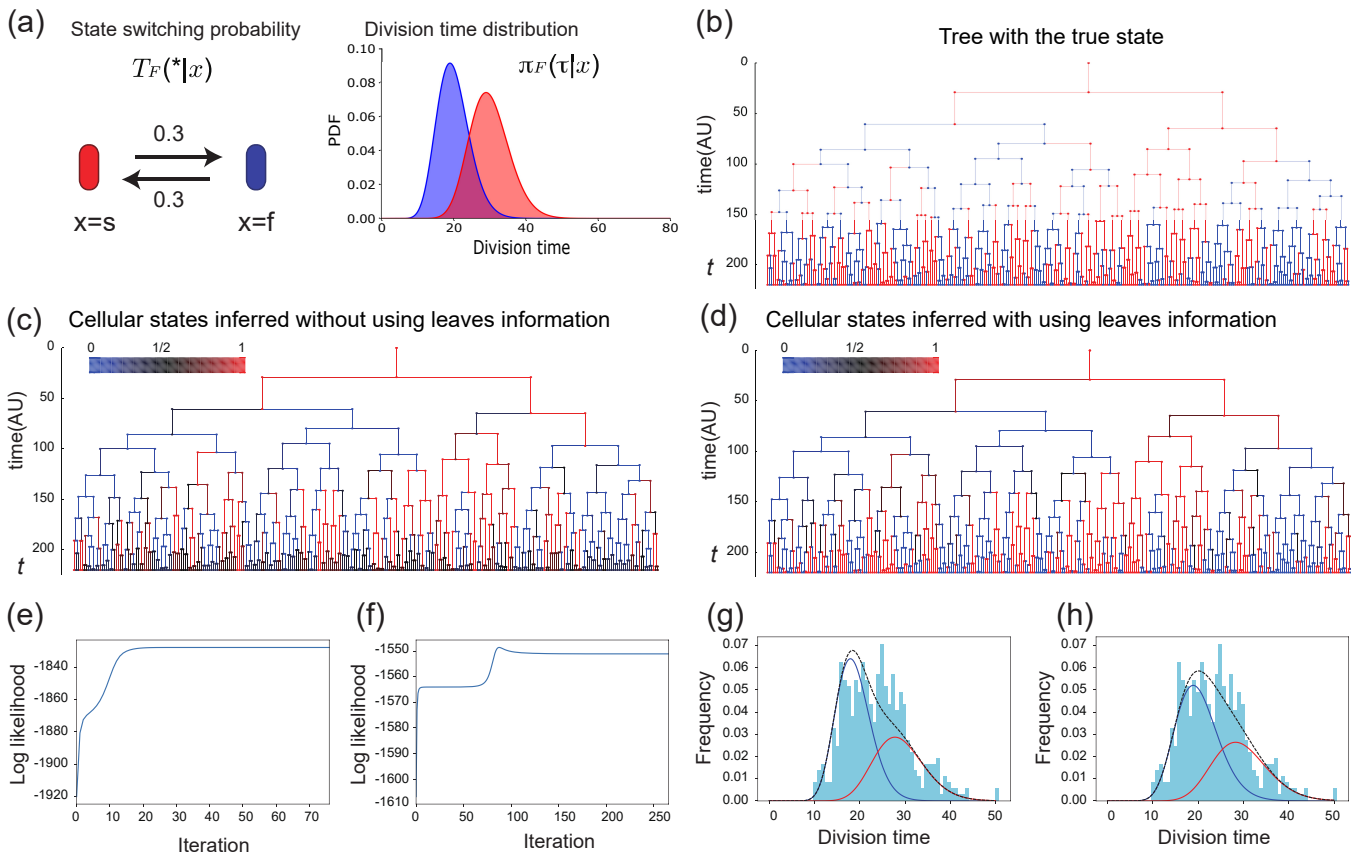

Fig. 5. A performance evaluation of LEM by comparing a simulated lineage tree of the discrete model and the corresponding inferred states of the cells with and without using the information of the states of the leaves. (a) A schematic diagram of the state-switching dynamics and their division time distributions used for the evaluation. Each cell is supposed to have either slowly growing (red) or quickly growing (blue) state, and the transitions between them occur with the probability 0.3. (b) A synthetic lineage tree of the model in (a) obtained by simulating the corresponding branching process. (c, d) The lineage trees with the latent states inferred from the tree in (b) without (c) and with (d) the information of the actual states of the leaf cells. The color on a segment indicates the probability that the state of the cell corresponding to the segment is in the red state. Red (blue) means that the cell is estimated to be in the red (blue) state with a high probability, whereas black means that the estimated state is ambiguous. (e, f) The convergence of the log-likelihoods when the states of the cells at the leaves are not available (e) and are available (f). (g, h) The empirical and the inferred distributions of the division time when we do not know the states of the leaf cells (g) and when we know the states of the leaf cells (h). The red and the blue curves show the relative frequencies of the division time of the red and blue states, and the black one is their mixture. The histogram is the empirical distribution of the division times of the cells on the tree.

B. Validation of LEM for continous state model. We checked the validity of LEM for the continuous state model by estimating parameters from synthetic data. We generated synthetic data as follows. The lineage tree generated was a perfect binary tree with 10 generations. The state space is two dimensional as $\Omega = \mathbb{R}^2$. The state of the root cell was set to be $(1, 0)^{\mathrm{T}}$. We adopted the following parameters to generate the synthetic lineage tree:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{C} = \begin{pmatrix} 1 & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{w}} = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}, \quad \Sigma_v = 0.01. \tag{13}$$

We estimated the parameters, i.e., $\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Sigma}_{\boldsymbol{w}}$, and Σ_v, by LEM. The initial values of the parameters used in the algorithm were randomly chosen. See Supplementary Information (Section S13). We ran the estimation algorithm for 1000 different initial values and adopted the parameters with the largest likelihood. LEM terminates in about an hour in this validation.

The estimated parameters became

$$\boldsymbol{A} = \begin{pmatrix} 1.040 & 1.093 \\ -0.002 & 0.962 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{w}} = \begin{pmatrix} 0.010 & 0 \\ 0 & 0.011 \end{pmatrix}, \quad \Sigma_v = 0.008. \tag{14}$$

The estimated parameters are sufficiently close to the true parameters. This result validates our LEM for the continuous model.

To see the dependence of the estimated parameters on the sample, we estimated parameters by LEM from 100 synthetic lineage trees generated independently from the same model and the same parameter set (Eq. [13]) as above. To check the

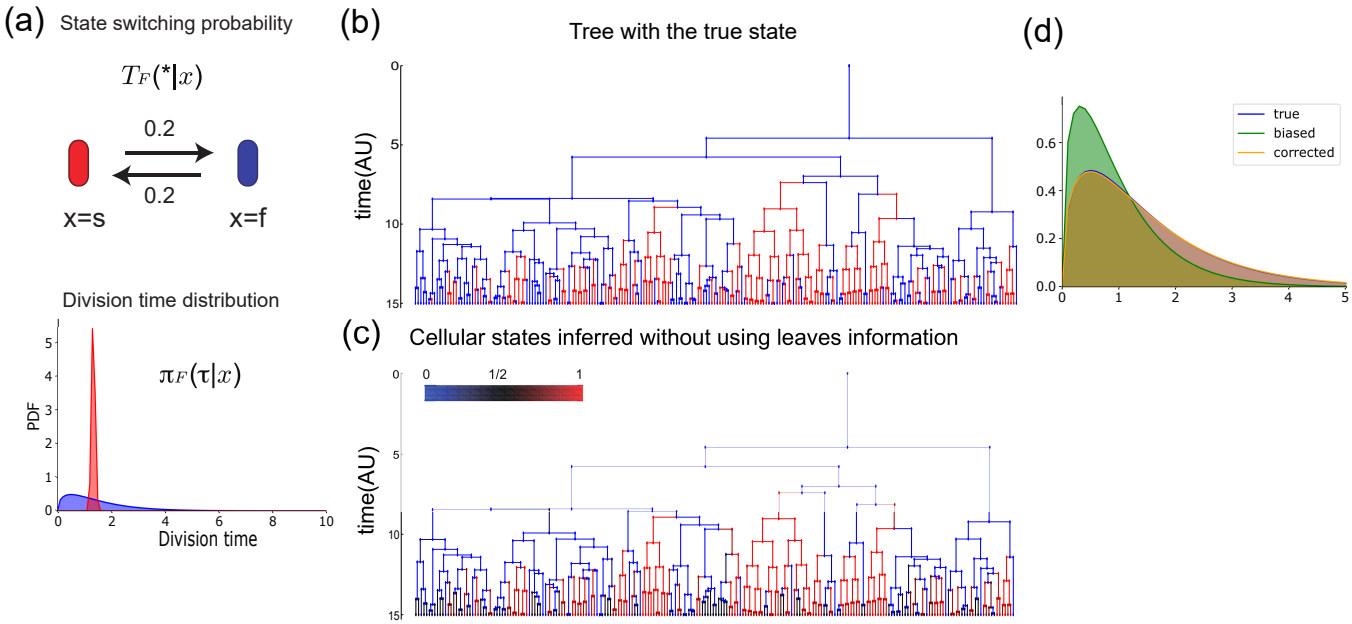

Fig. 6. A performance evaluation of LEM by comparing a simulated lineage tree of the discrete model in which the survivorship bias is larger. (a) A schematic diagram of the state-switching dynamics and their division time distributions used for the evaluation. This is the same model as Fig. 2 (a). (b) A synthetic lineage tree of the same model as Fig. 2 obtained by simulating the corresponding branching process. (c) The lineage trees with the latent states inferred from the lineage tree without the information of the actual states of the leaf cells. (d) Correction of the survivorship bias. The blue distribution shows the true distribution of division time (which is almost identical to the corrected distribution), the green shows the estimated distribution by naïve EM-algorithm (without correction of the survivorship bias), and the orange shows the estimated distribution by LEM.

sample-dependency efficiently, we reduced the number of the initial values in the estimation algorithm from 1000 to 100. The histogram of the estimated parameter values are summarized in Fig. 7. For each parameter, the estimated value distributed around the true parameters. Moreover, the estimated $\boldsymbol{A}$ conserved the non-zero structure of the true parameter $\boldsymbol{A}$. This result also shows the validity of LEM for the continuous model. We also checked that the estimated states by LEM are close to the true states. See Supplementary Information (Section S14).

C. Application to E.coli lineage tree. Finally, in order to demonstrate applicability of our method to real data, we inferred the latent states of the *E. coli* cells in the lineage trees observed by using the dynamics cytometer in Hashimoto *et al.* (18) (Fig. 2 (d)). By applying LEM of the continuous-state model, in which the dimension k of the state space $\Omega = \mathbb{R}^k$ was determined by AIC (42), we found $k = 3$ to be the dimension of the best continuous model. LEM terminate in about an hour in this estimation. Here, we used the continuous-state model since the discrete-state model cannot explain the correlation between the division time of a mother cell and that of the daughter cells. See Supplementary information (Section S15). We also obtained the inferred dynamics of the latent state $\boldsymbol{x}$ over the lineage tree as in Fig. 8 and its parameter values as follows:

$$\boldsymbol{A} = \begin{pmatrix} -0.731 & 0.438 & 0.032 \\ -2.51 & 1.124 & 0.062 \\ -0.262 & 0.0068 & 1.007 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix},$$
$$\boldsymbol{\Sigma}_w = \begin{pmatrix} 0.055 & 0 & 0 \\ 0 & 0.038 & 0 \\ 0 & 0 & 0.016 \end{pmatrix}, \quad \Sigma_v = 0.04. \qquad [15]$$

We verified the statistical significance of the estimated dimension and the parameter values in Supplementary Information (Section S15). Of the three components of the inferred latent state, the first one has the fastest time-scale of approximately one generation (Fig. 8 (a)), whereas the third one changes slowly over generations (Fig. 8 (c)). Investigation of the inferred dynamics also indicates that the third component of a cell encodes a certain amount of information on the division pattern of its descendants. See also Supplementary Information (Section S15) for the detail. This result is consistent with the slow dynamics of certain intracellular proteins, which might affect the noisy behavior of the division times and being inherited over generations (43). Moreover, the inferred states here can be interpreted as the underlying effective inheritance dynamics that can capture the correlation of the division times between the mother-daughter pairs. Such correlation between generations can also be modeled more directly without the latent state by assuming the conditional dependence of the division time of a daughter τ_0 on that of the mother τ as $\pi_{\mathrm{F}}(\tau_0|\tau)$ (12, 13). However, the latent states can offer a way to link the identified states with intracellular physical quantities such as the expressions of candidate proteins. This link may substantially facilitate our understanding of how the reproductive capabilities of the cells are determined, regulated, and inherited as the consequences of the intracellular networks.

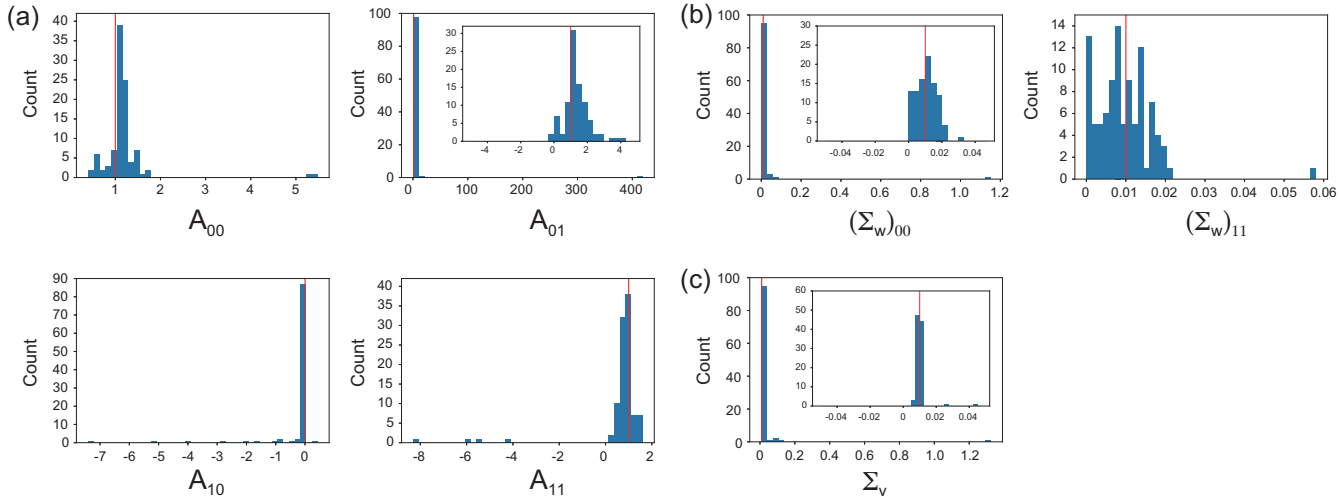

Fig. 7. The histograms of the estimated values of the parameters, $\boldsymbol{A}'$ (a), $\boldsymbol{\Sigma}'_{\boldsymbol{w}}$ (b), and Σ'_v (c). The true value of each parameter is indicated by the red line. The vertical axes show the counts of the estimated values of the parameters of the 100 trials. The insets are the blow-ups of the corresponding histograms around their peaks.

6. Summary and Discussion

In this study, we have derived and proposed LEM, a statistical method to infer the latent states of the cells, the associated state-switching, and division dynamics from lineage tree data. To this end, we combined the correction method of the survivorship bias with the EM algorithm for trees. The accuracy and consistency of the method were verified by using the synthetic tree data with two distinct states or by a continuous state model. By applying the method to the lineage tree of *E. coli*, we have identified the latent low-dimensional states of the cells, which are inherited over a couple of generations at least.

LEM provides a data-driven way to identify and to characterize individual cells in apparently similar yet latently distinct states in a growing population. Cells in the distinctive modes of the growth, e.g., vegetative and dormant ones, have been identified manually and shown to have different susceptibility to stresses (5, 44). Recent experimental investigations have further suggested that more subtle differences are still ruling the fates of the cells under the challenge of antibiotics (6). LEM combined with the dynamics cytometer may play the pivotal roles to investigate the more complicated processes of the cellular natural selections occurring in the populations of bacteria, pathogen, immune cells, and cancer cells (14).

LEM still leaves room for further improvements that extend its applicability to various problems, some of which may be addressed by using existing techniques of the hidden Markov models. For instance, we may relax the assumption of the independence of the state-switching between the daughter cells (26, 45). This generalization may be useful when we include the size of a cell as a state, which naturally correlates between the daughters (46, 47). We may also extend LEM either to include other experimentally observed quantities than the division times for the estimation of the latent states or to combine the observed quantities as the visible state with the latent states. The assumption of the linear dynamics in the continuous model or that of the exponential families for the division time distribution can be generalized to incorporate realistic nonlinear dynamics or non-parametric distributions by using Monte-Carlo or ensemble methods at the cost of heavy computational loads (28, 48).

Moreover, we still have biologically important but theoretically challenging problems: One of the problem is the state-dependent death rate of the cells. We anticipate that the analysis of the survivorship bias still can be carried over to such situation and conjectures that if a cell dies with a rate $\gamma(\boldsymbol{x})$, then the empirical distributions of the generation time converges to $\pi_{\mathrm{B}}(\tau \mid \boldsymbol{x}) = \pi_{\mathrm{F}}(\tau \mid \boldsymbol{x}) e^{-(\lambda+\gamma(\boldsymbol{x}))\tau}$. This π_{B} may be again characterized as the retrospective path from a uniformly chosen cell at the end of an infinitely large lineage. A proof of this conjecture is indispensable for addressing the impact of the antibiotics. Another is that the feedback from the division time to the latent state transition, which naturally occurs when the latent state is affected how long the next division occurs. The feedback inevitably destroys the prerequisite of the BW algorithm that there is no feedback. All these problems together with the potential applicability of LEM may open up a new target of the machine learning and the statistics, which will provide quantitative and data-drive ways to address the problems of evolutionary and systems biology.

Availavility of Materials

An implementation of LEM is available at https://github.com/so-nakashima/Lineage-EM-algorithm.

ACKNOWLEDGMENTS. We greatly thank Yuichi Wakamoto for providing the data of *E. coli* for this work and helpful discussion on our results. This research is supported by JSPS KAKENHI Grant Number JP16K17763, 16H06155, 19H03216, and 19H05799 and by JST Grant Number JPMJPR15E4 and JPMJCR1927 Japan.

1. Kærn M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* 6:451 EP –. Review Article.

2. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* 135(2):216 – 226.
3. Shahrezaei V, Swain PS (2008) The stochastic nature of biochemical networks. *Current Opinion in Biotechnology* 19(4):369 – 374. Protein technologies / Systems biology.
4. Bigger JW (1944) Treatment of staphyloeoeeal infections with penicillin by intermittent sterilisation. *Lancet* pp. 497–500.
5. Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. *Science* 305(5690):1622–1625.
6. Wakamoto Y, et al. (2013) Dynamic persistence of antibiotic-stressed mycobacteria. *Science* 339(6115):91–95.
7. Harms A, Maisonneuve E, Gerdes K (2016) Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science* 354(6318).
8. van Boxtel C, van Heerden JH, Nordholt N, Schmidt P, Bruggeman FJ (2017) Taking chances and making mistakes: non-genetic phenotypic heterogeneity and its consequences for surviving in dynamic environments. *Journal of The Royal Society Interface* 14(132).
9. Brock A, Chang H, Huang S (2009) Non-genetic heterogeneity – a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics* 10:336 EP –. Perspective.
10. Sharma SV, et al. (2010) A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 141(1):69 – 80.
11. Fisher RA, Gollan B, Helaine S (2017) Persistent bacterial infections and persister cells. *Nature Reviews Microbiology* 15:453 EP –. Review Article.
12. Jolly MK, Kulkarni P, Weninger K, Orban J, Levine H (2018) Phenotypic plasticity, bet-hedging, and androgen independence in prostate cancer: Role of non-genetic heterogeneity. *Frontiers in Oncology* 8:50.
13. Müller V, de Boer RJ, Bonhoeffer S, Szathmáry E (2017) An evolutionary perspective on the systems of adaptive immunity. *Biological Reviews* 93(1):505–528.
14. Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. *Nature Ecology &Amp; Evolution* 1:0077 EP –. Perspective.
15. Reyes J, Lahav G (2018) Leveraging and coping with uncertainty in the response of individual cells to therapy. *Current Opinion in Biotechnology* 51:109 – 115. Systems biology • Nanobiotechnology.
16. Rowat AC, Bird JC, Agresti JJ, Rando OJ, Weitz DA (2009) Tracking lineages of single cells in lines using a microfluidic device. *Proceedings of the National Academy of Sciences* 106(43):18149–18154.
17. Wang P, et al. (2010) Robust growth of escherichia coli. *Current Biology* 20(12):1099 – 1103.
18. Hashimoto M, et al. (2016) Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113(12):3251–3256.
19. Wakamoto Y, Grosberg AY, Kussell E (2011) Optimal lineage principle for age-structured populations. *Evolution* 66(1):115–134.
20. Lambert G, Kussell E (2015) Quantifying selective pressures driving bacterial evolution using lineage analysis. *Phys Rev X* 5(1):011016. 26213639[pmid].
21. Hormoz S, Desprat N, Shraiman BI (2015) Inferring epigenetic dynamics from kin correlations. *Proceedings of the National Academy of Sciences* 112(18):E2281–E2289.
22. Hormoz S, et al. (2016) Inferring cell-state transition dynamics from lineage trees and endpoint single-cell measurements. *Cell Systems* 3(5):419 – 433.e8.
23. Hicks DG, Speed TP, Yassin M, Russell SM (2018) Statistical inference in cell lineage trees. *bioRxiv*.
24. Hicks DG, Speed TP, Yassin M, Russell SM (2018) Maps of variability in cell lineage trees. *bioRxiv*.
25. Olariu V, et al. (2009) Modified variational bayes em estimation of hidden markov tree model of cell lineages. *Bioinformatics* 25(21):2824–2830.
26. Failmezger H, et al. (2018) Clustering of samples with a tree-shaped dependence structure, with an application to microscopic time lapse imaging. *Bioinformatics* p. bty939.
27. Kuzmanovska I, Milias-Argeitis A, Mikelson J, Zechner C, Khammash M (2017) Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC Systems Biology* 11(1):52.
28. Kuchen EE, Becker N, Claudino N, Hofer T (2018) Long-range memory of growth and cycle progression correlates cell cycles in lineage trees. *bioRxiv*.
29. Marc H, Adélaïde O (2016) Nonparametric estimation of the division rate of an age dependent branching process. *Stochastic Processes and their Applications* 126(5):1433 – 1471.
30. Thomas P (2017) Making sense of snapshot data: ergodic principle for clonal cell populations. *Journal of The Royal Society Interface* 14(136).
31. Marguet A (2016) Uniform sampling in a structured branching population. *ArXiv e-prints*.
32. Marguet A (2017) A law of large numbers for branching Markov processes by the ergodicity of ancestral lineages. *ArXiv e-prints*.
33. Thomas P (2018) Population growth affects intrinsic and extrinsic noise in gene expression. *bioRxiv*.
34. Thomas P (2018) Analysis of cell size homeostasis at the single-cell and population level. *Frontiers in Physics* 6:64.
35. Nozoe T, Kussell E, Wakamoto Y (2017) Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data. *PLOS Genetics* 13(3):1–25.
36. Harris TE (1963) *The Theory of Branching Processes*. (Springer-Verlag Berlin Heidelberg).
37. Sughiyama Y, Nakashima S, Kobayashi TJ (2018) Fitness response relation of a multi-type age-structured population dynamics. *ArXiv e-prints*.
38. Sughiyama Y, Kobayashi TJ, Tsumura K, Aihara K (2015) Pathwise thermodynamic structure in population dynamics. *Phys. Rev. E* 91(3):032120.
39. Christopher B (2006) *Pattern Recognition and Machine Learning*. (Springer-Verlag New York).
40. Rubinow SI (1968) A maturity-time representation for cell populations. *Biophys J* 8(10):1055–1073. 5679389[pmid].
41. Frieda KL, et al. (2016) Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541:107 EP –.
42. Akaike H (1998) *Information Theory and an Extension of the Maximum Likelihood Principle*, eds. Parzen E, Tanabe K, Kitagawa G. (Springer New York, New York, NY), pp. 199–213.
43. Taniguchi Y, et al. (2010) Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.
44. Balaban N (2011) Persistence: mechanisms for triggering and enhancing phenotypic variability. *Current Opinion in Genetics & Development* 21(6):768 – 775. Genetics of system biology.
45. Paskin MA (2002) Thin junction tree filters for simultaneous localization and mapping, (EECS Department, University of California, Berkeley), Technical Report UCB/CSD-02-1198.
46. Taheri-Araghi S, et al. (2015) Cell-size control and homeostasis in bacteria. *Current Biology* 25(3):385 – 391.
47. Susman L, et al. (2018) Individuality and slow dynamics in bacterial growth homeostasis. *Proceedings of the National Academy of Sciences* 115(25):E5679–E5687.
48. Särkkä S (2013) *Bayesian filtering and smoothing*. (Cambridge University Press) Vol. 3.

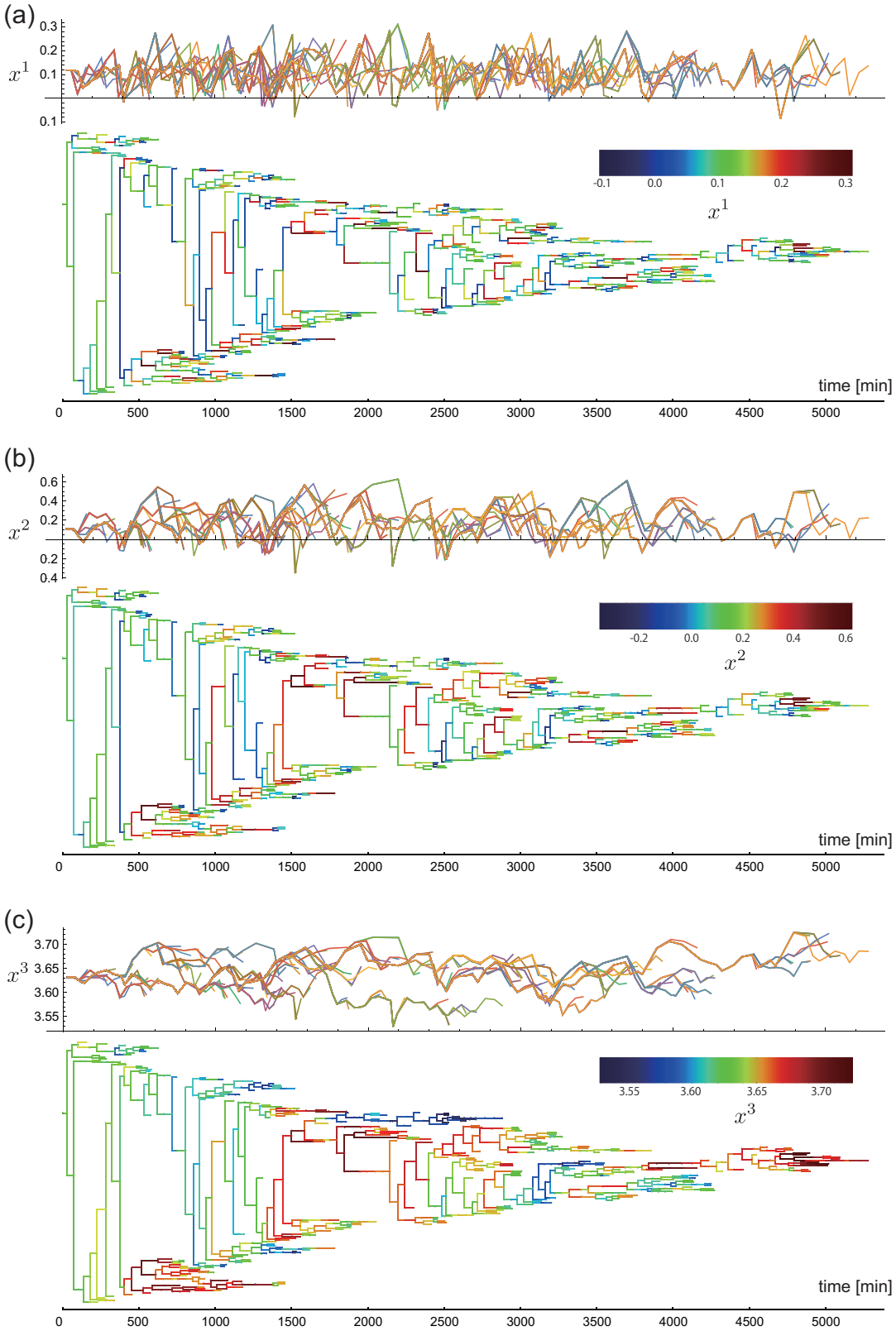

Fig. 8. The dynamics of the first (a), the second (b), and the third (c) components of the inferred three-dimensional state depicted as time series (upper panel) and overlaid on the lineage tree (lower panel). The color codes over the trees represent the actual values of the components.