Supplementary Note 1: CELLMA: Cell-state transition Estimation by Lineage Leaf-state Markov Analysis

1 Introduction of the model and data

In this Supplementary Note, we formulate a model that describes the behavior of a pool of cells that can transition between discrete, mutually exclusive "states" and divide, generating two daughter cells. Specifically, we consider the scenario where the relative abundances of cells in different states do not vary significantly over time, while cells continuously undergo transitions (i.e., the cell pool is at steady-state).

The model we present is thought in such a way to allow the estimation of the rates of cell state transitions from incomplete recordings. In particular, we are interested in the situation where one can measure the time-resolved division history of hundreds of cells in a pool but not their evolving state. If the time-resolved information on the cell state were available, the estimation of the cell-state transition rates would be direct and rather trivial. Instead, the limitation to computationally circumvent is that cell state cannot be probed for its state "live" and can only be recorded at an endpoint after measuring the division history.

Pragmatically, measurements of this sort could be the result of slightly different experimental set-ups. In this work, we consider a case where live imaging microscopy has been used to record the division history and stainings to assess the endpoint state of the cells. Live imaging recordings, appropriately processed by cell segmentation and tracking, provide both kinship relationships between cells and the time of division events, a variable important for the modeling framework presented here. The information of the cell staining can be summarized in a modest number of discretized cell states, among which we want to estimate the rate of transition. For example, in the context of the present paper, we have six lipotypes: ChTxB+, ShTxB1+, ShTxB2+, ShTxB1+ShTxB2+ (double), ShTxB1+ShTxB2+ChTxB+ (triple), Other (e.g. configurations that cannot be uniquely resolved by toxin stainings alone).

Furthermore, in the biological context considered in the present study, we have no ways of measuring "live" casual factors or covariates correlating with the transition choice of an individual cell. Thus, for the purpose of our model, the state transition occurring in a particular cell can be modeled as a stochastic process, yet with fixed rates. We also note that there is no biological reason to assume that the past state of a cell should affect the likelihood of the cell to transition from its current to a new, future state. Also, we have no biological evidence suggesting the system might not be ergodic.

These considerations naturally lead to the choice of a Markov chain to model the transition process. Finally, we assume that cell division per se does not induce an immediate state transition. In biological terms, this assumption means that we do not allow for asymmetric segregation of the lipids during cell division. However, we note that if this were not true, our fit should accommodate that occurrence by skewing the transition rates to recapitulate a similar behavior.

In summary, the model we propose considers a discrete-time Markov chain evolution of the cell state in each cell of the pool. Transitions are not affected by cell divisions. The model postulates that two daughter cells will inherit the same cell state from their mother at the moment of division. Only then will their states start drifting apart, as they will be more likely to transition to other states according to the same transition probability.

In the following section, we present a more careful mathematical formulation of the model and derive a maximum likelihood estimation of the transition matrix between cell states. We anticipate that it will formally show that one can use the knowledge of the time at which sister-cell drift started together with the recording of the cell's final state to estimate the transition matrix of the Markov chain.

2 Formulation

Let us indicate with $\mathbf{c}_i \in [0,1]^n \ : \ \mathbf{c}_i^T \cdot \mathbb{1} = 1$ the vector containing the probabilities that cell i is in each of n possible states. Upon measurement we become sure that a cell is in state s and therefore $\mathbf{c}_i = \mathbf{e}_s$, where the notation $\mathbf{e}_s$ indicates a unit vector with its s-th element equal to one and the other entries equal zero. We also define for convenience $\mathcal{S} = \{\mathbf{e}_s \ \forall s \in \mathbb{N} : 1 \leq s \leq n\}$.

We consider a discrete-time setting and postulate that for a single time step, the state of a cell changes following a memoryless Markov process, defined by the transition matrix $\mathbf{\Theta} \in [0,1]^{n \times n}$, where each entry Θ_{ij} is the probability of transition from state j to state i. Thus, the updated probability vector after a series of k time steps can be simply obtained by left-multiplying the vector with the matrix. For example, the updated probability vector of a cell i after $k \in \mathbb{N}$ time steps can be written as:

$$\mathbf{c}_{i|t+k} = \mathbf{\Theta}^k \cdot \mathbf{c}_{i|t} \tag{1}$$

Where $\mathbf{\Theta}^k$ indicates the matrix power and the notation $_{|t}$ is used to specify that the state that a particular variable is considered at a particular time point t. We note that $\mathbf{\Theta}$ is a stochastic matrix, that is $\mathbf{\Theta}^T \cdot \mathbb{1} = \mathbb{1}$.

From the above, it follows that the probability of a cell to be in state s starting from a state probability vector $\mathbf{c}_i$ can be written as:

$$P(\mathbf{c}_{i|t+k} = \mathbf{e}_s \mid \mathbf{c}_{i|t}, \mathbf{\Theta}) = \mathbf{e}_s^T \cdot \mathbf{\Theta}^k \cdot \mathbf{c}_{i|t} \tag{2}$$

Let us now consider the notation for data. For each cell at the end of our division tracking experiment, we readout a cell state. Consistently with the formalism defined above, we represent this state using a unit vector. The cell state measured for cell i will be, thus, indicated as $\mathbf{m}_i \in \mathcal{S}$. Conversely, the state of each parent cell, which we cannot measure, has to be considered as a latent

variable. For an easier discrimination between measurable and latent cell states, we indicate those two sets of variables with $\mathbf{c}$ and γ respectively. Furthermore, we indicate the sister relationship between two cells i and j writing $j = \alpha(i)$ and the mother-daughter relation between cell k and i writing $k = \mu(i)$ and conversely $i = \mu^{-1}(k)$. Finally, we indicate a set of relatives as a "clade" and we distinguish the set of measured states in the rth clade with $\mathcal{K}_r$ and the set of latent (e.g. parent) states in that clade with $\mathcal{Q}_r$. Considering a cell i we then have $\mathcal{K}_r = \kappa(i)$ and $\mathcal{Q}_r = \rho(i)$, so that we can write expressions such as $\alpha(\mu(i)) \in \mathcal{Q}_r$.

From now on we will omit the notation $_{|t}$ and, instead, indicate with $\mathbf{c}_i$ the state of the i-th cell at the time of the final measurement and with $\boldsymbol{\gamma}_j$ the state of the j-th latent cell at the time when it divides generating the two daughter cells. We note that all the equations below can be written as a function of those variables only and of the number of time-separating pairs of related cells i and j that we indicate with $\chi_{ji} \in \mathbb{N}$.

The aim of CELLMA is to produce a Maximum Likelihood estimate $\hat{\boldsymbol{\Theta}}$ of the Markov transition matrix of the state transition process from the set of data $\mathcal{M} = \{\mathbf{m}_i\}$. In order to achieve this, we need to consider the following facts and assumptions:

Firstly, sister cell states are not independent because at time of their generation have inherited the same cell type from their mother cell.

$$P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q) \neq P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z) \cdot P_{\boldsymbol{\Theta}}(\mathbf{c}_j = \mathbf{e}_q) \quad \forall (i,j) : j = \alpha(i) \tag{3}$$

Instead, sister cells are conditionally independent given the mother cell state at the time of division and the Markov transition matrix, as the two cells in a identical state after division transition to new state independently following a Markov process.

$$P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q | \boldsymbol{\gamma}_b) = P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z | \boldsymbol{\gamma}_b) \cdot P_{\boldsymbol{\Theta}}(\mathbf{c}_j = \mathbf{e}_q | \boldsymbol{\gamma}_b) \ \ \forall (i,j,b) : j = \alpha(i) \ \& \ b = \mu(i) \tag{4}$$

The state of cells for which we do not observe any evidence of kinship during the time-lapse imaging and are, thus, from different "clades" are considered independent beside having to obey the same probabilistic process determined by the Markov matrix.

$$P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q) = P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z) \cdot P_{\boldsymbol{\Theta}}(\mathbf{c}_j = \mathbf{e}_q) \quad \forall (i,j) : j \notin \kappa(i) \tag{5}$$

We have now all the ingredients to simplify the expression of the Maximum likelihood problem to estimate $\boldsymbol{\Theta}$.

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{maximize}} \ \ & \mathcal{L}(\boldsymbol{\Theta} | \{\mathbf{m}_i\}) \\ \text{s.t.} \ \ & \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1} \\ & \boldsymbol{\Theta} \succcurlyeq 0 \end{aligned} \tag{6}$$

With $\mathcal{L}(\boldsymbol{\Theta} | \{\mathbf{m}_i\}) = P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\})$. First we consider the assumption expressed equation (5) to factor the likelihood function by clades as follows:

$$P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\}) = \prod_{r=1}^{N_K} P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, \forall i \in \mathcal{K}_r\}) \tag{7}$$

where N_K is the number of clades.

Now we note that we can compute the likelihood for a clade considering the joint probability of the data and the latent variables of a clade, and by marginalizing over all the possible combination of states. More specifically we have:

$$\prod_{r=1}^{N_K} P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, \forall i \in \mathcal{K}_r\}) = \sum_{l=1}^{n} \sum_{k=1}^{n} ... \sum_{q=1}^{n} P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, \forall i \in \mathcal{K}_r\}, \boldsymbol{\gamma}_a = \mathbf{e}_l, \boldsymbol{\gamma}_b = \mathbf{e}_k, ... \boldsymbol{\gamma}_q = \mathbf{e}_q) \tag{8}$$

Where only the latent variables in that clade are considered e.g. $\{a, b, ..., z\} = \rho(i)$. We can write the formula (8) more compactly considering two compounded variables, one indicating the possible combination of latent states $\mathbf{E} = (\mathbf{e}_l, \mathbf{e}_k, ..., \mathbf{e}_q) \in \mathcal{S}^{|\rho(i)|}$ that will be indexed for convenience as follows $\mathbf{E} = (\boldsymbol{\varepsilon}_a, \boldsymbol{\varepsilon}_b, ... \boldsymbol{\varepsilon}_z) : \{a, b, ..., z\} = \rho(i)$ and the other corresponding to all stacked latent variables of a clade $\mathbf{\Gamma} = (\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b, ... \boldsymbol{\gamma}_z) : \{a, b, ..., z\} = \rho(i)$. Note that $\mathcal{S}^{|\rho(i)|}$ indicates a Cartesian power and $|\cdot|$ indicates the carnality of a set.

$$P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, \forall i : \, i \in \mathcal{K}_r\}) = \sum_{\mathbf{E} \in \mathcal{S}^{|\rho(i)|}} P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, \forall i : \, i \in \mathcal{K}_r\}, \mathbf{\Gamma} = \mathbf{E}) \tag{9}$$

We note that relying on the conditional independence of daughter cell state given the mother (e.g. equation (4)), we can factorize the term inside the sum as the product of probabilities for each cell, summed over all possible states $\mathcal{S}$ for the parent cells, which are latent variables.

$$P_{\mathbf{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \, i \in \mathcal{K}_r\}) = \sum_{\mathbf{E} \in \mathcal{S}^{|\rho(i)|}} \prod_{\forall i \in \mathcal{K}_r} P_{\mathbf{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \; \mathbf{\Gamma} = \mathbf{E}) \tag{10}$$

Despite the appearance, the computation of equation (10) is less daunting than it may seem at first sight as the expression also factorizes in the "mother-daughter" terms shown below (either latent-measured or latent-latent). Those terms have an analogous form of the expression in equation (2) and allow us to explicit the function of the only unknown quantity $\mathbf{\Theta}$, resulting in a functional form that is the product of different terms with .

$$\begin{aligned}
&\sum_{\mathbf{E} \in \mathcal{S}^{|\rho(i)|}} \prod_{i \in \mathcal{K}_r} P_{\mathbf{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \; \mathbf{\Gamma} = \mathbf{E}) = \\
&= \sum_{\mathbf{E} \in \mathcal{S}^{|\rho(i)|}} \prod_{i \in \mathcal{K}_r} P_{\mathbf{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \; \boldsymbol{\gamma}_{\mu(i)} = \boldsymbol{\varepsilon}_{\mu(i)}) \prod_{u \in \mathcal{Q}_r} P_{\mathbf{\Theta}}(\boldsymbol{\gamma}_u = \boldsymbol{\varepsilon}_u, \; \boldsymbol{\gamma}_{\mu(u)} = \boldsymbol{\varepsilon}_{\mu(u)}) = \\
&= \sum_{\mathbf{E} \in \mathcal{S}^{|\rho(i)|}} \prod_{i \in \mathcal{K}_r} \mathbf{m}_i^T \cdot \mathbf{\Theta}^{\chi_{i,\mu(i)}} \cdot \boldsymbol{\varepsilon}_{\mu(i)} \prod_{u \in \mathcal{Q}_r} \boldsymbol{\varepsilon}_u^T \cdot \mathbf{\Theta}^{\chi_{u,\mu(u)}} \cdot \boldsymbol{\varepsilon}_{\mu(u)}
\end{aligned} \tag{11}$$

Where we have indicated with $\mathbf{E}_{\mu(i)}$ the unitary vector $\mathbf{e}_s$ extracted at the $\mu(i)$-th row of E.

We can finally combine everything we have learned in a single expression of the likelihood function:

$$\mathcal{L}(\boldsymbol{\Theta}|\{\mathbf{m}_i\}) \;=\; \prod_{r=1}^{N_K} \sum_{\mathbf{E}\in\mathcal{S}^{|\rho(i)|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i,\; \mathcal{K}_{\mathbf{r}_r} = \mathbf{E}) \tag{12}$$

Explicating all the terms as a function of $\boldsymbol{\Theta}$ and the data, and listing all the constraints, we arrive to the final form of the CELLMA optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{maximize}} & \prod_{r=1}^{N_K} \sum_{\mathbf{E}\in\mathcal{S}^{|\rho(i)|}} \prod_{\forall i:\, i\in\mathcal{K}_r} \mathbf{m}_i^T \cdot \boldsymbol{\Theta}^{\chi_{i,\mu(i)}} \cdot \boldsymbol{\varepsilon}_{\mu(i)} \prod_{\forall u:\, u\in\mathcal{Q}_r} \boldsymbol{\varepsilon}_u^T \cdot \boldsymbol{\Theta}^{\chi_{u,\mu(u)}} \cdot \boldsymbol{\varepsilon}_{\mu(u)} \\ \text{s.t.} \quad & \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1} \\ & \boldsymbol{\Theta} \succcurlyeq 0 \end{aligned} \tag{13}$$

3 Comments on the implementation

In practice, at the implementation level we actually minimize the following negative log-likelihood function:

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{minimize}} \quad & -\log \mathcal{L}(\boldsymbol{\Theta}|\{\mathbf{m}_i\}) \\ \underset{\boldsymbol{\Theta}}{\text{minimize}} \quad & -\sum_{r=1}^{N_K} \log \sum_{\mathbf{E}\in\mathcal{S}^{|\rho(i)|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i,\; \boldsymbol{\Gamma}_r = \mathbf{E}) \end{aligned} \tag{14}$$

For numerical stability we use log-probabilities whenever possible to avoid numerical cancellation problems. In particular we solve the problem in this form:

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{minimize}} \quad & -\sum_{r=1}^{N_K} \log \sum_{\mathbf{E}\in\mathcal{S}^{|\rho(i)|}} \exp\Bigg(\sum_{i\in\mathcal{K}_r} \log(\mathbf{m}_i^T \cdot \boldsymbol{\Theta}^{\chi_{i,\mu(i)}} \cdot \boldsymbol{\varepsilon}_{\mu(i)}) + \\ & \qquad\qquad \sum_{u\in\mathcal{Q}_r} log(\boldsymbol{\varepsilon}_u^T \cdot \boldsymbol{\Theta}^{\chi_{u,\mu(u)}} \cdot \boldsymbol{\varepsilon}_{\mu(u)}) \Bigg) \\ \text{s.t.} \quad & \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1} \\ & \boldsymbol{\Theta} \succcurlyeq 0 \end{aligned} \tag{15}$$

We solve the above problem using a standard interior point solver. We note that the gradient of the function above is not tractable analytically, nonetheless we can compute it empirically at each iteration of the solver. Clearly, this adds to the computational cost quadratically with respect to the number of states. More influential on time complexity is, however, is the Cartesian product that brings in a factorial contribution. However, we notice that more distant ancestors are progressively less crucial to the estimation, and therefore we can partition big clades by dropping all lineage relation deeper than three layers without losing much information. Overall, for several hundreds of cells and a small number of states (e.g. ≤ 10) the method converges in just few minutes.