Deciphering Latent Growth-States from Cellular Lineage Trees

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

[a]Graduate School of Information Science and Technology, Department of Mathematical Informatics, 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

Individual cells in a population generally have different replicative capability, presumably due to the phenotypic variability of the cells. Identifying the latent states that rule the replicative capability and characterizing how the states are inherited over generations are crucial for understanding how the self-replication of the cells is modulated and controlled for achieving higher fitness and resistance to different kinds of perturbations. Even with technological development to monitor the proliferation of single cells over tens of generations and to trace the lineages of cells, estimating the state of the cells is still hampered by the lack of statistical methods that can appropriately account for the lineage specific problems. In this work, we develop a statistical method to infer the growth-related latent states of cells over a cellular lineage tree concurrently with the switching dynamics of the states and the statistical law how the state determines the division time. An application of our method to a lineage data of *E.coli* has identified a three dimensional effective state in the cells, one component of which seems to capture slow fluctuation of cellular state over generations.

1. Introduction

A population of cells is phenotypically heterogeneous even if they are genetically identical (1–3). Such a phenotypic variability can work as the bet-hedging of the cells under an unpredictably changing environment, the typical example of which is the bacterial persistence, the survival of the slowly growing but resistant cells against challenges of antibiotics (4–8). More generally, the heterogeneities in the self-replication speed and the death rate as well as their inheritance from a mother to daughter cells constitute the Darwinian natural selection among the cells. The natural selection at the cellular level is also highly relevant for drug-resistances of pathogens and cancers, the establishment of immunological memories, and cell competitions in tissues (9–13). Therefore, quantification of the replicative and survival capabilities, which are often identified with the fitness, from data is crucial for predicting and controlling these phenomena ruled by the micro-evolution of the cells (14, 15).

However, defining the replicative capabilities at the level of individual cells from data is by no mean trivial in the face of the stochastic nature of the cellular replication, even if we can access to the observations of the actual division times. The division times of cells in a presumably identical phenotypic state can still vary stochastically, and thereby, the division times of a cell in an unknown state cannot be used as the proxy of the capability of the cell. Observations of phenotypic states of the cell, e.g., by bioimaging, may not always help to resolve this problem, because the replication is a consequence of the tangled interplay among high dimensional metabolic and regulatory networks in the cell (16–18). The observed low dimensional quantities may not be related sufficiently to the replicative capability of the cell under a given situation. Even in the context of the evolutionary biology, moreover, defining fitness to individual agents in a population is one of the central problems

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

that have not yet been solved (19, 20).

We address this problem in this work by framing it as an inference of growth-related latent states of cells from data of cellular lineage trees. The determinant of discriminating the replicative capability of a cell from mere stochasticity in the division time is its inheritance to the descendants. The replicative capability of a mother cell should be somehow inherited to its daughter cells, whereas the stochasticity should be independent among the mother and the daughter cells. Such structure can be effectively captured by considering the inheritance of the latent state of the mother $\boldsymbol{x}$ to that of a daughter $\boldsymbol{x}'$ and the stochastic determination of the division time τ conditioned by the state of the cell $\boldsymbol{x}$ (Fig. 1 (a)). While any latent state $\boldsymbol{x}$ of a cell is determined by the high dimensional dynamics of the intracellular metabolites and molecules, the state $\boldsymbol{x}$ supposed here is an effective low dimensional one that is relevant overall for the replication speed of the cell. We introduce a stochastic transition matrix, $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$, and a conditional distribution, $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$, to represent the inheritance of the states and the stochasticity of the division time, respectively (Fig. 1 (a)).

Recent advancements in the microfluidic technology enable us to trace replicating cells over a hundred generations, which offer the data samples to be used for the inference (21–23). Among others, the most widely used device is the mother machine in which a replicating cell is trapped at the bottom of a narrow chamber (23). By flowing the daughter cells away, we can trace the founder cell at the bottom over tens of generation as long as it is alive, and can obtain samples of the division times over the lineages from multiple founders in parallel (24–26). Another devise is the dynamics cytometer, in which a population of cells are accommodated in a more spacious chamber (21, 27, 28) (Fig. 1 (b)). Tracking of the cells in the chamber reconstitutes the tree of lineages, which contains more detailed information on the parent-daughter relationship of the cells and on the actual competition among the cells (Fig. 1 (c,d)). The estimation of the latent states of the cells from the tree enables us to capture which cells with which states have survived in the population; that is an indispensable step towards understanding how natural selection works over the population.

However, the inference of the states from a lineage tree is accompanied by two difficulties. First, the estimation with respect to 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 (29, 30), an algebraic invariance of the lineage tree (31, 32), clustering algorithms (33, 34), Monte-Carlo algorithms (35), and model selection (36). 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 (Fig. 1 (c,d)). Therefore, clades of the cells replicating faster are represented more than the others in the tree, which inevitably introduces bias in the data sample. This is the so-called survivor (or survivorship) bias in statistics such that winners are overrepresented in a population whereas the losers are underrepresented (21, 27, 37, 38) (Fig. 1 (c)). Previous works circumvent this difficulty 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. This process inevitably loses the information on the replicative capabilities. The construction of a correction method of this bias contributes not only to accurate inference but also to estimation of the selection pressure as the strength of the bias (28, 39). The survivor bias in a growing system with states has be analyzed only recently (40–43). Thus, this topic is still immature and an appropriate correction method is yet to be developed.

To address these problems, we first clarify how the survivor bias distorts statistical estimations depending on the way to collect a sample of cells from the tree. By deriving explicit relations between unbiased and biased estimates, we establish a correction method of the survivor bias under the condition that the sates of the cells are known. Then, we propose an estimation algorithm of the latent states from lineage trees based on an estimation-maximization (EM) algorithm, which we call **Lineage EM algorithm (LEM)**. We verify the effectiveness of LEM by using synthetic data. 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. The inferred dynamics suggests that the homeostasis and heterogeneity of the division times are controlled with multiple time scales.

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 the state switching. 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. Upon the division of the mother cell, each daughter cell switches its state stochastically. 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. 1 (a)). 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. 1 (a)). By supposing the generation of two daughter cells upon the division of a mother, $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ and $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ define a multi-type age-dependent branching process(Fig. 1 (c)) (44), whereas they also constitute a continuous semi-Markov process if one of the daughter cells is ignored (Fig. 1 (a)) (45). See Supplimentary Informaion (Section 1 and 2). 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 determination of 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.

3. Correction of the survivor 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 $\boldsymbol{x}'$ from a given data set, respectively:

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

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

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. 2 (a,b,c)), 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 (23). 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. 1 (a)). See Supplementary Information (Section 1 and 2). 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 (21, 28, 45). We can also effectively obtain a chronologically sampled lineage from the tree by using the weighting technique proposed in (39).

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. 2 (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 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 the 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. 2 (b)) (21). 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 (21, 27). However, the cells in a retrospective lineage are subject to the survivor bias, because the lineage is sampled from a survived cell. 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 survivor 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 (45). See Supplementary Information for the proof (Section 3). This is an extension of Wakamoto *et al.* 2011 (27) 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 (45) for the derivation. This is also an extension of our previous work (46) in which the division time was not considered. $\mathbb{T}_{\mathrm{B}}(\boldsymbol{x}'|\boldsymbol{x})$ 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 3) for the details on the retrospective process.

Equation [3] shows that the correction of the bias in $\pi_{\mathrm{B}}(\tau|\boldsymbol{x})$ requires the population growth rate λ, which is easily estimated in the dynamics cytometer experiment. On the other hand, Eq. [4] indicates that the correction of $\mathbb{T}_{\mathrm{B}}(\boldsymbol{x}'|\boldsymbol{x})$ necessitates $u(\boldsymbol{x}')$, 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 the related 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. 2 (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(37, 40), 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 4 and 5).

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 converged distributions of the chronological, retrospective, and tree sampling are summarized in Tab. 1.

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 a couple of genes over lineages, such genes may not be sufficiently relevant for 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 (47) 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. 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 (47). 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 (27, 48). 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 Euclidian 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 cells, respectively. $\boldsymbol{w}$ is a multidimensional Gaussian random variable with a mean vector $\boldsymbol{0}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_w$.

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 $\boldsymbol{\Sigma}_v$. In this model, the estimation problem is reduced to estimating parameters $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{\Sigma}_w$, and $\boldsymbol{\Sigma}_v$, simultaneously. 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 $\boldsymbol{\theta}_{\boldsymbol{x}}^{B}$ 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 $(\mathbb{T}_{\mathrm{F}}, \{\boldsymbol{\theta}_{\boldsymbol{x}}^{B}\})$ 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 (47). 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 6 and 8). For the continuous state-space model, we can employ the well-established estimation technique of the Kalman filter (47). In the M-step, the parameters $\Theta^{(n)} = (\mathbb{T}_{\mathrm{F}}, \{\boldsymbol{\theta}_{\boldsymbol{x}}^{B}\})$ is updated so that $\pi_{\mathrm{B}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{F}}$ are fitted to the following modification of the empirical distributions, respectively (1):

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

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

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 7). It is known that each update always increases the likelihood (47). In the continuous case, we update $\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Sigma}_w$, and $\boldsymbol{\Sigma}_v$ in the same way, that is, update the parameters so that $\pi_{\mathrm{B}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{F}}$ are fitted to $\pi_{\mathrm{emp}}^{\mathrm{BW}}(\cdot|\boldsymbol{x})$ and $\mathbb{T}_{\mathrm{emp}}^{\mathrm{BW}}(\boldsymbol{x}'|\boldsymbol{x})$, respectively.

5. Applications

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. 3 (a) and obtained a synthetic lineage tree as shown in Fig. 3 (b). By applying LEM to the lineage tree in Fig. 3 (b), we could recover the states of the cells from the tree as in Fig. 3 (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 9 and 10). The states of the leaf cells cannot be inferred in Fig. 3 (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. 3 (d). Such information on the states of the leaves may be obtained by conducting single-cell staining or scFISH (49), or scRNA sequencing at the end of the experiment, as assumed in the previous attempts of the state inference from lineage trees (29, 30). The convergence of the log-likelihoods was also checked for both situations (Figs. 3 (e) and (f)). We have further compared the empirical and estimated retrospective distributions of the division times (Figs 3 (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. 3 (a) and another with a different parameter set for 1000 times each to evaluate the accuracy of our estimation. See Supplementary Information (Section 9 and 10). We observed that the estimation is consistent with the true parameter in total by virtue of the correction of the survivor bias. Similarly, we have also tested LEM for the continuous-model to confirm that LEM also works for that situation. See Supplementary Information (Section 11).

B. Deciphering the latent states of *E. coli* cells from lineage trees. We next inferred the latent states of the *E. coli* cells in the lineage trees observed by using the dynamics cytometer in Hashimoto *et al.* (21) (Fig. 1 (d)). The population of *E. coli* (F3 rpsL-gfp strain) was observed every one minute in the M9 minimum medium supplemented with 0.2% glucose at 37°C. We first applied LEM for the discrete model and determined the number of the latent states by Akaike Information Criteria (AIC) (50). The best number of the discrete states was estimated to be 1, which means that the discrete model with no latent state fits the data the best (data not shown). However, this result cannot explain the non-zero correlation ($r = 0.2082$) between the division times of the mother-daughter pair observed in our data set. A potential reason why the discrete model could not capture this correlation and the associated latent states may be because the latent states are not distinct enough to be detected by the discrete-state model, suggesting that the latent state is better represented by the continuous rather than the discrete model.

To validate this hypothesis, we applied LEM of the continuous-state model, in which the dimension k of the state space $\Omega = \mathbb{R}^k$ was again determined by AIC. Then, we found $k = 3$ to be the dimension of the best continuous model. We also obtained the inferred dynamics of the latent state $\boldsymbol{x}$ over the lineage tree as in Fig. 4 and its parameter values as follows:

$$\begin{aligned}\boldsymbol{A} &= \begin{pmatrix} -0.731 & 0.438 & 0.032 \\ -2.51 & 1.124 & 0.062 \\ -0.262 & 0.0068 & 1.007 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \\ \boldsymbol{\Sigma}_w &= \begin{pmatrix} 0.055 & 0 & 0 \\ 0 & 0.038 & 0 \\ 0 & 0 & 0.016 \end{pmatrix}, \quad \Sigma_v = 0.04. \end{aligned} \qquad [13]$$

For the details of the analysis, see Supplementary Information (Section 12). Of the three components of the inferred latent state, the first one has the fastest time-scale of approximately one generation, whereas the third one changes slowly over generations (Fig. 4).

As shown in Fig. 5 (a), the likelihood increases monotonically in terms of k, and $k = 3$ is the dimension above which the likelihood starts saturating, indicating that LEM convergences and the inference is achieved appropriately. In order to validate the significance of $k = 3$, we firstly simulated the continuous-state model without latent state ($k = 0$) for two parameter sets to obtain synthetically lineage tree data, and then applied LEM to infer the dimensionality from the synthetic data. For all 100 independent simulations and the subsequent inferences, we have obtained $k = 0$ as the inferred dimensionality (data not shown), demonstrating that LEM rarely detect a wrong latent state if it does not exist. To check the validity further, we also conducted a bootstrap analysis in which we generated surrogate trees by randomly swapping the division times of the cells in the *E. coli* lineage tree (Fig. 1 (d)) and applied LEM to the surrogates. Because the division times of the cells in the surrogate trees can be approximated to be mutually independent due to the random swapping, the surrogate trees can effectively work as the data from the null hypothesis of no latent state. Of 100 trials, $k = 0$ was inferred in most of cases (Fig. 5 (b)). In the rest of the trials, $k = 1$ and $k = 4$ were obtained. All the trials with $k = 4$ inferred are accompanied by much higher likelihoods than the case of $k = 0$ ((Fig. 5 (a)) and irregularly large variances for the latent state $\boldsymbol{w}$ (Fig. 5 (d)). Such large variances effectively allow the latent state to arbitrarily fit to the observations. Therefore, $k = 4$ is probably due to an inappropriate convergence of the EM algorithm, which has also been reported to occur when it is applied to the MLE of models with latent states (47). In contrast, the results of $k = 1$ show the likelihood and the variance, comparable to those of $k = 0$. This suggests that the model with $k = 0$ sometime generates samples being similar to those from $k = 1$. However, the lack of $k = 2$ indicates that the probability to obtain $k > 1$ from the model of $k = 0$ by chance is much less than 1/100. Lastly, we also applied LEM to another *E. coli* tree and obtained $k = 3$ for this data set (data not shown). Therefore, $k = 3$ inferred from Fig. 1 (d) should have a high statistical significance.

Next, we investigated how the latent state represents the stochastic behavior of the division times. From the assumption of the continuous model, the posterior average of the division time of the cell i in the tree is obtained as

$$\langle \log \tau_i \rangle = \boldsymbol{C}\boldsymbol{x}_i = x_i^1 + x_i^2 + x_i^3. \quad [14]$$

The comparison of $\langle \log \tau_i \rangle$ with the actual observation of the division time $\log \tau_i$ shows that the intercellular variation of the division times is mainly accounted by the fluctuation of the latent state (Fig. 6 (a)), which is also reflected in the small value of the state-independent fluctuation Σ_v (Eq. (13)). The dissection of $\langle \log \tau_i \rangle$ into each component of the latent state also indicates that x^1 and x^2 mainly represent the fluctuation of the division time whereas x^3 encodes its average value (Fig. 6 (a)).

Then, we also analyzed how the latent state conveys the information on the division statistics over generations. By using $\boldsymbol{A}$ and $\boldsymbol{x}$ inferred, we can predict the division time of the daughter cells from the latent states of their mothers as

$$\langle \log \tau_{i+1} \rangle = \boldsymbol{C}\boldsymbol{A}\boldsymbol{x}_i. \quad [15]$$

where we abuse the notation $i + 1$ to mean the label of a daughter cell of the cell i. Similarly, we can predict the division times of the grand daughter cells. As shown in Fig. 6 (c), the latent state effectively captures the relationship of the division times over generation, and thereby, the posterior averages of the division times, $\langle \log \tau_i \rangle$ and $\langle \log \tau_{i+1} \rangle$, also reproduce the correlation between the mother-daughter pairs as $r = 0.2034$ (Fig. 6 (b)).

Finally, we clarify how the inter-generation information is encoded in the latent state and its dynamics by plotting the phase space dynamics of the latent state (Fig. 6 (d)). The latent dynamics had fast and slow components: the fast one is basically the projective dynamics to an one-dimensional sub-manifold in the x^1-x^2 plane(Fig. 6 (e)), whereas the slow one is a dynamics formed in the the sub-manifold and x^3(Fig. 6 (d)). This result demonstrates that x^3 is not only encoding the average value of the division time, but also the information of the division times of its descendants. Moreover, the slow dynamics suggests an existence of a slow regulatory factor underlying the noisy behavior of the division times and being inherited over generations.

6. Summary and Discussion

In this study, we have derived and proposed LEM, a statistical method to infer the latent states of the cells and the associated state-switching and division dynamics from lineage tree data, which combines the correction method of the survivor 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. 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. The inferred states successfully capture the underlying effective inheritance dynamics of the division times over generations even though the correlation of the observed division times between the mother-daughter pairs is subtle presumably because of the stochastic nature of the cellular replication.

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' on that of the mother τ as $\pi_{\mathrm{F}}(\tau'|\tau)$ (51, 52). 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 how the reproductive capabilities of the cells are determined, regulated, and inherited as the consequences of the intracellular networks.

Moreover, 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, 53). 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 indispensable roles to investigate the more complicated processes of the cellular natural selections occurring in the populations of bacteria, pathgens, 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 (34, 54). This generalization may be useful when we include the size of a cell as a state, which naturally correlates between the daughters (25, 55). 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 (36, 56).

On the other hand, 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 survivor bias still 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 ancestral 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.

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, JP16H06155 and JST PRESTO Grant Number JPMJPR15E4, 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. Pugatch R (2015) Greedy scheduling of cellular self-replication leads to optimal doubling times with a log-frechet distribution. *Proceedings of the National Academy of Sciences* 112(8):2611–2616.
17. Wehrens M, Büke F, Nghe P, Tans SJ (2018) Stochasticity in cellular metabolism and growth: Approaches and consequences. *Current Opinion in Systems Biology* 8:131 – 136. • Regulatory and metabolic networks • Special Section: Single cell and noise.
18. Kaneko K, Furusawa C (2018) Macroscopic theory for evolving biological systems akin to thermodynamics. *Annual Review of Biophysics* 47(1):273–290. PMID: 29792817.
19. Crewe P, Gratwick R, Grafen A (2018) Defining fitness in an uncertain world. *Journal of Mathematical Biology* 76(5):1059–1099.
20. Lehmann L, Mullon C, Akçay E, Cleve J (2016) Invasion fitness, inclusive fitness, and reproductive numbers in heterogeneous populations. *Evolution* 70(8):1689–1702.
21. 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.
22. 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.
23. Wang P, et al. (2010) Robust growth of escherichia coli. *Current Biology* 20(12):1099 – 1103.
24. Norman TM, Lord ND, Paulsson J, Losick R (2013) Memory and modularity in cell-fate decision making. *Nature* 503:481 EP –. Article.
25. Taheri-Araghi S, et al. (2015) Cell-size control and homeostasis in bacteria. *Current Biology* 25(3):385 – 391.
26. Tanouchi Y, et al. (2015) A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature* 523:357 EP –.
27. Wakamoto Y, Grosberg AY, Kussell E (2011) Optimal lineage principle for age-structured populations. *Evolution* 66(1):115–134.
28. Lambert G, Kussell E (2015) Quantifying selective pressures driving bacterial evolution using lineage analysis. *Phys Rev X* 5(1):011016. 26213639[pmid].
29. Hormoz S, Desprat N, Shraiman BI (2015) Inferring epigenetic dynamics from kin correlations. *Proceedings of the National Academy of Sciences* 112(18):E2281–E2289.
30. 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.
31. Hicks DG, Speed TP, Yassin M, Russell SM (2018) Statistical inference in cell lineage trees. *bioRxiv*.
32. Hicks DG, Speed TP, Yassin M, Russell SM (2018) Maps of variability in cell lineage trees. *bioRxiv*.
33. Olariu V, et al. (2009) Modified variational bayes em estimation of hidden markov tree model of cell lineages. *Bioinformatics* 25(21):2824–2830.
34. 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.
35. 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.
36. Kuchen EE, Becker N, Claudino N, Hofer T (2018) Long-range memory of growth and cycle progression correlates cell cycles in lineage trees. *bioRxiv*.
37. 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.
38. Thomas P (2017) Making sense of snapshot data: ergodic principle for clonal cell populations. *Journal of The Royal Society Interface* 14(136).
39. 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.
40. Marguet A (2016) Uniform sampling in a structured branching population. *ArXiv e-prints.*
41. Marguet A (2017) A law of large numbers for branching Markov processes by the ergodicity of ancestral lineages. *ArXiv e-prints.*
42. Thomas P (2018) Population growth affects intrinsic and extrinsic noise in gene expression. *bioRxiv*.
43. Thomas P (2018) Analysis of cell size homeostasis at the single-cell and population level. *Frontiers in Physics* 6:64.
44. Harris TE (1963) *The Theory of Branching Processes.* (Springer-Verlag Berlin Heidelberg).
45. Sughiyama Y, Nakashima S, Kobayashi TJ (2018) Fitness response relation of a multi-type age-structured population dynamics. *ArXiv e-prints.*
46. Sughiyama Y, Kobayashi TJ, Tsumura K, Aihara K (2015) Pathwise thermodynamic structure in population dynamics. *Phys. Rev. E* 91(3):032120.
47. Christopher B (2006) *Pattern Recognition and Machine Learning.* (Springer-Verlag New York).
48. Rubinow SI (1968) A maturity-time representation for cell populations. *Biophys J* 8(10):1055–1073. 5679389[pmid].
49. Frieda KL, et al. (2016) Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541:107 EP –.
50. 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.
51. Lin E, et al. (2017) High-throughput microfluidic labyrinth for the label-free isolation of circulating tumor cells. *Cell Systems* 5(3):295–304.e4.
52. Lin J, Amir A (2018) Population growth with correlated generation times at the single-cell level. *ArXiv e-prints.*
53. Balaban N (2011) Persistence: mechanisms for triggering and enhancing phenotypic variability. *Current Opinion in Genetics & Development* 21(6):768 – 775. Genetics of system biology.
54. Paskin MA (2002) Thin junction tree filters for simultaneous localization and mapping, (EECS Department, University of California, Berkeley), Technical Report UCB/CSD-02-1198.
55. Susman L, et al. (2018) Individuality and slow dynamics in bacterial growth homeostasis. *Proceedings of the National Academy of Sciences* 115(25):E5679–E5687.
56. Särkkä S (2013) *Bayesian filtering and smoothing.* (Cambridge University Press) Vol. 3.

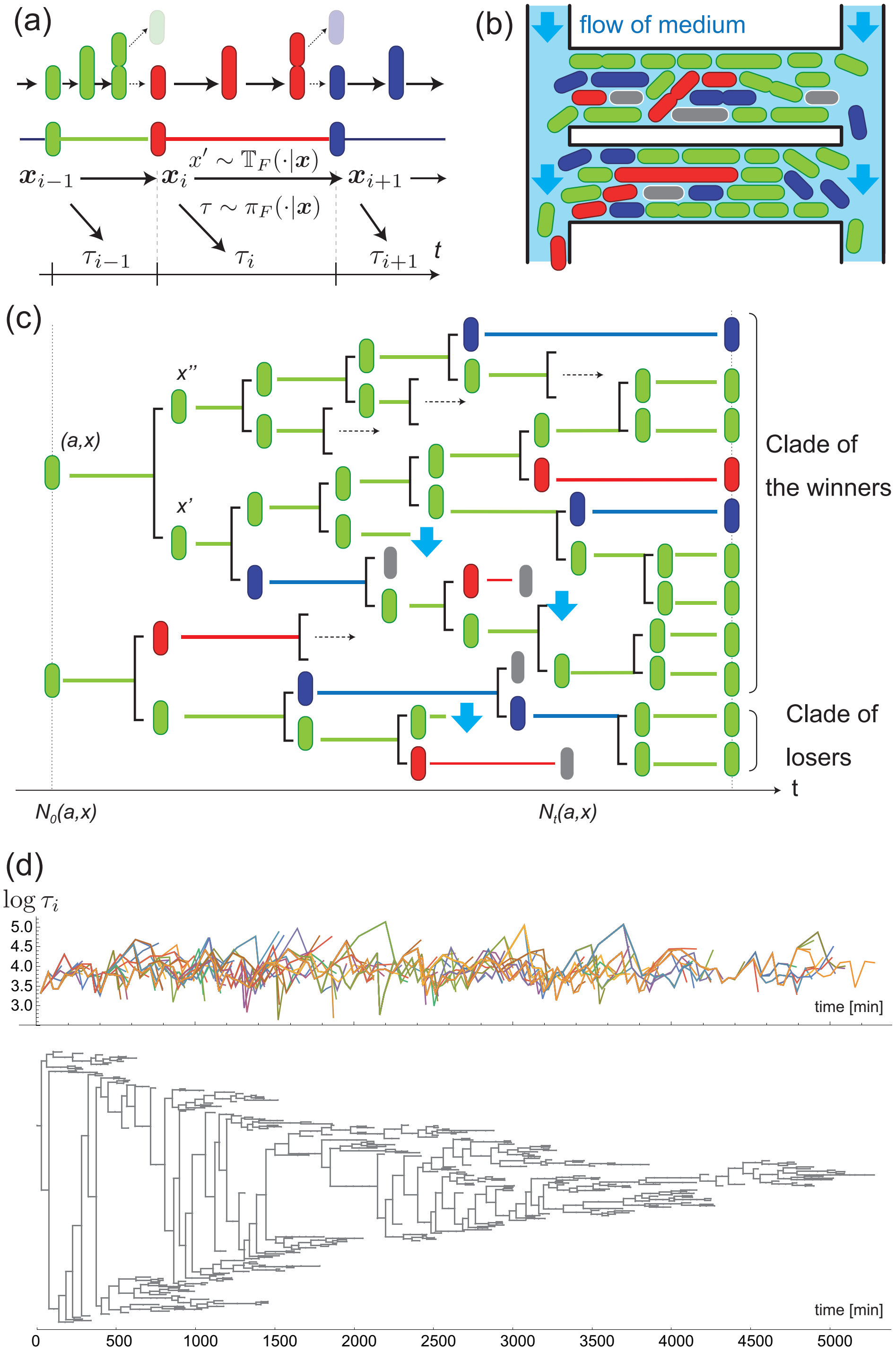

Fig. 1. (a) A schematic diagram of the stochastic state-switching and fluctuating division time of a cell traced by ignoring its sister cells. $\boldsymbol{x}_i$ represents the state of the cell i and τ_i is its division time. (b) The outline of the dynamics cytometer (21). (c) A schematic representation of a lineage tree obtained by the dynamics cytometer, and an illustration of the survivor bias. The dynamics of individual cells follows the state-switching and the division time statistics described in (a). The lineage tree is composed of two kinds of information: parent-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. (d) A lineage tree of *E. coli* (F3 rpsL-gfp strain) cells grown with M9 minimum medium supplemented with 0.2% glucose at 37°C. (lower panel) and a time series plot of the division times (upper panel). The data is adopted from (21) and replotted.

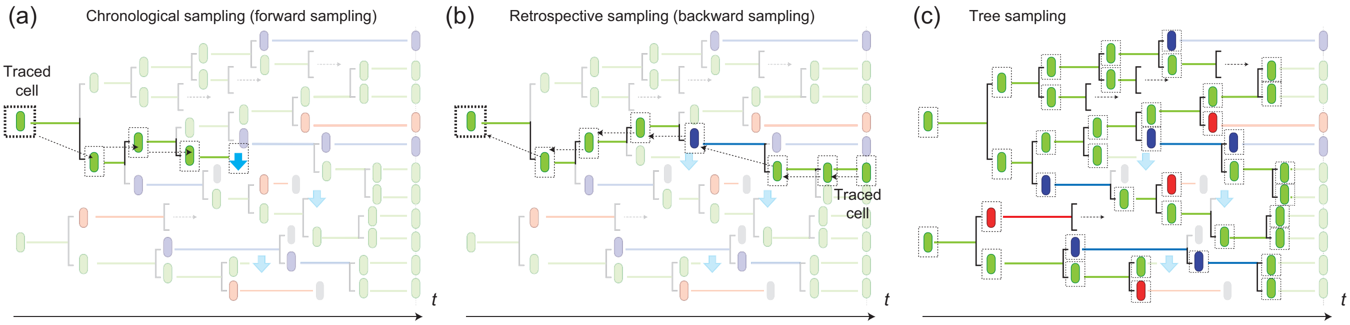
(a)
Chronological sampling (forward sampling)
Traced cell
t
(b)
Retrospective sampling (backward sampling)
Traced cell
t
(c)
Tree sampling
t

Fig. 2. 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 a 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 cell. 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.

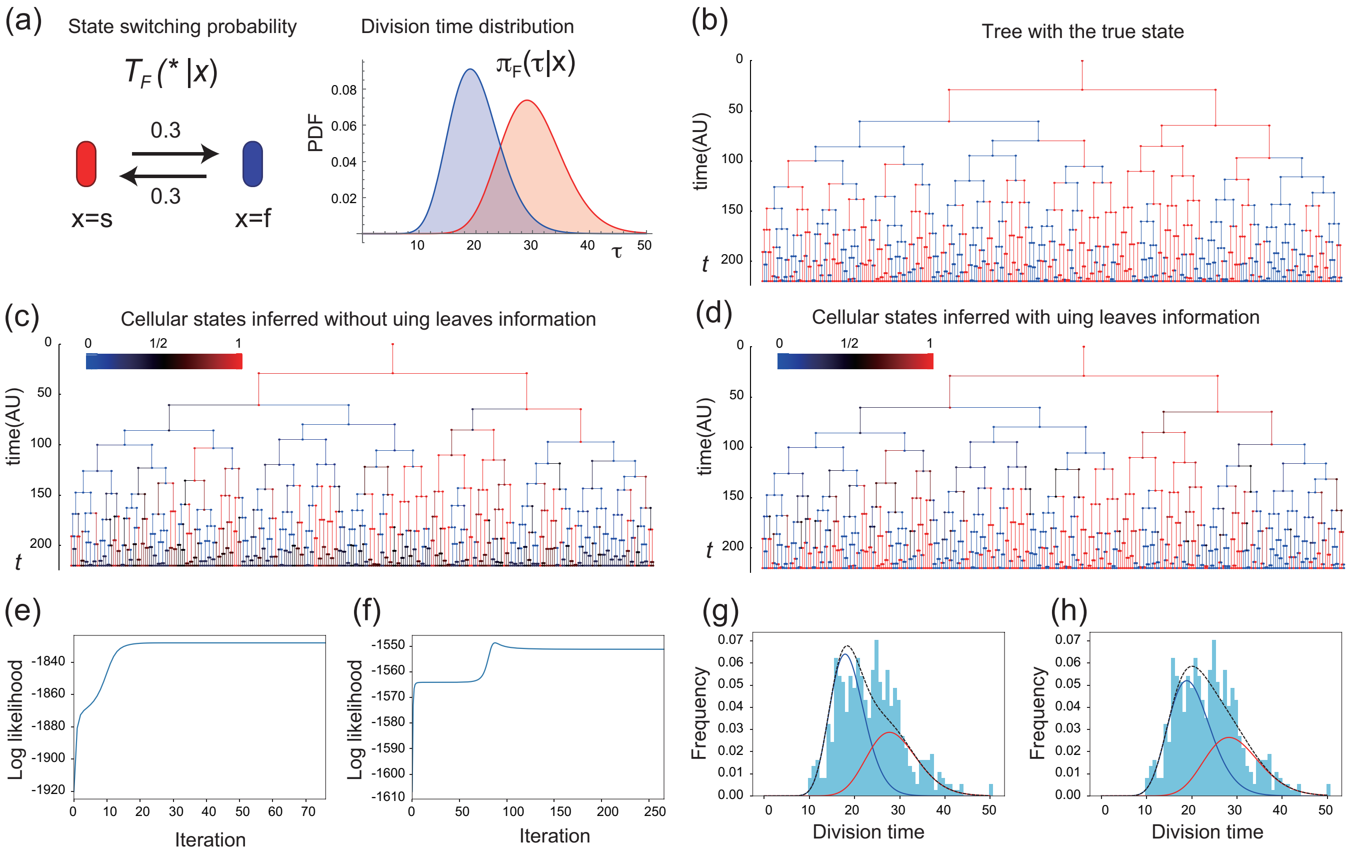

Fig. 3. A performance evaluation of LEM by comparing a simulated lineage tree of the discrete model and the corresponding inferred states of the cells with and without using the information of the states of the leaves. (a) A schematic diagram of the state-switching dynamics and their division time distributions 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 linage 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 .

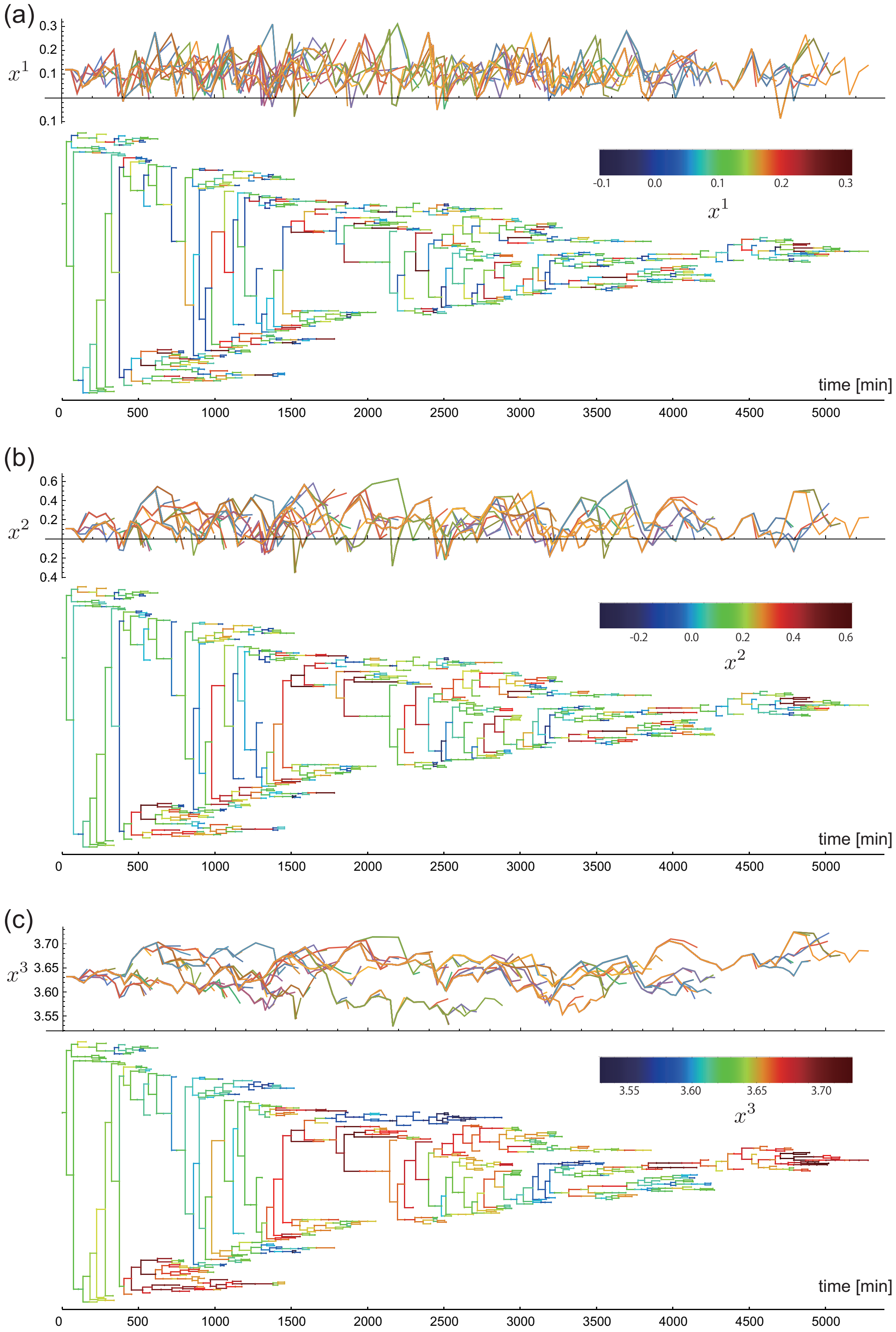

Fig. 4. The dynamics of the first (a), the second (b), and the third (c) components of the inferred three-dimensional state depicted as time series(upper panel) and overlaid on the lineage tree (lower panel). The color codes over the trees represent the actual values of the components.

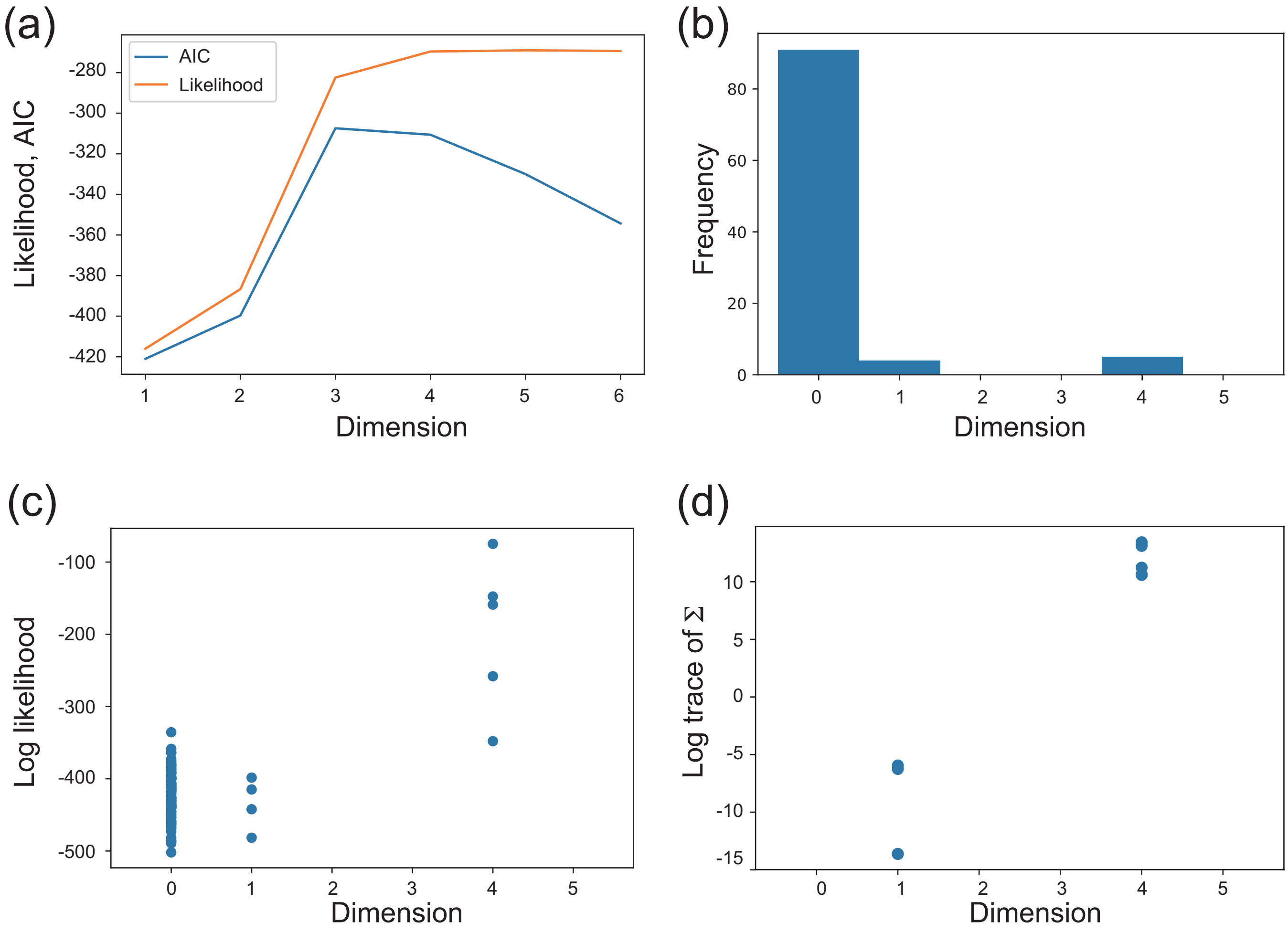

Fig. 5. (a) The log-likelihood and AIC as functions of the dimension k. (b) Histogram of AICs obtained by the bootstrap analysis using 100 independently surrogated trees. (c, d) A scatter plots of the likelihoods (c) and the variances of the latent state $\mathbf{\Sigma}_w$ (d) from the surrogated trees. Each point correspond to the value obtained from each surrogated tree.

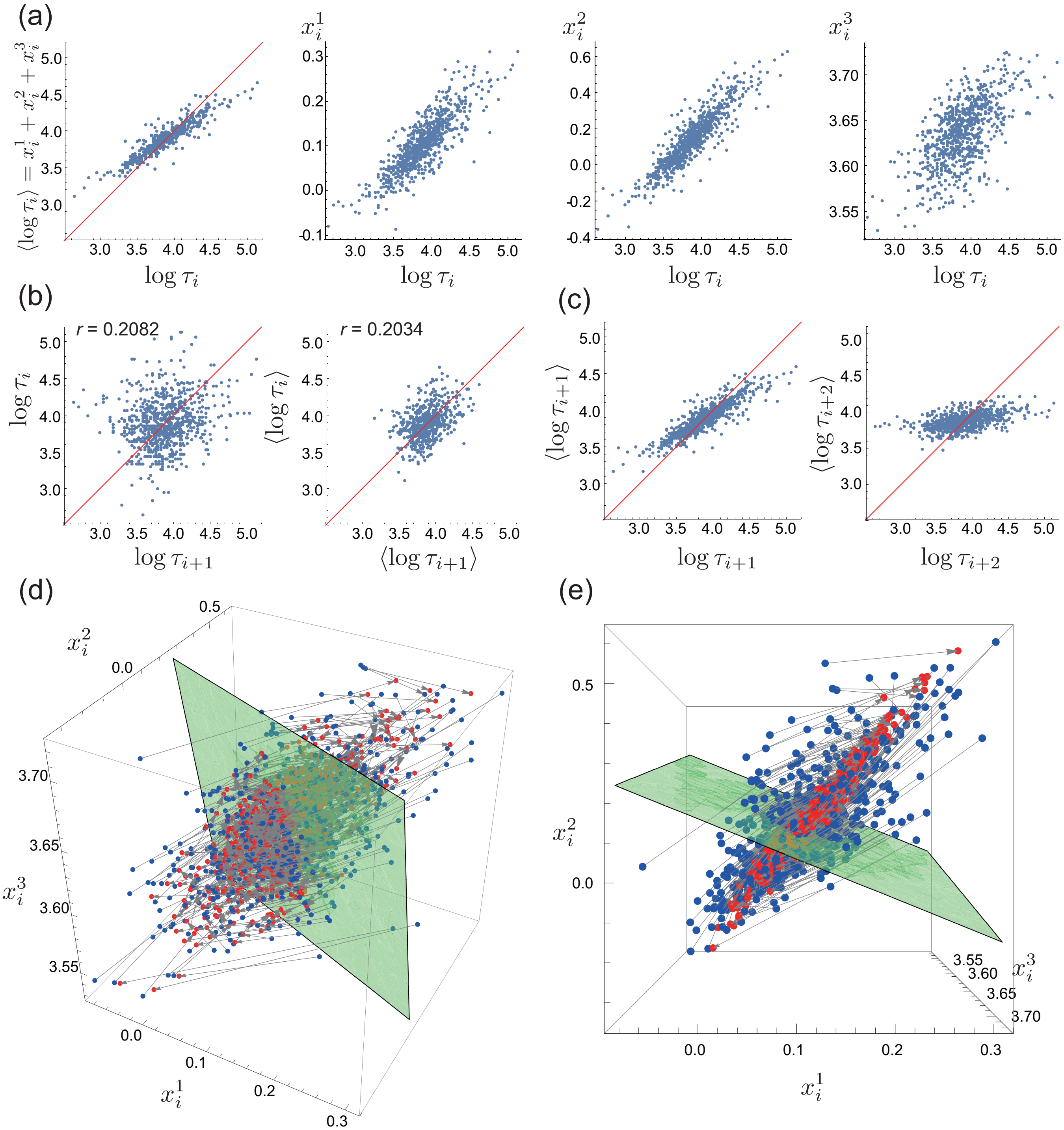

Fig. 6. (a) Comparisons of the actual division times, $\log \tau_i$, with the predicted division times, $\langle \log \tau_i \rangle$, and with the three components of $\boldsymbol{x}_i$. Each point corresponds to each cell in the tree. (b) Comparisons of the mother's and its daughter's division times by using the actual observations (left) and the predicted values (right). Each point corresponds to each mother-daughter pair. (c) Comparisons of the actual division times of the daughter (left, $\log \tau_{i+1}$) and grand-daughter (right, $\log \tau_{i+2}$) cells with their predicted values $\langle \log \tau_{i+1} \rangle$ and $\langle \log \tau_{i+2} \rangle$ obtained from the latent states of the corresponding mother cells $\boldsymbol{x}_i$. (d) State-space representation of the dynamics of the latent state. Each blue point represents the latent state of each cell $\boldsymbol{x}_i$, and the corresponding red point connected by the gray arrow is its mapped state $A\boldsymbol{x}_i$. The green plane is an instance of the surface satisfying $x^1 + x^2 + x^3 = \mathrm{const.}$. A subset of cells that are on the same $x^1 + x^2 + x^3 = \mathrm{const.}$ plane generates the same predicted value of the division time. The green plane in the plot is obtained for $\mathrm{const.} = \tau_{\mathrm{av}}$ where τ_{av} is the sample average of the division times. (e) The same 3D plot as (d) but rendered from the top-view.