PAPER • OPEN ACCESS

Coupled differentiation and division of embryonic stem cells inferred from clonal snapshots

View the article online for updates and enhancements.

Physical Biology

PAPER

Coupled differentiation and division of embryonic stem cells inferred from clonal snapshots

Liam J Ruske[1] , **Jochen Kursawe**[2] , **Anestis Tsakiridis**[3,6] , **Valerie Wilson**[4] ,
Alexander G Fletcher[5,6] , **Richard A Blythe**[7] and **Linus J Schumacher**[4,*]

1. Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Parks Road, Oxford, OX1 3PU, United Kingdom
2. School of Mathematics and Statistics, University of St Andrews, North Haugh, St Andrews KY16 9SS, United Kingdom
3. Centre for Stem Cell Biology, Department of Biomedical Science, University of Sheffield, Sheffield S10 2TN, United Kingdom
4. Centre for Regenerative Medicine, University of Edinburgh, Edinburgh BioQuarter, 5 Little France Drive, Edinburgh, EH164UU, United Kingdom
5. School of Mathematics and Statistics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield, S3 7RH, United Kingdom
6. Bateson Centre, University of Sheffield, Sheffield S10 2TN, United Kingdom
7. SUPA, School of Physics and Astronomy, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom
* Author to whom any correspondence should be addressed.

E-mail: Liam.Ruske@physics.ox.ac.uk, jochen.kursawe@st-andrews.ac.uk, a.tsakiridis@sheffield.ac.uk, V.Wilson@ed.ac.uk, a.g.fletcher@sheffield.ac.uk, R.A.Blythe@ed.ac.uk and Linus.Schumacher@ed.ac.uk

Abstract

The deluge of single-cell data obtained by sequencing, imaging and epigenetic markers has led to an increasingly detailed description of cell state. However, it remains challenging to identify how cells transition between different states, in part because data are typically limited to snapshots in time. A prerequisite for inferring cell state transitions from such snapshots is to distinguish whether transitions are coupled to cell divisions. To address this, we present two minimal branching process models of cell division and differentiation in a well-mixed population. These models describe dynamics where differentiation and division are coupled or uncoupled. For each model, we derive analytic expressions for each subpopulation's mean and variance and for the likelihood, allowing exact Bayesian parameter inference and model selection in the idealised case of fully observed trajectories of differentiation and division events. In the case of snapshots, we present a sample path algorithm and use this to predict optimal temporal spacing of measurements for experimental design. We then apply this methodology to an *in vitro* dataset assaying the clonal growth of epiblast stem cells in culture conditions promoting self-renewal or differentiation. Here, the larger number of cell states necessitates approximate Bayesian computation. For both culture conditions, our inference supports the model where cell state transitions are coupled to division. For culture conditions promoting differentiation, our analysis indicates a possible shift in dynamics, with these processes becoming more coupled over time.

1. Introduction

Changes in gene expression underlie many aspects of cellular behaviour in tissue development, home-ostasis, and regeneration. The concept of discrete cell states is intended to capture the distinct patterns of gene expression that are observed within tissues over time. The deluge of single cell data is leading to an increasingly detailed description of cell state [1, 2]. There are various ways to interrogate a cell's state in different contexts *in vivo* and *in vitro*. Modern technologies such as scRNAseq produce vast amounts of data but are costly, laborious to analyse, and relatively noisy. Older techniques, such as immunofluo-rescent stainings, where cell state can be defined by the co-expression of a small number of genes and/or proteins, are cheaper and simpler and thus still remain heavily used.

While there are numerous techniques for classifying cell states based on single cell data, and ordering them along pseudotime trajectories [3, 4], such classification only offers very limited insight into the

dynamics of cell state transitions [4, 5]. The ability to quantify the dynamics of cell state transitions promises greater insights into cell population heterogeneity and how differentiation varies with culture conditions, and could help steer targeted differentiation of stem cells *in vitro*. Currently, the data remain limited to snapshots in most cases, with only one or few time-point measurements available, and a loss of cell identity across measurements due to their destructive nature. Fluorescent reporters of gene expression for live-imaging circumvent this, but their availability is relatively limited and they allow observation of at most a few genes.

Correct tissue development and regeneration requires balanced cell proliferation and differentiation at the tissue level to ensure the right cell type in the right place at the right amount. However, which dynamics at the cellular level give rise to this population behaviour is in many cases still unclear. In particular, deciding how or whether cell state transitions are coupled to, or occur independently of, cell division is a pre-requisite for quantifying cell state transition rates. Determining evidence for such coupling from snapshot data is an incompletely solved problem [6].

If we only have access to snapshots of clonal colony growth, how can one distinguish whether cell division and cell state transitions are coupled? The answer to this question will affect how well one can infer cell state transition rates, as assuming a 'wrong' underlying model will impair the quality of the inference. Conversely, can we inform experimental design to maximise the information gained from experimental data? To address this, we combine biophysical theory of stochastic population dynamics with Bayesian parameter inference and model comparison. As an example we choose the well-established system of epiblast stem cells (EpiSCs) [7]. EpiSCs are self-renewing, pluripotent cell lines that are an *in vitro* equivalent of the embryonic epiblast tissue. *In vivo*, these cells begin to differentiate (undergo lineage commitment) during gastrulation.

We start with minimal branching process models of cell division and cell state transition (e.g. reversible differentiation), and derive the population dynamics and exact likelihoods for inference based on complete trajectory data (observing every cell division and state transition). We then present a sample path algorithm for inference based on snapshot data. We verify our methods using simulated data and show how the information gained from experimental data can be improved by adaptively spacing the snapshots in time. Next, we extend our models and inference pipeline to EpiSC clonal assay data. In this case, the sample path algorithm becomes too computationally intensive, due to the larger cell state space. Using an approximate Bayesian computation sequential Monte-Carlo scheme, we find that inference favours coupling of cell

division and transitions. We compare inferred transition rates in two experimental conditions, promoting self-renewal and differentiation, and find a bias towards increased expression of two transcription factors (T(Bra) and FoxA2) and decreased expression of another (Sox2) under conditions promoting differentiation, in line with published results of a previous experimental study that cultured EpiSCs *in vitro* [7].

2. Methods

We begin this section by introducing two continuous-time Markov processes as minimal stochastic models for stem cell dynamics. These simple models are then generalized to describe the dynamics of epiblast stem cells (EpiSC) in more detail. We will close this section with a brief summary of Bayesian parameter inference and model selection.

2.1. Minimal models

We begin by defining minimal models of cell division and differentiation in a well-mixed population. We assume that the population dynamics can be adequately described by a continuous-time Markovian stochastic process where state transition rates depend only on the current state of the system. This in turn is specified by a set of population numbers $\{\mathbf{N}\}$ of different cell states. In this section we will consider systems with two different cell states and assume that cells within the population divide and change state independently of one another.

We consider two distinct models of population dynamics: one in which cell differentiation is either coupled to cell division (model C) one in which processes are uncoupled (model U). In both models there are two cell states, A and B. The essential difference is that in model C, cells of a different state can be produced only through cell division, whereas in model U cells can change state reversibly without dividing. The dynamics in the state space spanned by $N_A(t)$ and $N_B(t)$ are illustrated in figure 1. Initial populations $N_A(0)$ and $N_B(0)$ will be denoted as N_{A0} and N_{B0}, respectively.

2.1.1. Model C
Model C is defined by the following continuous-time transition rates:

$$[N_A, N_B] \rightarrow \begin{cases} [N_A+1, N_B], & \text{with rate } h_1^C := N_A \lambda_{AA}, \\ [N_A, N_B+1], & \text{with rate } h_2^C := N_A \lambda_{AB}, \\ [N_A-1, N_B+2], & \text{with rate } h_3^C := N_A \lambda_{BB}, \end{cases}$$

$$(1)$$

where the system's transition rates are denoted as h_k^C. In each of these three reactions, an A-state cell divides, producing two A-state cells, an A- and a B-state cell, or two B-state cells, at per-capita rates λ_{AA}, λ_{AB}, and λ_{BB} respectively. In this model, B-state cells do not divide.

Figure 1. Cell transitions in models C and U. (a) The state of the system is defined by the number of A-state cells (i) and B-state cells (j). Population dynamics can be interpreted as a 2D random walk with position-dependent jump rates. (b) Model C allows symmetric and asymmetric divisions. (c) Model U features only symmetric divisions, and state changes are made through reversible transitions that do not change cell number.

Let $p_{i,j}(t)$ be the probability that our cell population consists of i A-state and j B-state cells at time t, then the master equation for model C is given by

$$\frac{\mathrm{d}p_{i,j}}{\mathrm{d}t} = (i-1)\lambda_{AA}p_{i-1,j} + (i+1)\lambda_{BB}p_{i+1,j-2}$$

$$+ i\lambda_{AB}p_{i,j-1} - i(\lambda_{AA} + \lambda_{AB} + \lambda_{BB})p_{i,j}. \quad (2)$$

This master equation is a differential equation defining the time evolution of population probabilities $p_{i,j}(t)$ caused by stochastic transitions.

2.1.2. Model U

In model U, the state transition rates are:

$$[N_A, N_B] \rightarrow \begin{cases} [N_A+1, N_B], & \text{with rate } h_1^U := N_A\lambda_A, \\ [N_A-1, N_B+1], & \text{with rate } h_2^U := N_A k_{AB}, \\ [N_A, N_B+1], & \text{with rate } h_3^U := N_B\lambda_B, \\ [N_A+1, N_B-1], & \text{with rate } h_4^U := N_B k_{BA}, \end{cases}$$
$$(3)$$

with transition rates h_k^U. These reactions correspond to division of A-state and B-state cells into two daughter cells of the same state at per-capita rates λ_A and λ_B, respectively, and to dynamic state changes of an A-state into a B-state at per-capita rate k_{AB}, and of a B-state into an A-state at per-capita rate k_{BA}.

The master equation of model U is given by

$$\frac{\mathrm{d}p_{i,j}}{\mathrm{d}t} = (i-1)\lambda_A p_{i-1,j} + (j-1)\lambda_B p_{i,j-1}$$

$$+ (j+1)k_{BA}p_{i-1,j+1} + (i+1)k_{AB}p_{i+1,j-1}$$

$$- (i\lambda_A + j\lambda_B + ik_{AB} + jk_{BA})p_{i,j}. \quad (4)$$
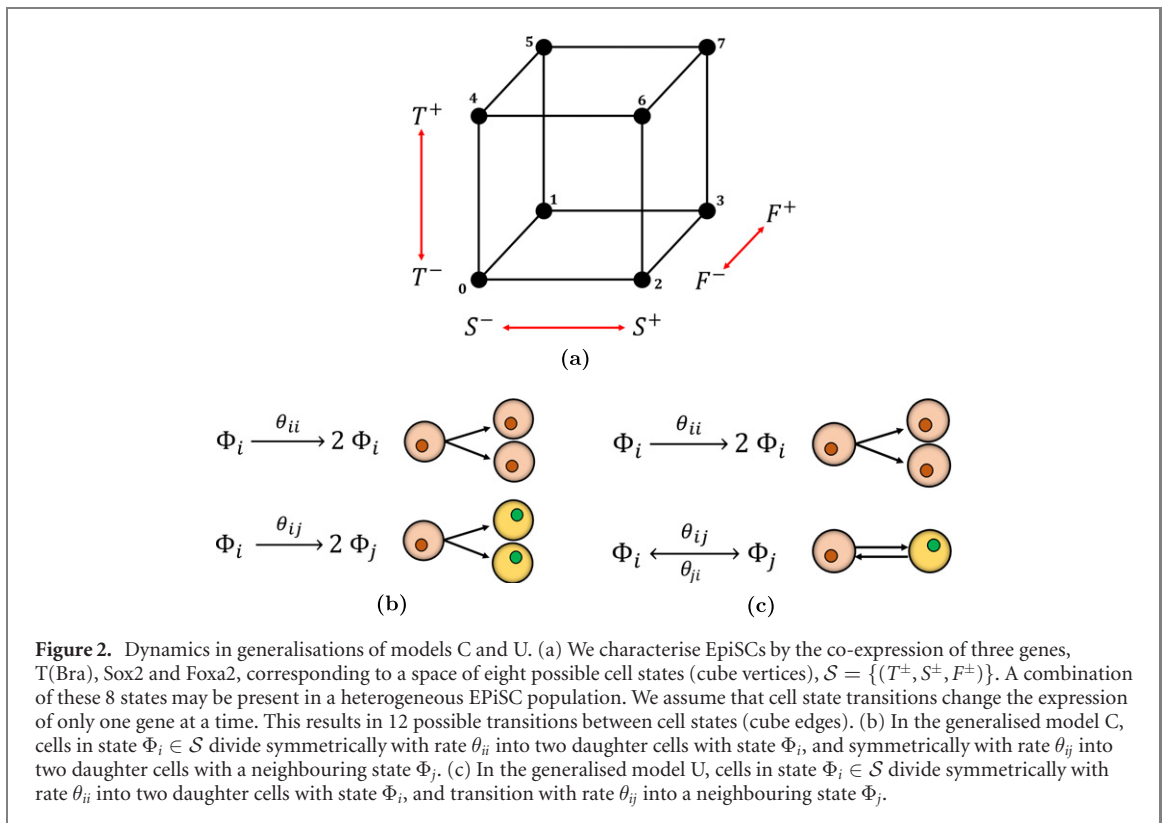
2.1.3. Generalisations of models C and U: epiblast stem cell models

While models C and U are useful for illustrative purposes, we require more cell states to make use of experimental data on epiblast-derived stem cell (EpiSC) dynamics. Here, we define EpiSC states by the binary expression of the transcription-factor genes T(Bra), Sox2, and Foxa2, which is what the experimental dataset contains information about [7]. There are thus 8 distinct cell states accessible for modelling, which we will denote by

$$\Phi_i := \begin{pmatrix} [T^-, S^-, F^-] \\ [T^-, S^-, F^+] \\ [T^-, S^+, F^-] \\ [T^-, S^+, F^+] \\ [T^+, S^-, F^-] \\ [T^+, S^-, F^+] \\ [T^+, S^+, F^-] \\ [T^+, S^+, F^+] \end{pmatrix}_i, \quad (5)$$

where the index T, S, F represents gene T(Bra), Sox2 and Foxa2, respectively. For simplicity, we make the assumption that cell state transitions change only one of the three genes T, F, S at a time. This leads to 12 allowed transitions between cell states, which can be conveniently visualised as a cube where vertices represent cell states and edges represent transition paths (see figure 2(a)).

The generalisation of model C retains the defining feature that new cell states can be created only at division of the parent. Specifically, we allow symmetric division of cells in each state, $\Phi_i \rightarrow 2\Phi_i$, at

Figure 2. Dynamics in generalisations of models C and U. (a) We characterise EpiSCs by the co-expression of three genes, T(Bra), Sox2 and Foxa2, corresponding to a space of eight possible cell states (cube vertices), $\mathcal{S} = \{(T^{\pm}, S^{\pm}, F^{\pm})\}$. A combination of these 8 states may be present in a heterogeneous EpiSC population. We assume that cell state transitions change the expression of only one gene at a time. This results in 12 possible transitions between cell states (cube edges). (b) In the generalised model C, cells in state $\Phi_i \in \mathcal{S}$ divide symmetrically with rate θ_{ii} into two daughter cells with state Φ_i, and symmetrically with rate θ_{ij} into two daughter cells with a neighbouring state Φ_j. (c) In the generalised model U, cells in state $\Phi_i \in \mathcal{S}$ divide symmetrically with rate θ_{ii} into two daughter cells with state Φ_i, and transition with rate θ_{ij} into a neighbouring state Φ_j.

per-capita rate θ_{ii}. Note that we can recover model C by setting all but one of these rates to zero. To keep the number of parameters small, we accommodate only the other symmetric cell division processes, $\Phi_i \to 2\Phi_j$ for any j sharing an edge with i, at per-capita rate θ_{ij}. These dynamics are illustrated in figure 2(b). The fact that we chose to include only symmetric cell divisions should not change the outcome of model comparison too much, considering the following argument: the asymmetric division process $\Phi_i \to \Phi_i + \Phi_j$ with $i \neq j$ effectively increases the Φ_j population by one. The same outcome can also be achieved either by a symmetric division $\Phi_j \to 2\Phi_j$ or by the sequence of two symmetric cell divisions $\Phi_i \to 2\Phi_j$, $i \neq j$ and $\Phi_j \to 2\Phi_k$, $j \neq k$.

Since cells in any state can divide, there are eight different reactions of the form $\Phi_i \to 2\Phi_i$. State-changing divisions of the form $\Phi_i \to 2\Phi_j$ can happen between all neighbouring cell states (i, j) in the network shown in figure 2(a); there are $3 \times 8 = 24$ such reactions, since each cell state has three neighbouring states. The total number of reaction types in extended model C is therefore $8 + 24 = 32$.

In principle therefore there are 32 parameters in this model. To reduce the size of the parameter space, we assume that the self-renewal rates are the same for all cell states, i.e., $\theta_{ii} = \theta_0$ for all i. We further assume that switching each gene $M \in \{T, S, F\}$ on or off does not depend on the state of the other two genes which are not changed. This assumption implies the transition rates of genes are independent

of each other, which is a first approximation only. We discuss relaxing this assumption in section 4.2. Each gene M thus contributes two independent rate constants θ_{M+}, θ_{M-} for switching on and off, respectively. The extended model C therefore has 7 free parameters:

$$\boldsymbol{\theta} = (\theta_0, \theta_{T+}, \theta_{T-}, \theta_{S+}, \theta_{S-}, \theta_{F+}, \theta_{F-}). \quad (6)$$

We now turn to the generalisation of model U, recalling that the original two-state model allowed symmetric division of both states, and dynamic differentiation in which one state could turn into the other. This is easily extended to 8 cell states Φ_i, wherein each cell state can perform symmetric cell division at rate θ_{ii}, and dynamical state changes between all pairs of neighbouring cell states can occur at rates θ_{ij} and θ_{ji}. See figure 2(c).

As in the extended model C, there are eight symmetric division rates, which we again set to a universal value $\theta_{ii} = \theta_0$. Similarly the rates of dynamic cell state changes are defined by the gene that is switched on or off, which again leads to a total of $8 + 24 = 32$ reaction types and 7 free parameters:

$$\boldsymbol{\theta} = (\theta_0, \theta_{T+}, \theta_{T-}, \theta_{S+}, \theta_{S-}, \theta_{F+}, \theta_{F-}). \quad (7)$$

An appealing feature of models C and U from a classical model-selection perspective is that they have the same number of free parameters. A problem when comparing models with different numbers of free parameters is that the models which offer more free

parameters will usually show better agreement with the data. However, this is because it can be more accurately fitted to the data, not because the underlying assumptions are more accurate in describing nature. Bayesian model selection generally eliminates this problem since it embodies Occam's razor: simpler models have higher posterior probabilities than more complex models if both fit the data equally well [8].

2.2. Bayesian inference and model selection

We use Bayesian methods to perform model selection and parameter inference. Given some experimental data ξ, we seek the probability distribution $p(\theta|\xi)$ over the parameters θ of either model C or model U. This is achieved by appealing to Bayes' theorem,

$$p(\theta|\xi) = \frac{p(\xi|\theta)p(\theta)}{p(\xi)}. \tag{8}$$

Here $p(\theta)$ encapsulates our *prior* beliefs as to appropriate values for the parameters θ (e.g., in the absence of any data), and $p(\xi|\theta)$ is the *likelihood* of the data ξ given a parameter choice. In this context, the likelihood is given by solving the master equation for the model under consideration. Lastly $p(\xi)$ is the probability of observing the data ξ, and is usually obtained indirectly by ensuring that the *posterior* distribution is correctly normalised. It is worth mentioning that data ξ can take very different forms, such as continuous trajectory data of cell population or population snapshots at a finite number of times. The form of ξ does not affect Bayes' theorem, but enters the calculation via the likelihood function $p(\xi|\theta)$.

If not only the model parameters θ are to be inferred, but also the model itself, we can use Bayesian methods for model comparison. Given data ξ and the two competing models M_C and M_U, we wish to calculate the probability of each model $p(M_j|\xi)$ given the data. Following Bayes' theorem, the *posterior odds* of the models are given by

$$\frac{p(M_C|\xi)}{p(M_U|\xi)} = \frac{p(\xi|M_C)q(M_C)}{p(\xi|M_U)q(M_U)}, \tag{9}$$

where $q(M_i)$ is the prior belief that model M_i is the appropriate description of the data. The selection of the 'best' model is usually not based on posterior odds, but the ratio of posterior to prior odds: that is, how much more the data point to one model over the other, given our prior beliefs:

$$B_{CU} := \frac{p(M_C|\xi)/p(M_U|\xi)}{q(M_C)/q(M_U)} = \frac{p(\xi|M_C)}{p(\xi|M_U)}. \tag{10}$$

This ratio is called the *Bayes factor* [9]. Following (10), the *marginal likelihoods* $p(\xi|M_j)$ are the quantities of key interest. They can be obtained by integrating over the parameter space θ^j of the corresponding model:

$$p(\xi|M_j) = \int p(\xi|\theta^j, M_j)p(\theta^j|M_j)\mathrm{d}\theta^j, \tag{11}$$

where $p(\theta^j|M_j) = q(\theta^j)$ is the parameter prior of model M_j.

All algorithms shown in this work were implemented using python3 with standard libraries (numpy, pandas, scipy, pyabc). All code can be accessed via GitHub (https://github.com/real-save/Stem-cell-inference).

3. Results

3.1. Coupling of differentiation and division leads to qualitatively different population dynamics

Our first set of results pertain to how populations evolve over time in models C and U in their original formulation with two cell states. In particular, we will see the utility of *moment generating functions*. In the following we will derive equations of the probability generating function and mean expressions for N_A and N_B for each of models C and U. In the last section we will also derive approximate expressions for mean populations $\langle\Phi\rangle$ in the extended models.

3.1.1. Population dynamics of model C

Applying the procedure of probability generating functions (see appendix B) to the master equation (2) for model C yields

$$\frac{\partial G}{\partial t} = \left[\lambda_{AA}(z_A^2 - z_A) + \lambda_{BB}(z_B^2 - z_A) + \lambda_{AB}(z_A z_B - z_A)\right]\frac{\partial G}{\partial z_A}. \tag{12}$$

This is a linear, first-order PDE for $G(z, t)$ and can be solved by the method of characteristics [10]. We assume the system initially consists of N_{A0} A-state cells and N_{B0} B-state cells, which translates to the initial condition $G(z_A, z_B, t = 0) = z_A^{N_{A0}} z_B^{N_{B0}}$. Using this initial condition, we find

$$G(z_A, z_B, t) = z_B^{N_{B0}}\left[\frac{1}{\rho\lambda_{AA}}\right.$$
$$\times \tan\left(\frac{t}{\rho} + \arctan\left(\rho\lambda_{AA}z_A + \frac{\rho\xi}{2}\right)\right)$$
$$\left. - \frac{\xi}{2\lambda_{AA}}\right]^{N_{A0}}, \tag{13}$$

where
$$\xi(z_B) = \lambda_{AB}(z_B - 1) - \lambda_{AA} - \lambda_{BB},$$
$$\rho(z_B) = \frac{2}{\sqrt{4\lambda_{AA}\lambda_{BB}z_B^2 - \xi^2}}. \tag{14}$$

The full time-dependence of mean population sizes then follow from (B.2):

$$\langle N_A(t)\rangle = N_{A0}e^{\Delta\lambda\,t}, \tag{15}$$

$$\langle N_B(t)\rangle = N_{B0} + N_{A0}\left(\frac{2\lambda_{BB} + \lambda_{AB}}{\Delta\lambda}\right)\left(e^{\Delta\lambda\,t} - 1\right), \tag{16}$$

where $\Delta\lambda := \lambda_{AA} - \lambda_{BB}$ is the difference of symmetric division rates. To understand these results, we note that cells in state A divide at per-capita rate λ_{AA}, and are lost at per-capita rate λ_{BB}. These competing processes lead to either an exponential growth ($\lambda_{AA} > \lambda_{BB}$) or exponential decay ($\lambda_{AA} < \lambda_{BB}$) in the mean number of cells in state A. Since there is no loss of cells from state B, the mean number of cells in state B either grows exponentially with time ($\lambda_{AA} > \lambda_{BB}$) or plateaus at a finite value ($\lambda_{AA} < \lambda_{BB}$). Mean population dynamics are therefore sensitive to changes in symmetric division rates since $\Delta\lambda$ controls the exponential growth/decay of mean populations. On the other hand, the rate of asymmetric cell divisions, λ_{AB}, weakly affects mean dynamics as it enters equation (16) only as a constant prefactor.

A particular feature of model C is that there is an absorbing state. This arises because A-state cells can be lost via the process $A \to 2B$, and B-state cells do not divide. Therefore, there is a finite probability that the number of cells in state A will eventually decrease to zero and the system dynamics will come to a halt. We can obtain $p_{ext}(t)$, the *extinction probability* that this has occurred by time t, from the moment generating function as follows:

$$p_{ext}(t) = \Pr\left[N_A(t) = 0\right] = \sum_{j=0}^{\infty} p_{0,j}(t) = G(0,1,t).$$
(17)

From (13) we find

$$p_{ext}(t) = \left(\frac{2\lambda_{BB}}{\lambda_{AA} + \lambda_{BB} + |\Delta\lambda|\coth\left(t|\Delta\lambda|/2\right)}\right)^{N_{A0}}.$$
(18)

In the special case $\lambda_{AA} = \lambda_{BB} = \lambda$ this simplifies to

$$p_{ext}(t) = \left(\frac{\lambda t}{1 + \lambda t}\right)^{N_{A0}}.$$
(19)

The extinction probability is monotonically increasing with time and converges to a finite value $\lim_{t\to\infty} p_{ext}(t) = p_{ext}^{max}$. For $\lambda_{BB} \geqslant \lambda_{AA}$, it follows $p_{ext}^{max} = 1$ and A-state cells will always die out. For $\lambda_{BB} < \lambda_{AA}$ however, $p_{ext}^{max} = (\lambda_{BB}/\lambda_{AA})^{N_{A0}} < 1$.

3.1.2. Population dynamics of model U
We can solve model U by following the same sequence of steps. We first arrive at the PDE

$$\frac{\partial G}{\partial t} = \left[\lambda_A(z_A^2 - z_A) + k_{AB}(z_B - z_A)\right]\frac{\partial G}{\partial z_A}$$
$$+ \left[\lambda_B(z_B^2 - z_B) + k_{BA}(z_A - z_B)\right]\frac{\partial G}{\partial z_B}.$$
(20)

The solution is given by

$$G(z_A, z_B, t) = \left[1 + e^{\Gamma t}\left(c_1(t)(z_A - 1) + c_2(t)(z_B - 1)\right)\right]^{N_{A0}}$$
$$\times \left[1 + e^{\Gamma t}\left(c_3(t)(z_A - 1) + c_4(t)(z_B - 1)\right)\right]^{N_{B0}}.$$
(21)

Here, c_1, c_2, c_3, c_4 are functions of time t:

$$c_1(t) = \cosh\left(\frac{\omega}{2}t\right) + \theta\,\sinh\left(\frac{\omega}{2}t\right),$$

$$c_2(t) = \frac{2k_{AB}}{\omega}\,\sinh\left(\frac{\omega}{2}t\right),$$

$$c_3(t) = \frac{2k_{BA}}{\omega}\,\sinh\left(\frac{\omega}{2}t\right),$$

$$c_4(t) = \cosh\left(\frac{\omega}{2}t\right) - \theta\,\sinh\left(\frac{\omega}{2}t\right),$$
(22)

and Γ, ω, θ are constants depending on model parameters ($\lambda_A, \lambda_B, k_{AB}, k_{BA}$):

$$\Gamma = \frac{1}{2}\left((\lambda_A + \lambda_B) - (k_{AB} + k_{BA})\right),$$

$$\omega = \sqrt{(\lambda_A - \lambda_B)^2 + (k_{AB} + k_{BA})^2 + 2(\lambda_A - \lambda_B)(k_{BA} - k_{AB})},$$

$$\theta = \frac{1}{\omega}\left((\lambda_A + k_{BA}) - (\lambda_B + k_{AB})\right).$$
(23)

The mean numbers of cells in each state are given by

$$\langle N_A(t)\rangle = e^{\alpha t}\left[N_{A0}\,\cosh(\beta t)\right.$$
$$\left. + \left(\frac{k_{BA}}{\beta}N_{B0} + \frac{\gamma}{2\beta}N_{A0}\right)\sinh(\beta t)\right],$$
(24)

$$\langle N_B(t)\rangle = e^{\alpha t}\left[N_{B0}\,\cosh(\beta t)\right.$$
$$\left. + \left(\frac{k_{AB}}{\beta}N_{A0} - \frac{\gamma}{2\beta}N_{B0}\right)\sinh(\beta t)\right],$$
(25)

where α, β, γ are constants depending on model parameters ($\lambda_A, \lambda_B, k_{AB}, k_{BA}$):

$$\alpha = \frac{1}{2}\left((\lambda_A + \lambda_B) - (k_{AB} + k_{BA})\right),$$
(26)

$$\beta = \frac{1}{2}\sqrt{\gamma^2 + 4k_{AB}k_{BA}},$$
(27)

$$\gamma = (\lambda_A + k_{BA}) - (\lambda_B + k_{AB}).$$
(28)

Over long times, the population means grow approximately exponentially, as in model C, but with a different rate $\alpha + \beta$. In contrast to model C, however, the population means never decay since $\alpha + \beta \geqslant 0$ as can be easily shown. This is due to the absence of an absorbing state in model U. It is worth mentioning that for any given parameter set ($\lambda_A, \lambda_B, k_{AB}, k_{BA}$) in model U, there exists a set of parameters ($\lambda_{AA}, \lambda_{AB}, \lambda_{BB}$) in model C for which the asymptotic behaviour ($t \to \infty$) of the means $\langle N_A\rangle$, $\langle N_B\rangle$ is identical in both models (see appendix A). Thus, we cannot use information about mean populations alone to distinguish model U from model C. The converse is not true, since $\langle N_A\rangle$ decays for some parameter regimes in model C, which is impossible in model U. This is part of the rationale for appealing to Bayesian methods to infer model parameters and facilitate model comparison.

3.1.3. Population dynamics of extended models C & U

As we showed in the last two sections, obtaining an exact solution of the probability generating function is already quite challenging for only two distinct cell states A and B. To obtain the probability generating function for the extended models including eight distinct cell states we would need to solve PDEs in 8 dimensions with 7 free parameters. Therefore we restrict the analysis of population dynamics in the extended models on obtaining approximate expressions of the mean populations $\langle \mathbf{\Phi} \rangle(t)$. Following the definition of allowed transitions in the extended models, on can derive a system of ODEs for the time evolution of moments $\langle \mathbf{\Phi} \rangle$ of the form [11]:

$$\frac{d\langle \mathbf{\Phi} \rangle}{dt} = \mathbf{M} \cdot \langle \mathbf{\Phi} \rangle, \tag{29}$$

where $\mathbf{M}$ is a 8×8 matrix which depends on the allowed transitions in the respective model (shown in appendix B). The general solution of mean populations is a linear combination $\langle \mathbf{\Phi} \rangle(t) = \sum_m c_m \mathbf{v}_m e^{\lambda_m t}$ where $\lambda_m, \mathbf{v}_m$ are the mth eigenvalue and eigenvector of matrix $\mathbf{M}$, respectively, and c_m are constants depending on initial conditions. For the extended models U and C, the matrix $\mathbf{M}$ is listed in the supplementary material (https://stacks.iop.org/PB/17/065009/mmedia), equations (B.4) and (B.5), and only has real eigenvalues. At late times $\lambda_m t \gg 1$, the term with the largest eigenvalue $\max(\{\lambda_m\})$ dominates the dynamics. For extended model U, the late time evolution according to the largest eigenvalue is given by

$$\langle \mathbf{\Phi} \rangle(t) \propto \mathbf{v}_0 \cdot e^{\theta_0 t}, \quad \mathbf{v}_0 = \begin{pmatrix} \theta_{T-}\theta_{S-}\theta_{F-} \\ \theta_{T-}\theta_{S-}\theta_{F+} \\ \theta_{T-}\theta_{S+}\theta_{F-} \\ \theta_{T-}\theta_{S+}\theta_{F+} \\ \theta_{T+}\theta_{S-}\theta_{F-} \\ \theta_{T+}\theta_{S-}\theta_{F+} \\ \theta_{T+}\theta_{S+}\theta_{F-} \\ \theta_{T+}\theta_{S+}\theta_{F+} \end{pmatrix}. \tag{30}$$

This result can be interpreted as follows: all cell states grow with the same rate θ_0 and therefore the total population size also grows with rate θ_0. The shape of $\mathbf{v}_0$ reflects that after sufficiently many division events, pairs of cell states are distributed according to their equilibrium distribution, $\langle \Phi_{M+} \rangle / \langle \Phi_{M-} \rangle = \theta_{M+}/\theta_{M-}$, where $\langle \Phi_{M\pm} \rangle$ is the mean population size of cells with gene $M \in \{T, S, F\}$ switched on and off, respectively. For extended model C, we can find two kinds of approximations for the late time evolution of mean populations. If rates for switching genes on and off are very unbalanced, $\theta_{M+} \gg \theta_{M-}$ or $\theta_{M+} \ll \theta_{M-}$, it follows

$$\langle \mathbf{\Phi} \rangle(t) \propto \mathbf{e}_i \cdot e^{\theta_0 t}, \tag{31}$$

where $\mathbf{e}_i$ equals one for those cell state favoured by the combination of imbalanced rates $\theta_{M+} \gg \theta_{M-}$ and zero for all other cell states. The effective growth rate

of the total population equals θ_0 and is therefore the same as in model U. On the other hand, if rates for switching genes on and off are approximately equal, $\theta_{M+} \approx \theta_{M-} = \theta_M$, late time dynamics are given by

$$\langle \mathbf{\Phi} \rangle(t) \propto \mathbf{1} \cdot e^{(\theta_0 + \theta_T + \theta_S + \theta_F)t}, \tag{32}$$

where $\mathbf{1}$ is a 7×1 all-ones vector. Cell states are therefore uniformly distributed and the overall population grows with effective growth rate $(\theta_0 + \theta_T + \theta_S + \theta_F)$. Differentiation events can therefore trigger increased population growth in extended model C. Populations grow with an increased rate $\tilde{\theta}_0, \theta_0 \leqslant \tilde{\theta}_0 \leqslant (\theta_0 + \theta_T + \theta_F + \theta_S)$, whose exact value depends on the rate imbalance θ_{M+}/θ_{M-} of the marker switching rates.

3.1.4. Summary

In this section we derived exact solutions of the probability generating function for both model C and model U. This allows an efficient and systematic way of calculating moments of the population distribution. We explicitly derived expressions for the time evolution of mean populations in both models. Neglecting cases for which A-state populations decrease in model C, mean populations grow exponentially over long time scales in both models with rates depending on the underlying model parameters. These expressions will be used in the next section for deriving an efficient algorithm for reaction rate inference (see also appendix D.2) and for proposing experimental measurement procedures which optimise parameter inference (see section 3.2.3).

3.2. Bayesian inference on minimal models

In the previous section we analysed the time-evolution of the distribution over cell states $N(t)$ as a function of model parameters θ. Now we will look at the inverse problem: obtaining the distribution over model parameters θ given population data N. In the first section we consider parameter inference given complete knowledge about population sizes $N(t)$ at all times. In the second section we generalise parameter inference to cases where we have incomplete data and population sizes $N(t_i)$ are only known at certain time points $\{t_i\}$.

3.2.1. Longer observations decrease uncertainty exponentially for complete data

We first suppose that we have complete knowledge of how the population sizes $N(t)$ have changed over time. In other words, each cell division and differentiation event is evident in the trajectory $N(t)$. Such data could, for example, come from live-imaging experiments in which each cell state can be identified and the full lineage be tracked. Since each reaction as defined by equations (1), (3) leads to a unique change in the number of cells in each state, we can infer the type of an occurred reaction from changes in $N(t)$. We define

τ_i, ν_i and $h_{\nu_i}(\tau_i)$ as the reaction times, the corresponding reaction types and the transition rates, respectively. We define the combined transition rate $h_0(t)$ as the sum over transition rates for all possible reactions that can take place at time t (which will depend on the state of the system at that time). Since each reaction event is modelled as a Poisson process with corresponding rate h_{ν_i}, the likelihood of a sequence of n reactions is given by [12]

$$L_{\mathrm{comp}}(\xi|\boldsymbol{\theta}) = \left[\prod_{i=1}^{n} h_{\nu_i}(\tau_{i-1})\right] \cdot e^{-\sum_{i=0}^{n} h_0(\tau_i)\cdot(\tau_{i+1}-\tau_i)},$$
(33)

where the subscript comp denotes complete data and $\tau_0 = 0$, $\tau_{n+1} = T$. Now, in the models we consider here, each transition rate h_{ν_i} is proportional to a per-capita rate θ_{ν_i} multiplied by the number of cells in state Φ_ν, which is the cell state associated with a reaction of type ν. If we define now $N_\nu(t)$ as the total number of cells in state Φ_ν at time t, we have that

$$L_{\mathrm{comp}}(\xi|\boldsymbol{\theta}) \propto \prod_\nu \left[\theta_\nu^{r_\nu}\ \exp\left(-\theta_\nu \int_0^T N_\nu(t)\,\mathrm{d}t\right)\right],$$
(34)

where the product is over each type of reaction ν (i.e. division or differentiation of each cell state) and r_ν is the number of times that reaction ν occurs during the observation window $0 \leqslant t \leqslant T$.

Note that what matters for Bayesian inference is the dependence of the likelihood on the parameters $\boldsymbol{\theta}$. Therefore factors that do not depend on these parameters (such as the population sizes N_ν that multiply each of the transition rates θ_k) can be subsumed into the constant of proportionality in (34).

A key feature of (34) is that it factorises into terms that each depend on a different reaction rate θ_ν. This means that each parameter value can be inferred independently, assuming that parameter priors are independent: $p(\boldsymbol{\theta}) = \prod_\nu p(\theta_\nu)$. That is, the posterior distribution of θ_ν is

$$p(\theta_\nu|\xi) \propto \theta_\nu^{r_\nu}\ \exp\left(-\theta_\nu \int_0^T N_\nu(t)\,\mathrm{d}t\right) p(\theta_\nu),$$
(35)

where $p(\theta_\nu)$ is the prior distribution over the reaction rate θ_ν. As before, the constant of proportionality can be obtained by normalising the posterior distribution $p(\theta_\nu|\xi)$. It is convenient for the prior distribution to be taken conjugate to the likelihood, which means that the product of prior and likelihood is of the same functional form as the prior. Here, the likelihood is a gamma distribution, and therefore the conjugate prior is also a gamma distribution [13]. If the prior is a gamma distribution with shape a_ν and rate b_ν, then the posterior is a gamma distribution with shape $a'_\nu = a_\nu + r_\nu$ and rate $b'_\nu = b_\nu + \int_0^T N_{\Phi_\nu}(t)\,\mathrm{d}t$, with mean $\langle\theta_\nu|\xi\rangle = a'_\nu/b'_\nu$ and variance $\mathrm{Var}(\theta_\nu|\xi) = a'_\nu/b'^2_\nu$ [14].

The confidence of parameter estimation can be quantified by the dispersion of the posterior distribution $p(\theta_\nu|\xi)$, measured by the *coefficient of variation* c_v (COV)

$$c_v(\theta_\nu|\xi) = \frac{\sqrt{\mathrm{Var}(\theta_\nu|\xi)}}{\langle\theta_\nu|\xi\rangle} = \frac{1}{\sqrt{a_\nu + r_\nu}}.$$
(36)

For experimental design it is essential to estimate how much data is needed to achieve parameter inference of a certain precision. More precisely, if we observe n identical systems over a time period T, we would like to know how the COV scales on average with sample size n and observation time T. As shown in appendix C, the asymptotic scaling of the mean COV with observation time T in model C is given by

$$\langle c_v\rangle(\theta_\nu|\xi_1,\ldots,\xi_n) \propto \begin{cases} \dfrac{\exp\left(-\frac{\Delta\lambda}{2}T\right)}{\sqrt{N_{\mathrm{A0}}n}} & \Delta\lambda > 0, \\[3mm] \dfrac{1}{\sqrt{N_{\mathrm{A0}}n}} & \Delta\lambda < 0, \end{cases}$$
(37)

where $\Delta\lambda = \lambda_{\mathrm{AA}} - \lambda_{\mathrm{BB}}$. For model U, we obtain

$$\langle c_v\rangle(\theta_\nu|\xi_1,\ldots,\xi_n) \propto \frac{\exp\left(-\frac{(\alpha+\beta)}{2}T\right)}{\sqrt{n}},$$
(38)

where α and β are functions of rate constants as given by (26, 27). Note that the mean COV of θ_ν depends on the rate constant θ_ν itself. This result can be useful for experimental design if a rough estimate for the magnitude of θ_ν is available. If not, one has to rely only on the scaling with sample size n.

To summarise, if we have a complete history of all the population sizes $N_\nu(t)$ over time, we can infer the reaction rate θ_ν by counting how many times the reaction ν occurs, and by integrating the size of the population that drives each reaction, $N_\nu(t)$, over time t. Scaling laws for the confidence of parameter estimation can be obtained as a function of sample size n and observation time T. The uncertainty of the parameter estimation decreases exponentially with observation time, but only decays as a power law with sample size. Besides parameter inference, the aim of experiments is often to compare competing models based on data ξ. In the complete-data scenario, reaction types can be directly inferred from changes in population. Hence, if reactions are observed which are impossible in either model C or model U, we can rule out the respective model and model selection is trivial. This will be very different in scenarios where we only have access to snapshot data, as we now show.

3.2.2. Efficient sampling enables inference for analytically intractable snapshot data

The results of the previous section depend on having complete knowledge of population sizes over time. It is unlikely that this level of detail will be available in practice. More likely are a set of *snapshots*, that is, measurements of N at a finite set of times $t_0 = 0, t_1, t_2, \ldots, t_n$.

In the previous section we obtained an expression for the posterior distribution $p(\boldsymbol{\theta}|\xi)$ given a

Algorithm 1. Sample path MCMC for parameter inference.

1: Initialise algorithm with valid trajectory segments $\{\xi_i\}$, a parameter set $\boldsymbol{\theta}$ and set $i = 1$.

2: Propose trajectory segment $\xi_i^\star$ between time points t_{i-1} and t_i with fixed endpoints $\boldsymbol{N}(t_{i-1})$ and $\boldsymbol{N}(t_i)$.

3: Metropolis-Hastings step: replace original segment ξ_i by $\xi_i^\star$ with acceptance probability $\min(1, A_i)$, with acceptance ratio A_i
   from equation (D.6). Otherwise keep original segment ξ_i.

4: If $i < n$, set $i = i + 1$ and go to (2), otherwise go to (5)

5: Form an updated complete trajectory ξ by joining the updated segments ξ_i.

6: Sample a new parameter set $\boldsymbol{\theta}$ from the complete-data posterior $p(\boldsymbol{\theta}|\xi)$.

7: Output the current $\boldsymbol{\theta}$, set $i = 1$ and go back to (2).

Figure 3. Sample path Markov chain Monte Carlo (MCMC) algorithm for sampling N_{MC} reaction rate values $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{\theta}|\{\boldsymbol{N}(t_i)\})$, given snapshot-data $\{\boldsymbol{N}(t_i)\}$. Algorithm motivated by [12].
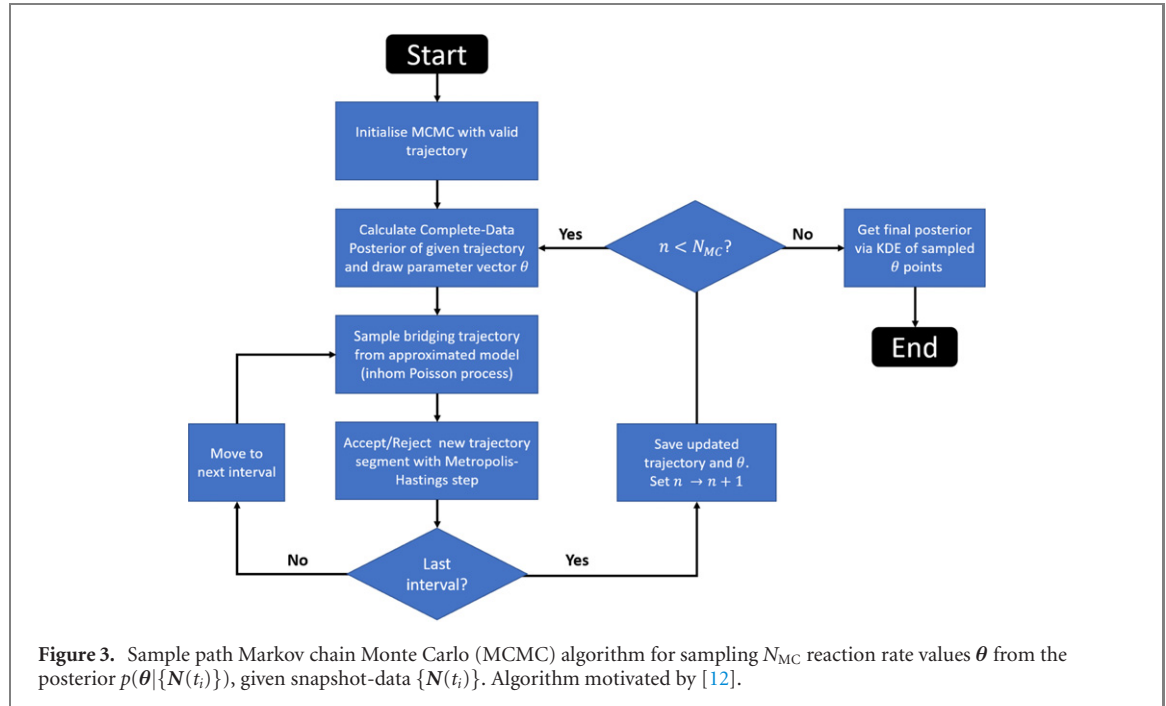
completely-specified trajectory ξ. To obtain the corresponding quantity from snapshot data, we integrate over all trajectories ξ that are consistent with the observations $\{\boldsymbol{N}(t_i)\}$, weighted by the probability $p(\xi|\{\boldsymbol{N}(t_i)\})$ which is conditioned on passing through the observation points. That is,

$$p(\boldsymbol{\theta}|\{\boldsymbol{N}(t_i)\}) = \int D\xi \, p(\boldsymbol{\theta}|\xi)p(\xi|\{\boldsymbol{N}(t_i)\}). \quad (39)$$

Since the underlying model is Markovian, we have that

$$p(\xi|\{\boldsymbol{N}(t_i)\}) = \prod_{i=1}^{n} p(\xi_i|\boldsymbol{N}(t_{i-1}), \boldsymbol{N}(t_i)), \quad (40)$$

where ξ_i is the portion of the trajectory corresponding to the time interval $[t_{i-1}, t_i]$.

This observation implies that the posterior distribution can be sampled using algorithm 1 (see also figure 3).
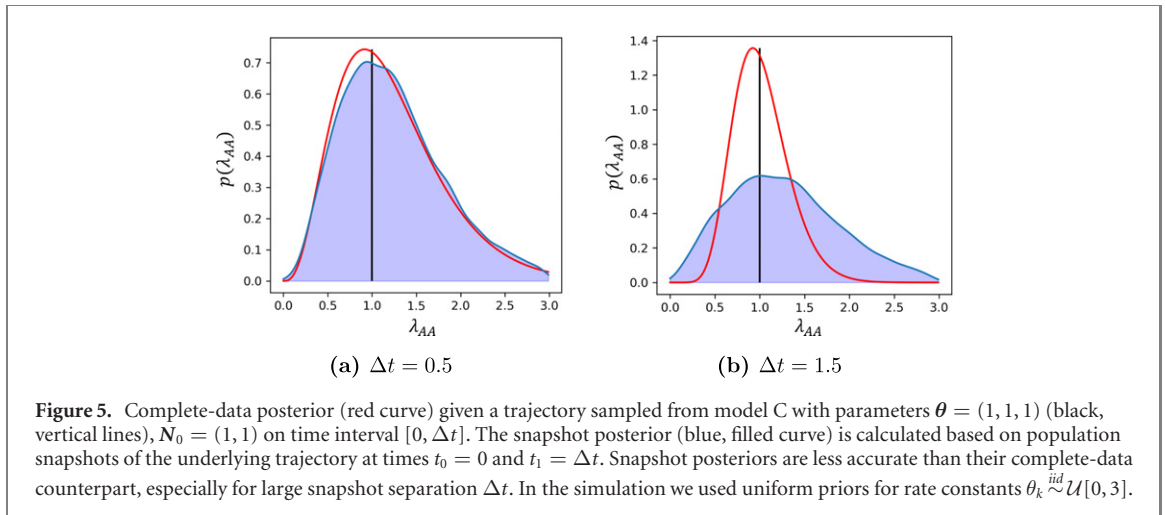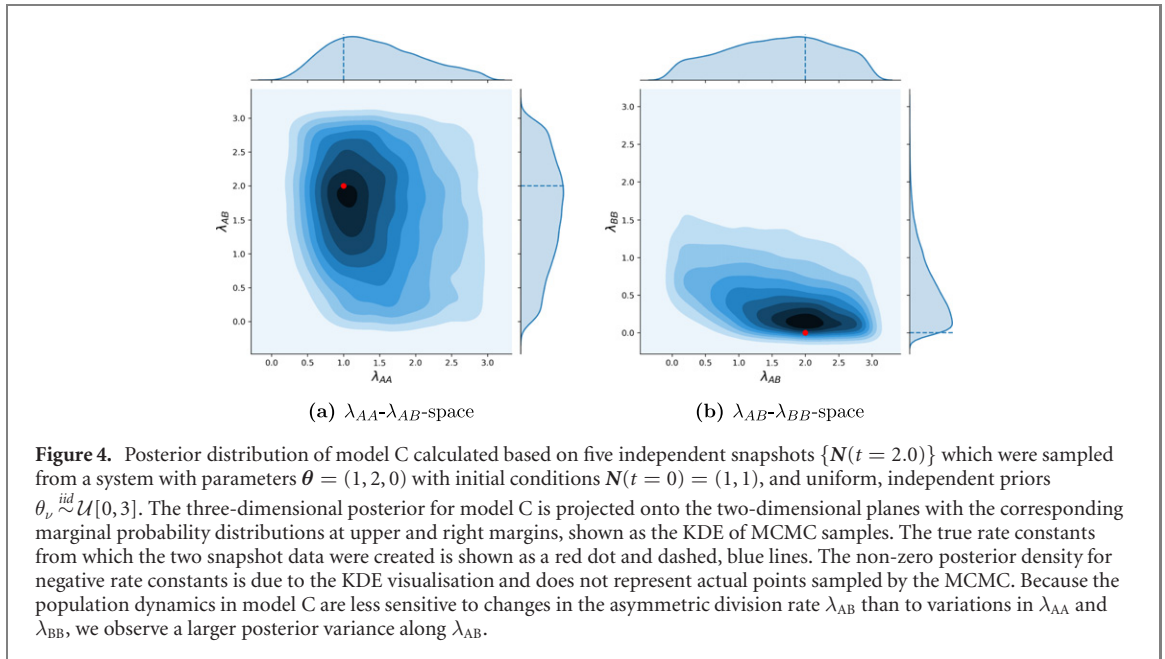
Previously we used the Gillespie algorithm to sample random trajectories with a fixed starting point. However, to propose new trajectory segments $\xi_i^\star$ in step 2 of algorithm 1, we need to sample a random trajectory where both end points are fixed. This is a much harder problem and we will use an approach based on the Metropolis-Hastings-Green algorithm [12, 15] to achieve this. The basic idea is to draw $\xi_i^\star$ from a

distribution that approximates the true distribution, and adjust the acceptance ratio A accordingly.

To construct the approximate distribution, we make use of the exact results for the mean population sizes that were derived in sections 3.1.1 and 3.1.2. We found that typically these grow or decline exponentially. Therefore we consider a set of reactions ν whose transition rates are *deterministic* functions of time,

$$h_\nu^*(t) = N_\nu(0)\theta_\nu \left(\frac{N_\nu(T)}{N_\nu(0)}\right)^{t/T}, \quad (41)$$

where t is the time from the start of the trajectory segment, and T is the length of the trajectory segment. The simplification here is that the probability a reaction occurs at time $0 \leqslant t \leqslant T$ is not affected by the sequence of reactions that has occurred up to time t, as would be the case for the true distribution. This amounts to sampling r_ν reactions of type ν from an inhomogeneous Poisson process with the time-dependent transition rates h_ν^*. The procedure is to first draw the numbers of reactions r_ν, and then sample the reaction times from an inhomogeneous Poisson process. To account for the approximation in sampling the trajectory, the acceptance ratio in the sample path MCMC algorithm must be chosen accordingly, which is shown in detail in appendix D.

(a) λ_{AA}-λ_{AB}-space   **(b)** λ_{AB}-λ_{BB}-space

Figure 4. Posterior distribution of model C calculated based on five independent snapshots $\{N(t = 2.0)\}$ which were sampled from a system with parameters $\theta = (1, 2, 0)$ with initial conditions $N(t = 0) = (1, 1)$, and uniform, independent priors $\theta_\nu \overset{iid}{\sim} \mathcal{U}[0, 3]$. The three-dimensional posterior for model C is projected onto the two-dimensional planes with the corresponding marginal probability distributions at upper and right margins, shown as the KDE of MCMC samples. The true rate constants from which the two snapshot data were created is shown as a red dot and dashed, blue lines. The non-zero posterior density for negative rate constants is due to the KDE visualisation and does not represent actual points sampled by the MCMC. Because the population dynamics in model C are less sensitive to changes in the asymmetric division rate λ_{AB} than to variations in λ_{AA} and λ_{BB}, we observe a larger posterior variance along λ_{AB}.

(a) $\Delta t = 0.5$ **(b)** $\Delta t = 1.5$

Figure 5. Complete-data posterior (red curve) given a trajectory sampled from model C with parameters $\theta = (1, 1, 1)$ (black, vertical lines), $N_0 = (1, 1)$ on time interval $[0, \Delta t]$. The snapshot posterior (blue, filled curve) is calculated based on population snapshots of the underlying trajectory at times $t_0 = 0$ and $t_1 = \Delta t$. Snapshot posteriors are less accurate than their complete-data counterpart, especially for large snapshot separation Δt. In the simulation we used uniform priors for rate constants $\theta_k \overset{iid}{\sim} \mathcal{U}[0, 3]$.

3.2.3. Optimal snapshot timing can improve parameter inference

We now apply the sample path MCMC algorithm to understand how well it performs given specific snapshots. Recall that this algorithm generates a sequence of rate-constant vectors, θ, which (when converged) are distributed according to the correct posterior distribution $p(\theta | \{N(t_i)\})$.

Our first test is a single run of model C, in which the initial condition is $N(0) = (1, 1)$ and after time $t = 1.5$ has evolved to $N(1.5) = (3, 7)$ with reaction rates $\theta = (\lambda_{AA}, \lambda_{AB}, \lambda_{BB}) = (1, 0, 1)$. With a large amount of data, one would expect the posterior distribution to be peaked around this point. Two-dimensional projections of the three-dimensional posterior distributions obtained via kernel density estimation (KDE) are shown in figures 4 and E3, and shows that the parameter estimates are reasonably well constrained despite the small amount of data (2 snapshots) used for the inference. Throughout, we use Scott's rule [16] for KDE bandwidth estimation. Unlike the complete-data scenario (section 3.2.1), the likelihood (and so the posterior) *does not* factorise into independent component likelihoods (cf equation (35)). This means given snapshot-data, rate constants are correlated and rate inference *cannot* be carried out separately. The shape of the joint posterior distributions in figure 4 indicates some correlation between λ_{AB} and λ_{BB}.

We now investigate how the parameter estimates are affected by the length of time, Δt, between the two snapshots. We anticipate that when Δt is small, only a small number of reactions could reasonably have occurred, and so one would not expect to lose much from having only the trajectory endpoints (as opposed to the full trajectory). On the other hand, when Δt is large, we would expect estimates based on snapshots to be less tightly constrained than those based on full trajectories.
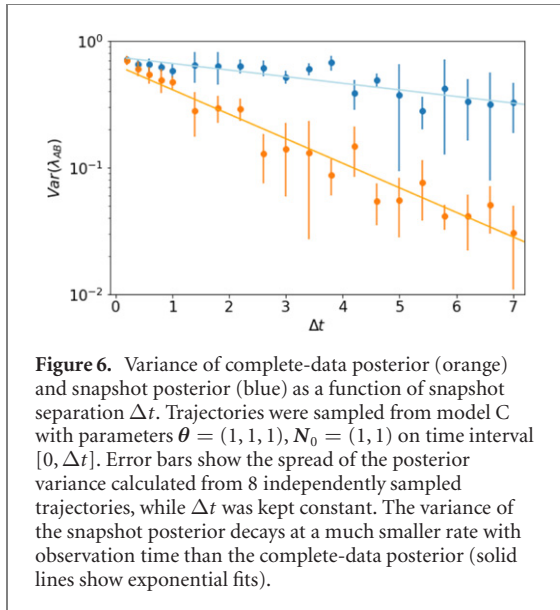
Figure 6. Variance of complete-data posterior (orange) and snapshot posterior (blue) as a function of snapshot separation Δt. Trajectories were sampled from model C with parameters $\boldsymbol{\theta} = (1, 1, 1)$, $\boldsymbol{N}_0 = (1, 1)$ on time interval $[0, \Delta t]$. Error bars show the spread of the posterior variance calculated from 8 independently sampled trajectories, while Δt was kept constant. The variance of the snapshot posterior decays at a much smaller rate with observation time than the complete-data posterior (solid lines show exponential fits).

Our procedure is to generate a trajectory from model C of length Δt, starting from $\boldsymbol{N}(0) = (1, 1)$ and with $\boldsymbol{\theta} = (\lambda_{\mathrm{AA}}, \lambda_{\mathrm{AB}}, \lambda_{\mathrm{BB}}) = (1, 1, 1)$. We then perform the parameter inference with both the full trajectory, and with the two endpoints. The resulting posterior distributions for the parameter λ_{AA} for different Δt are shown in figure 5.

The snapshot posteriors (blue, filled curves) are always less accurate with respect to the true parameters than their complete-data counterpart (red lines), as one would expect. As more reaction events are contained in the full trajectory ξ for longer times, it provides more information about the system and the variance of the complete-data posterior decreases. The *marginal information* stored in snapshot-data increases too with reaction numbers, however at a much smaller rate than the full trajectory ξ (see figure 6).

Suppose now we have the ability to take more than two snapshots, and have freedom to choose when they are taken. Is it more informative to space them out evenly, or to cluster them together? Given that the total rate at which reactions occur is proportional to the number of cells in the population, we expect to learn more from closely-spaced observations when the population size is large than when it is small.

More precisely, we can attempt to maximise the information content of multiple snapshots by keeping the mean number of reactions in between them constant. For example, in model C we know that the total reaction rate is $N_{\mathrm{A}}(t)(\lambda_{\mathrm{AA}} + \lambda_{\mathrm{AB}} + \lambda_{\mathrm{BB}})$ and from equation (16) that the mean number of A-state cells grows exponentially as $N_{\mathrm{A}0}e^{\Delta\lambda t}$, where $\Delta\lambda = \lambda_{\mathrm{AA}} - \lambda_{\mathrm{BB}}$. We can therefore deduce that the mean number of reactions between times t_{i-1} and t_i is

$$\langle r_0 \rangle_i = N_{\mathrm{A}0} \frac{\lambda_{\mathrm{AA}} + \lambda_{\mathrm{AB}} + \lambda_{\mathrm{BB}}}{\Delta\lambda} \left(e^{\Delta\lambda t_i} - e^{\Delta\lambda t_{i-1}} \right). \tag{42}$$

Setting this number equal for all i, with the constraint that $t_0 = 0$ and $t_n = T$, we find that the ith snapshot should be taken at time

$$t_i = \frac{1}{\Delta\lambda} \ln\left[1 + \frac{i}{n} \left(e^{\Delta\lambda T} - 1 \right) \right], \tag{43}$$

where n is the total number of snapshots taken in the time interval. We can compare the variance of posterior distributions obtained from differently spaced snapshots by replacing $\Delta\lambda$ with a variation parameter z in equation (43). For $z < 0$ snapshots are concentrated towards the start of the observation window, $z = 0$ represents uniformly spaced snapshots and $z > 0$ leads to a higher snapshot density towards the end of the observation window (see figure 7 for example). In figure 8 we compare the variance of posterior distributions obtained from differently spaced snapshot times over a range $-5 \leqslant z \leqslant 7$. We see that parameter estimates are most precise if the average number of reactions between snapshots are kept constant ($z = \Delta\lambda$), as proposed. This confirms our intuition that one is likely to be best served by sampling larger populations more frequently. A similar analysis is possible for model U, which yields the same optimal snapshot times as in equation (43), but $\Delta\lambda$ is replaced by $\alpha + \beta$ from equation (28).
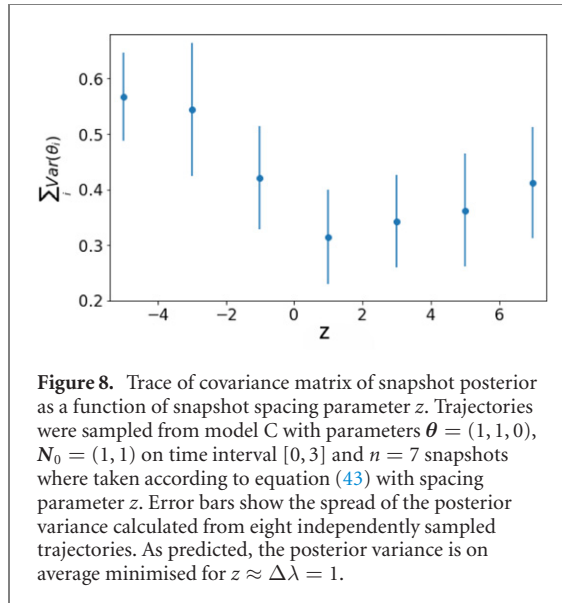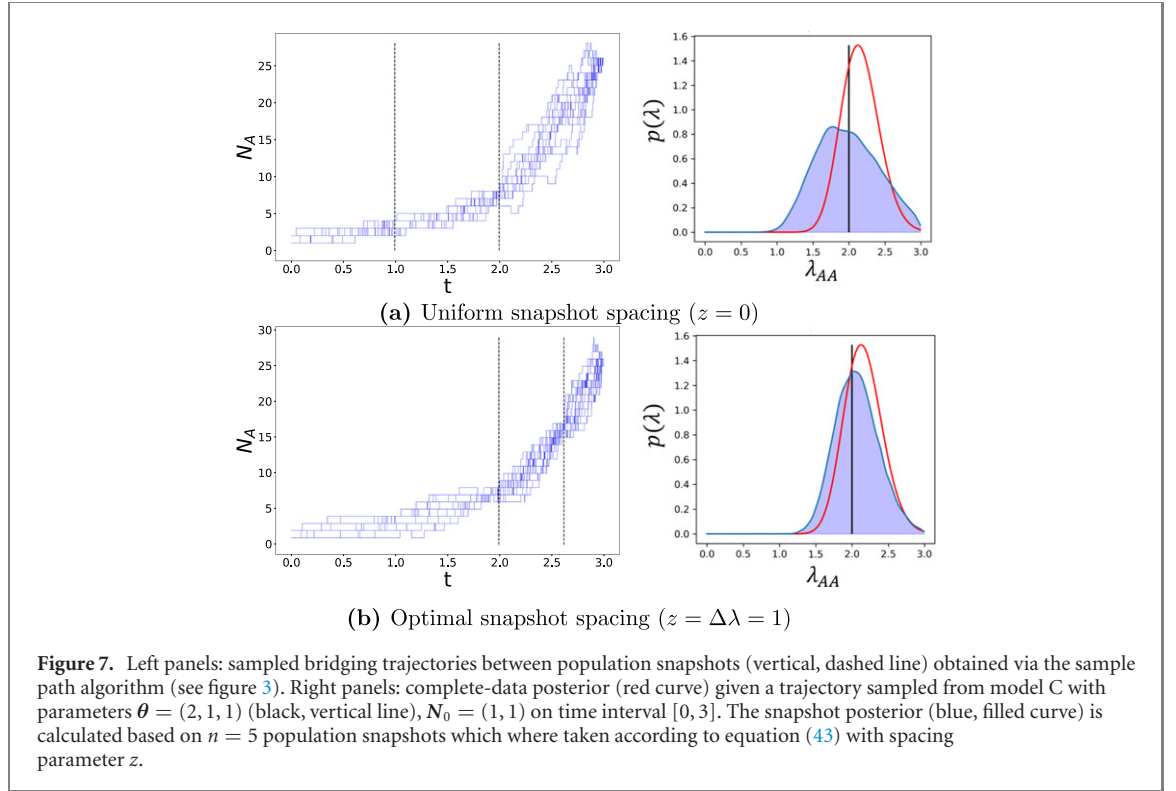
3.3. Application to epiblast stem cell data
Having established that parameter inference and model selection is viable with relatively limited amounts of data in the context of models C and U with two cell states, we turn now to the more challenging case of epiblast stem cell data with eight states as described in section 2.1.3. This may elucidate whether cell division and state changes are independent or coupled in this particular biological system.

3.3.1. Data structure
The structure of our dataset is largely influenced by measurement restrictions of the experimental setup (Schumacher *et al*, in preparation). The experiment starts with single cells of only partially known state: prior to the experiments cells have been sorted for the expression of marker $T^{\pm}$. From previous experiments measuring steady-state populations of $T^{\pm}$-sorted stem cells [7] we can calculate the distribution of marker expressions for F, S as $p(F^+|T^+) \approx 0.49$, $p(F^+|T^-) \approx 0.26$, $p(S^+|T^+) \approx 0.75$, $p(S^+|T^-) \approx 0.46$, with complementary probabilities $p(F^-|T) = 1 - p(F^+|T)$ and $p(S^-|T) = 1 - p(S^+|T)$, and marker expressions are assumed to be independent, so that $p(F, S|T) = p(F|T)p(S|T)$.

After time Δt, individuals cells will have grown into small colonies (clones). These are fixed, the number of cells in each clone counted and the expression of either T, F or T, S are measured. We can therefore only obtain information about the sum of underlying populations, where the sum is taken over the

(a) Uniform snapshot spacing ($z = 0$)

(b) Optimal snapshot spacing ($z = \Delta\lambda = 1$)

Figure 7. Left panels: sampled bridging trajectories between population snapshots (vertical, dashed line) obtained via the sample path algorithm (see figure 3). Right panels: complete-data posterior (red curve) given a trajectory sampled from model C with parameters $\boldsymbol{\theta} = (2, 1, 1)$ (black, vertical line), $\boldsymbol{N_0} = (1, 1)$ on time interval $[0, 3]$. The snapshot posterior (blue, filled curve) is calculated based on $n = 5$ population snapshots which where taken according to equation (43) with spacing parameter z.

Figure 8. Trace of covariance matrix of snapshot posterior as a function of snapshot spacing parameter z. Trajectories were sampled from model C with parameters $\boldsymbol{\theta} = (1, 1, 0)$, $\boldsymbol{N_0} = (1, 1)$ on time interval $[0, 3]$ and $n = 7$ snapshots where taken according to equation (43) with spacing parameter z. Error bars show the spread of the posterior variance calculated from eight independently sampled trajectories. As predicted, the posterior variance is on average minimised for $z \approx \Delta\lambda = 1$.

unknown marker (either S or F) in a given measurement. As an example: the number of cells possessing marker expression $[T^+, F^+]$ is given by the sum of cell populations with marker expression $[T^+, F^+, S^-]$ and $[T^+, F^+, S^+]$: $N_{[T+,F+]} = N_{[T+,F+,S-]} + N_{[T+,F+,S+]}$. In general,

$$N_{[M1,M2]} = \sum_{M3} N_{[M1,M2,M3]}, \tag{44}$$

where $M1$, $M2$ are measured marker expressions and $M3$ is the unknown marker. To compare snapshot data obtained from simulations to experimental data, we first have to calculate the *projected populations* $\{N_{[M1,M2]}\} := \mathcal{D}^{\text{sim}}$ before comparing it to

measured populations from experiments. Population snapshot data was experimentally recorded from cell colonies subject to two different experimental conditions: CHIR ($\mathcal{D}_{\text{CHIR}}$), which refers to the presence of the differentiation factor Chiron, and EPISC ($\mathcal{D}_{\text{EPISC}}$), which refers to the absence of this factor [7]. The first data set $\mathcal{D}_{\text{CHIR}}$ consists of eight measurement series recorded at different times and for marker expressions:

(a) Initial marker T^+, marker measured after $\Delta t = 2d$ and $\Delta t = 3d$: T, F;

(b) Initial marker T^+, marker measured after $\Delta t = 2d$ and $\Delta t = 3d$: T, S;

(c) Initial marker T^-, marker measured after $\Delta t = 2d$ and $\Delta t = 3d$: T, F;

(d) Initial marker T^-, marker measured after $\Delta t = 2d$ and $\Delta t = 3d$: T, S.

The second data set $\mathcal{D}_{\text{EPISC}}$ consists of four measurement series recorded at different times and for marker expressions:

(a) Initial marker T^+, marker measured after $\Delta t = 3d$: T, F;

(b) Initial marker T^+, marker measured after $\Delta t = 3d$: T, S;

(c) Initial marker T^-, marker measured after $\Delta t = 3d$: T, F;

(d) Initial marker T^-, marker measured after $\Delta t = 3d$: T, S.

Each measurement series in both sets contains between 98 and 142 observations of projected population snapshots.

3.3.2. Inference using ABC

The main challenge in applying the MCMC algorithm developed above to the EpiSC cell state network is the high dimensionality of the state space. The state space of our system is given by population numbers $\{N_\Phi\}$ of all 8 cell states Φ_i. We need to evaluate snapshot likelihoods of the form $L_{SS}(\{N_\Phi(\Delta t)\}|\boldsymbol{\theta})$ by an MC estimate, so a sufficiently large ensemble of trajectories is necessary to populate most parts of state space $\{N_\Phi\}$. The experimental data $\mathcal{D}_{CHIR}$ and $\mathcal{D}_{EPISC}$ contain up to 8 cells per cell state and therefore the MC estimate would need to sample up to $8^8 \approx 2 \times 10^7$ possible system states. Unfortunately, sampling a sufficiently large ensemble from $\sim 10^7$ system states turned out to be computationally not feasible. We therefore turn to *approximate Bayesian computation* (ABC) [17]. In our case, ABC speeds up calculations by several order of magnitudes by comparing data based on lower-dimensional *summary statistics* only, at the expense of introducing an additional error by approximating posterior distributions.

We choose the following summary statistics: the mean $\langle \boldsymbol{N}_\Phi \rangle$, median med$\{\boldsymbol{N}_\Phi\}$, and standard deviation $\boldsymbol{\sigma}_\Phi$. We use a sequential Monte Carlo algorithm (SMC-ABC) for parameter inference and model selection [18, 19]. We verify our results also using a rejection-ABC algorithm (modified from [17]) for parameter inference and model selection, which is presented in appendix E.

In ABC, a distance function $D(s(\mathcal{D}^{sim}), s(\mathcal{D}))$ classifies how 'close' the summary statistic of observed and simulated data are. A tolerance $\epsilon \geqslant 0$ controls the level of agreement between both summary statistics and $\epsilon > 0$ introduces the eponymous approximation in ABC. We choose the following summary statistic for a measurement series d: $\boldsymbol{s}_i(d) = [\langle N_i \rangle, \sigma_{\{N_i\}}, \text{Med}(N_i)]^T$, where $i \in \{1, 2, 3, 4\}$ is indexing the projected populations (expression of either T and F or T and S, starting from either $T+$ or $T-$ cells). This summary statistic captures the centre, spread, and skewness of the respective distribution $p(N_i)$. We use the sample median instead of the skewness as it can be calculated more easily than the third central moments needed for skewness estimation. We use the following norm $D(s(\mathcal{D}^{sim}), s(\mathcal{D}))$ to compare summary statistics of experimental and simulated data:

$$D(s(\mathcal{D}^{sim}), s(\mathcal{D})) = \sum_k \sum_{i=1}^4 \|\boldsymbol{s}_i(d_k^{sim}) - \boldsymbol{s}_i(d_k)\|_1,$$

(45)

where d_k^{sim} and d_k is the summary statistic of the kth measurement series of simulated and experimental data, respectively, and $\|.\|_1$ is the L_1-norm. The choice of norm does not significantly affect the shape of parameter posteriors if the acceptance threshold ϵ is sufficiently small (see figure G1). Note that $k \in \{1, \ldots, 8\}$ for data set D_{CHIR} and $k \in \{1, \ldots, 4\}$ for data set D_{EPISC}. The output of this ABC algorithm

is a set of parameters $\{\boldsymbol{\theta}\}$ distributed according to $p(\boldsymbol{\theta}|D(s(\mathcal{D}^{sim}), s(\mathcal{D})) < \epsilon$. For sufficiently small ϵ, this distribution should be a good approximation of the true posterior $p(\boldsymbol{\theta}|s(\mathcal{D}))$ which we confirmed by comparing the ABC posterior of cell division rate θ_0 in model U with the analytical solution (see figure F1).

For parameter inference and model selection we choose identical log-uniform priors for all rate constants in both models $\log_{10}(\theta_k) \sim \mathcal{U}[-2, 0])$ (in units of per day). The upper cut-off can be justified by inspection of the posterior distributions, which are close to zero before this threshold is reached. Cell transitions of a certain type should occur sufficiently often over the observation period of the experiment to be considered in our model. By constraining the prior distribution $p(\theta_k < 0.01) = 0$, we only discard very rare cell transitions which would occur less than once every 100 days in a single cell. Considering that total population sizes in the experiment are around 10 cells and the observation period is three days, this lower cut-off seems reasonable.

We run the SMC-ABC algorithm, which progressively lowers tolerance ϵ until the acceptance probability becomes too small and ϵ cannot be decreased much further without impractical computational costs. The Bayes factor B_{CU} can be calculated for every SMC-iteration from the marginal posterior distributions of models M_C and M_U. This was done for both datasets $\mathcal{D}_{CHIR}$ and $\mathcal{D}_{EPISC}$ and the resulting Bayes factors are shown in figure 9 as a function of ϵ/N_{set}. We normalised the tolerance value ϵ by the number of measurement series N_{set} in a dataset since the our distance function equation (45) sums over the deviation of every measurement series $1 \leqslant k \leqslant N_{set}$. For the lowest achievable tolerance (see red dots in figure 9) we obtain Bayes factors of $B_{CU} = 4.8 \pm 0.5$ for $\mathcal{D}_{CHIR}$ and $B_{CU} = 20 \pm 4$ for $\mathcal{D}_{EPISC}$). This constitutes moderate to strong evidence in favour of model C dynamics, i.e. a coupling of cell division and state transitions.

The posterior distribution of model C log-rates $\log_{10}(\theta_\nu)$ are shown in figure 10(a) for both data sets. While the cell division rate θ_0 depends only slightly on the experimental conditions ($\log_{10}(\theta_0) \approx -1.1$), the rates for switching markers on and off differ considerably in different experimental conditions. The rate imbalance θ_{M+}/θ_{M-} of marker switching rates is shown in figure 10(b), which implies:

- Cells in CHIR conditions are mostly in state $[T^-, S^-, F^+]$. Although only a small subpopulation of cells show marker expression T^+ in CHIR conditions, there are still more T^+ cells than under EPISC conditions which is consistent with the raw data from [7]. The rate-imbalance of markers F and S are quite small, so in steady-state there may be significant subpopulations with marker expression S^+ or F^-.
- Cells in EPISC conditions are biased towards the state $[T^-, S^+, F^-]$. The rate-imbalance of
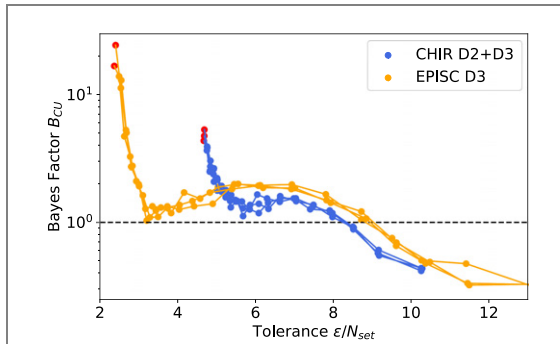
Figure 9. Bayes factors B_{CU} of datasets $\mathcal{D}_{CHIR}$ and $\mathcal{D}_{EPISC}$ over a range of tolerance values ϵ with identical log-uniform priors for all rate constants in both models $\log_{10}(\theta_k) \sim \mathcal{U}[-2, 0]$). The *x*-axis was re-scaled by the number of measurement series N_{set} in the respective dataset ($N_{set} = 8$ in $\mathcal{D}_{CHIR}$, $N_{set} = 4$ in $\mathcal{D}_{EPISC}$). Red dots show the minimal epsilon value to which the SMC-ABC algorithm converged after 20 iterations, resulting in $B_{CU} = 4.8 \pm 0.5$ for $\mathcal{D}_{CHIR}$ and $B_{CU} = 20 \pm 4$ for $\mathcal{D}_{EPISC}$. Three independent SMC-ABC runs were performed (shown as multiple lines), and the error of B_{CU} was estimated by the spread between runs. Since one of three runs for $\mathcal{D}_{EPISC}$ was still at a larger tolerance value after 20 iterations than the other two, we excluded it from the final BF calculation.

marker *F* is quite small, so in steady-state there may be sub-populations with marker expression F^+.

It is worth mentioning that although model U was shown to be less likely, the posterior distribution of model U shows a very similar rate imbalance (see figure F2). Furthermore, the inferred log-rates of model C allow us to obtain the complete distribution over cell states, which is not directly accessible by the experiments. As shown in figure 11, the cell state distribution depends on the initial cell conditions (T^--sorted/T^+-sorted stem cells) and the environmental conditions (CHIR/EPISC). The cell state distribution for initially unsorted stem cells (assuming every cell state is equally likely) is shown in appendix F.

So far we assumed cell transitions happen at a fixed rate θ_k which does not vary in time. We can test this assumption by using experimental data $\mathcal{D}_{CHIR}$ which recorded population snapshots at two different times $\Delta t = 2d$ and $\Delta t = 3d$. We used exclusively day 2 and day 3 data to obtain parameter posteriors separately for population dynamics at early times $0 < t < 2d$ and late times $2d < t < 3d$, respectively. The inferred reaction rates are compared in figure 12, which indicates:

- The universal reproduction rate θ_0 and marker rates θ_{S+}, θ_{S-} do not show a substantial time dependence and stay constant over the course of three days.
- The rate imbalance of marker T^+, θ_{T+}/θ_{T-}, significantly decreases at day 3 compared to day 1 and day 2 (difference $> 2\sigma$). The rate imbalance of marker F^+, θ_{F+}/θ_{F-}, seems to slightly increase with time, although the difference between day 2 and day 3 is not very substantial (difference $< \sigma$).
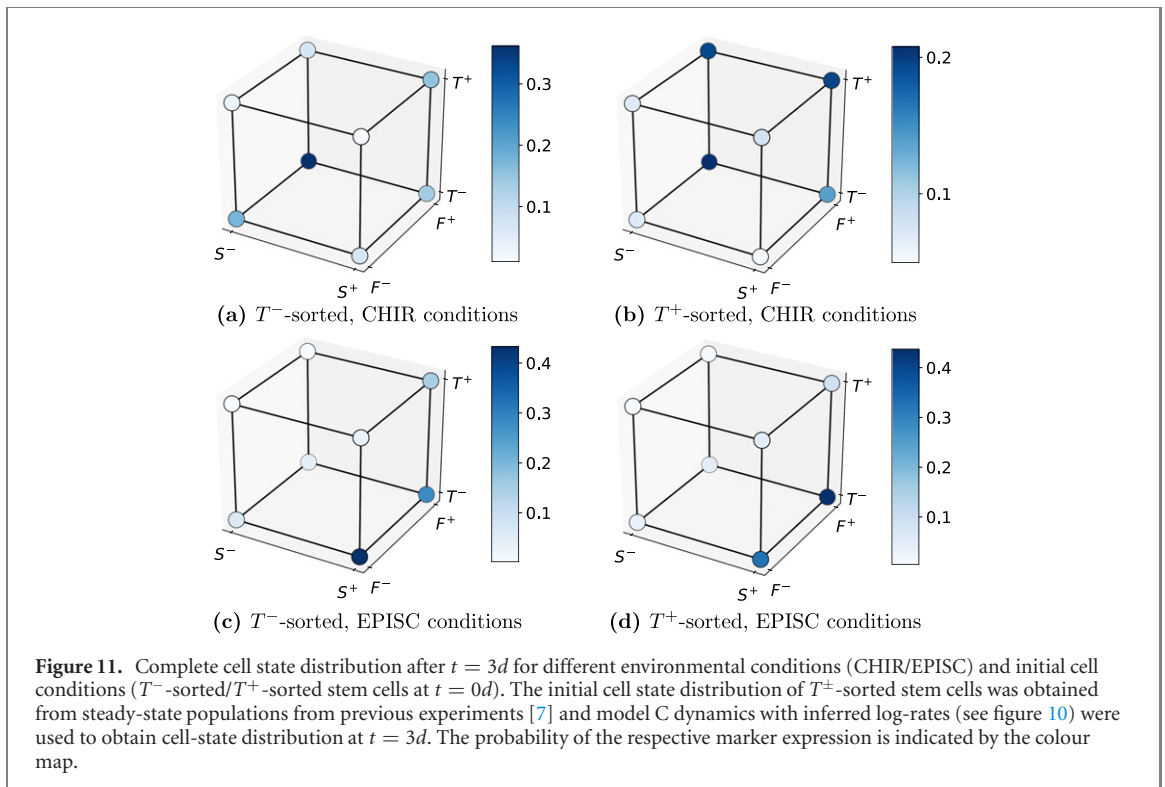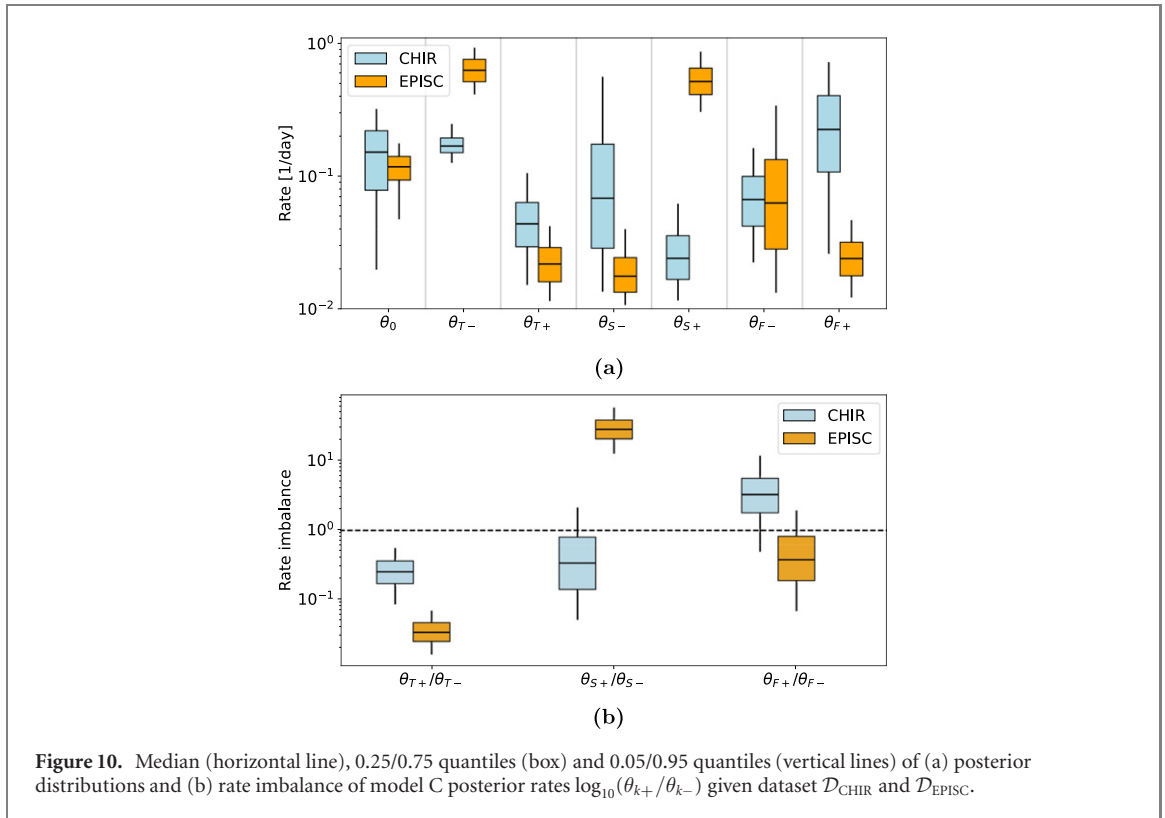
Again, the posterior distribution of model U shows a very similar time dependence of reaction rates (see figure F4).

4. Discussion

In this paper, we studied the population dynamics of two stochastic models of cell division and differentiation, or more generally, cell state changes: one in which cell state transitions are coupled to cell division, model C, defined by equation (2), and one in which these two processes are independent, model U, defined by equation (4). We solved the corresponding master equations analytically by using probability generating functions, which allow a systematic calculation of moments of the population distribution. The resulting solutions for both models C and U allowed us to analyse the dynamics of the population mean, based on which we showed that one cannot always differentiate between the two models. In model C, there is a finite probability that the proliferating sub-population will eventually die out and the dynamics will come to a halt. The probability of extinction (18) can be used to approximately infer the underlying reaction rate of model C solely based on the fraction of systems in which extinction was observed. Contrary, in model U both populations grow without bounds and extinction is impossible for finite reaction constants. This characteristic feature of extinction in model C might be useful to classify experimental observations or to infer cell division rates when there is negligible cell death.
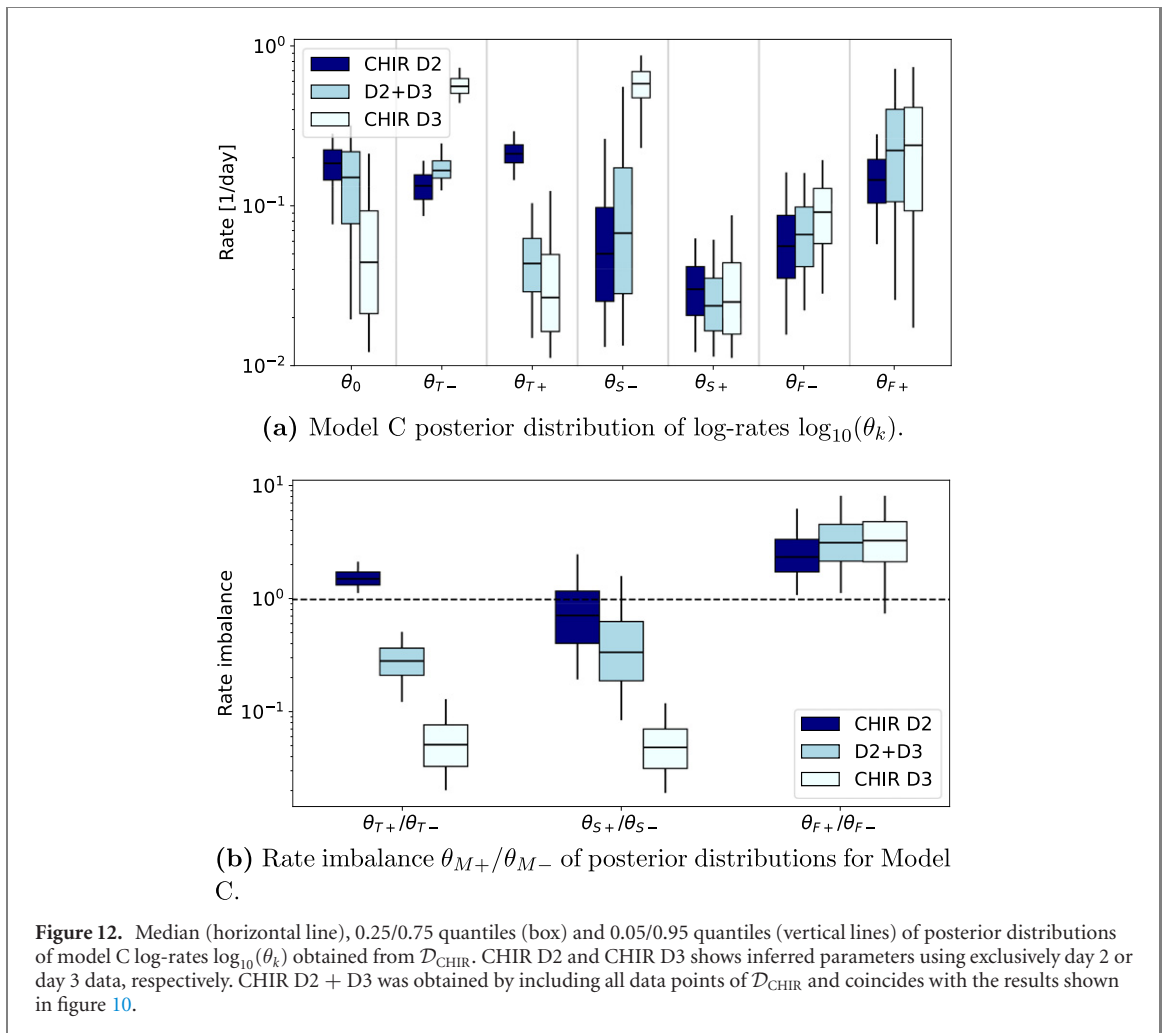
For Bayesian parameter inference we were able to obtain analytical expressions for the likelihood if continuous-time observations $[0, T]$ of all populations are available, e.g. as would be obtained from live imaging. Using conjugate priors, we obtained the posterior distributions analytically. When confronted with incomplete data in form of population snapshots, Bayesian inference is more challenging, and we have to rely on numerical results. As estimating the snapshot likelihood directly is computationally costly, we used a more elaborate, but efficient sample path method for obtaining posterior distributions of rate constants given snapshot-data. We further showed that the information content of snapshot-data depends on the times these snapshots were taken, and proposed an optimal spacing strategy which maximises the confidence of parameter inference.

For parameter and model inference based on experimental EpiSC data, we had to extend models C and U from two distinct cell populations to in total 8 cell states. Since our sample path method turned out to be computationally not feasible, we used approximate Bayesian computation (ABC) which speeds up calculation by several orders of magnitude. We used ABC rejection sampling and ABC-SMC to obtain posterior parameter distributions and Bayes factors for selecting between the two competing models C and

Figure 10. Median (horizontal line), 0.25/0.75 quantiles (box) and 0.05/0.95 quantiles (vertical lines) of (a) posterior distributions and (b) rate imbalance of model C posterior rates $\log_{10}(\theta_{k+}/\theta_{k-})$ given dataset $\mathcal{D}_{\text{CHIR}}$ and $\mathcal{D}_{\text{EPISC}}$.

Figure 11. Complete cell state distribution after $t = 3d$ for different environmental conditions (CHIR/EPISC) and initial cell conditions (T^--sorted/T^+-sorted stem cells at $t = 0d$). The initial cell state distribution of $T^\pm$-sorted stem cells was obtained from steady-state populations from previous experiments [7] and model C dynamics with inferred log-rates (see figure 10) were used to obtain cell-state distribution at $t = 3d$. The probability of the respective marker expression is indicated by the colour map.

U. The Bayes factors provide moderate to strong evidence in favour of model C in which cell division and differentiation are coupled (see figure 9). The posterior distribution of rate constants θ quantify how cell fate changes depending on the *in vitro* culture conditions (CHIR/EPISC, see figure 10). For cells in CHIR condition, we quantified how the rates of cell state

transitions might change over time (see figure 12). As we only have data from multiple time-points for the CHIR conditions, we cannot say whether this apparent time dependence is caused by a cell-intrinsic timing mechanism or whether the environment the cells experience changes over time because the cell colony is growing. Since cells kept under EPISC conditions

(a) Model C posterior distribution of log-rates $\log_{10}(\theta_k)$.

(b) Rate imbalance θ_{M+}/θ_{M-} of posterior distributions for Model C.

Figure 12. Median (horizontal line), 0.25/0.75 quantiles (box) and 0.05/0.95 quantiles (vertical lines) of posterior distributions of model C log-rates $\log_{10}(\theta_k)$ obtained from $\mathcal{D}_{\mathrm{CHIR}}$. CHIR D2 and CHIR D3 shows inferred parameters using exclusively day 2 or day 3 data, respectively. CHIR D2 + D3 was obtained by including all data points of $\mathcal{D}_{\mathrm{CHIR}}$ and coincides with the results shown in figure 10.

do not shift their expression of transcription factors over comparable timescales [7], it would seem more likely that these rates depend on the local environment and that cell-cell interactions are important, so that reaction rates change depending on the state of neighbouring cells. Interestingly, the data for cells under CHIR conditions indicates that cell division and differentiation are initially independent of each other (model U) up to day 2, after which cell division and differentiation appear to be coupled (model U, see figure F5). To account more rigorously for this time-dependent dynamics in systems which are not well-mixed or with strong cell–cell interactions, one would have to consider more complex models such as non-homogeneous Markov models or explicit representations of dynamic gene expression [20].

4.1. Limitations and alternative approaches

4.1.1. Assumption of Markovian dynamics

We began our investigation with minimal models of dynamics with two cell states. Although models C and U were motivated by biological arguments, the stochastic, two-dimensional jumping processes we defined could be applied to a variety of systems.

The assumption of Markovian dynamics is often used due to a range of useful properties which makes the models easier to analyse. Often, especially in the context of cell dynamics, the Markov assumption is at best a rough approximation [21, 22]. Waiting times τ between jumps are distributed according to an exponential distribution, which is biologically unrealistic: on a molecular level, many transport processes are required for a cell to replicate. Therefore, waiting times between two division events cannot be arbitrarily small [22].

Indeed, recent work has suggested that experimentally observed cell cycle time distributions differ substantially from a Markovian exponential distribution and are better described by the Erlang distribution [22]. How do population dynamics change for a non-Markovian process? In this scenario cell populations cannot grow arbitrarily fast, there is a time-dependent upper bound on population sizes $N < N_{\mathrm{lim}}(t)$. State probabilities $p_{i,j}(t) = p(N_{\mathrm{A}}(t) = i, N_{\mathrm{B}}(t) = j)$ are therefore zero for population numbers (i, j) exceeding limit $N_{\mathrm{lim}}(t)$. Cell population dynamics with Erlang distributed waiting times are non-Markovian stochastic processes with discrete states in continuous time. Non-Markovian processes like this can in general be analysed by converting them

into Markov processes by inclusion of supplementary variables [23].

4.1.2. Insight from minimal models

For Bayesian inference given snapshot data, our main focus was on building a Bayesian framework which can be applied to experimental data. However, the algorithms for parameter inference and model selection originally proposed for model C and U were computationally too intensive for the more complicated models with larger cell state networks. Nonetheless, the presented optimal snapshot spacing and scaling behaviour of Bayes factors with snapshot separation should generally be useful for experimental design. Measurement should be taken at times distributed non-uniformly according to equation (43). It is worth mentioning that equation (43) is only valid if all n snapshots are taken from the same system. If n snapshots were taken from independent systems, equation (43) cannot be applied and the posterior variance scales with snapshot separation Δt as shown in figure 6. Since the posterior of a set of independent trajectories is the product over posteriors obtained from individual trajectories, one should try to maximise Δt for every trajectory individually.

Here we have presented models for growing *in vitro* colonies of stem cells, for which neglecting cell death is a reasonable approximation. There are many biological systems in which cell death may be relevant, e.g. under homeostatic self-renewal of tissues, in which total cell numbers have to be kept constant, or even some applications in development where cell death is used for tissue size control [24, 25]. For these systems, the proposed models can be easily extended by including a cell death process of the form $A \to \emptyset$. Similarly, constraints on the population size can be implemented by introducing a cell division or death rate which depends on the total population size N, which acts as a crowding feedback mechanism [6]. This could represent contact inhibition or competition for niche access, for example. In spatial extensions of the models presented here, one could consider locally coupling division to differentiation and death events, as is relevant for example in epidermal homeostasis [26, 27]. Any such mechanisms can be easily added to the models and the same methods for inference and model selection can be applied. A disadvantage of these additional mechanisms is that the corresponding master equations would likely be intractable and one would have to rely solely on numerical results.

4.1.3. Sufficiency of summary statistics

We carried out parameter inference and model selection based on experimental EpiSC data using an ABC algorithm which works with summary statistics of the data. We obtained Bayes factors of $B_{CU} \approx 4.8$ for $\mathcal{D}_{CHIR}$ and $B_{CU} \approx 20$ for $\mathcal{D}_{EPISC}$ (see figure 9), which indicates moderate to strong evidence in favour of

model C. This result, however, has to be interpreted with care:

First, model selection based on Bayes factors depends on the prior distribution accurately reflecting the state of knowledge about the underlying model parameters $\boldsymbol{\theta}$. Here we chose log-uniform priors $\theta_k \sim \log(\mathcal{U}[-2, 0])$, which reflects not knowing the magnitude of the rate $\boldsymbol{\theta}$ (within bounds). The exact value of Bayes factors might therefore change slightly depending on the choice of prior.

Secondly, because we used summary statistics of the data instead of the full set of observations, we did not obtain the 'true' Bayes factor based on data $\mathcal{D}$, but one based on the summary statistic $s(\mathcal{D})$ of the data. Considering the full data would mean to look at each experimental replicate individually by calculating the parameter posterior given a single experimental realisation and iterate over all data points, using the posterior distribution as the prior distribution for the next data point. The 'true' Bayes factor and the one obtained using summary statistics only coincide if the summary statistic $s(\mathcal{D})$ is *sufficient* for comparing models C and U: $p(\mathcal{D}|s(\mathcal{D}), M_X) = p(\mathcal{D}|s(\mathcal{D}), M_Y)$. There are only few cases in which summary statistics are known to be sufficient, e.g. [28]. The more usual case is that we have to work with arbitrary summary statistics, which involve an unknown loss of information if they are insufficient. This 'curse of insufficiency' is the reason why some researchers caution against a naive use of ABC for model selection [29]. There are indeed three levels of approximation errors: one due to MC estimation, one due to the non-zero ABC tolerance $\epsilon > 0$, and one due to insufficient summary statistics. Alternatives to model selection with insufficient statistics were explored by [30], which selected models via a machine learning approach (random forests). An exhaustive investigation how the choice of summary statistics affects the shape of the posterior distributions is beyond the scope of this present work, but an interesting prospect for follow-on work.

Despite concerns about sufficiency of our summary statistics, we verified that the summary statistic used here is able to discriminate between extended models C and U. Using rejection-ABC, we obtained Bayes factors of $B_{CU} \approx 3.5$ on synthetic CHIR data $\mathcal{D}_{CHIR}^{sim}$ ($\epsilon/N_{set} = 3.75$) and $\langle B_{CU} \rangle \approx 3.7$ on synthetic EPISC data $\mathcal{D}_{EPISC}^{sim}$ ($\epsilon/N_{set} = 6$, note that one would likely achieve higher Bayes factors at lower ϵ with more efficient sampling schemes). Synthetic data was generated from model C with reaction rates set to posterior means obtained from experimental data (see figure 10). It is worth noting that for simulated data, the acceptance probability for a given ϵ value is much higher than for experimental data. This suggests that the models C and U do not perfectly describe the experimental data. More realistic models for cell state changes are therefore discussed in the following section.

4.1.4. Building models for single cell biology

Several approaches that aim to go beyond pseudotime ordering and connect single cell gene expression data to dynamic models have recently emerged. Examples include fitting drift-diffusion equations on a k-nearest neighbour graph [5] or along one-dimensional pseudotime trajectories [31], simulating cell trajectories in dimensionally reduced gene expression space based on RNA velocity estimates [32], and continuous state Markov process from single-cell time series [33]. Here, we have presented an alternative approach to these which allows for rigorous quantification of uncertainty. Working in a Bayesian framework further allows us to integrate multiple sources of data relatively easy, e.g. to combine bulk gene expression and single-cell data, which we discuss further below.

4.2. Future work

The models proposed in this work assumed that rates for switching genes on and off are independent of each other, thereby neglecting regulatory interactions between genes. A natural next step is to consider models with interdependencies between transcription factors, i.e., where the rate of transitioning into the on/off state for one transcription factor can depend on the expression state of another transcription factor. The number of possible pairwise dependencies alone leads to a model comparison problem that is computationally too intensive typical ABC inference approaches. However, systematic Bayesian model comparison can be performed on bulk population data, where likelihood evaluation is tractable, before parameterising the models on clonal data (Schumacher *et al*, in preparation). For Bayesian inference based on experimental EpiSC data, we had to extend models C and U from two distinct cell populations to in total 8 cell states. In order to reduce the number of free parameters, we only considered symmetric cell divisions, arguing that this should not significantly change population dynamics. Analysing model C we found that mean population dynamics are less sensitive to changes in the asymmetric cell division rate than to changes in the rate of symmetric cell division, which further supports our heuristic argument. A detailed investigation about how symmetric vs asymmetric cell divisions would affect population dynamics in the extended model, and how well the relative contribution of these two modes of division can be inferred from experimental data would be an interesting prospect for follow-on work.

Here we have have applied Bayesian inference and model comparison on *in vitro* data. It would be interesting to compare estimates of transition rates with *in vivo* case, where, in certain contexts, lineage tracing data can be used to identify clones. A potential application of our model comparison approach in this context would be to compare the number of stages in stem/progenitor cell lineages to best explain observed clone size distributions [34].

Contemporary experiments increasingly measure the expression of many more genes than considered here, for which one should consider models with a higher number of genes. As we showed in the last section, however, exact Bayesian inference is already computationally challenging for models considering only three genes. Since the number of cell states grows exponentially with the number of considered genes, one would run into computational scaling issues for methods such as ours. Thus, dimensionality reduction or careful consideration of the state space will be crucial to extend the inference of cell state transition rates to single-cell transcriptomic data.

The presented work focussed on total population sizes of cell states where individual cells are independent of each other, which implies a spatially homogeneous or well-mixed system. In many biological systems cell fate however depends on the local environment and signalling of cells. Here, there is potential to draw on and add to recent work inferring motility as well as proliferation in cell populations and tissue patterning [35–37]. The extension of these population models to spatially heterogeneous models including interactions between neighbouring cells, similar to [6], could thus be promising for inferring rates of differentiation in embryonic development.

ORCID iDs

Liam J Ruske ⓘ https://orcid.org/0000-0002-9731-1176
Jochen Kursawe ⓘ https://orcid.org/0000-0002-0314-9623
Anestis Tsakiridis ⓘ https://orcid.org/0000-0002-2184-2990
Valerie Wilson ⓘ https://orcid.org/0000-0003-4182-5159
Alexander G Fletcher ⓘ https://orcid.org/0000-0003-0525-4336
Richard A Blythe ⓘ https://orcid.org/0000-0001-8160-5068
Linus J Schumacher ⓘ https://orcid.org/0000-0003-0797-3406

References

[1] Casey M J, Stumpf P S and MacArthur B D 2019 Theory of cell fate *Wiley Interdiscip. Rev.: Syst. Biol. Med.* **12** e1471
[2] Morris S A 2019 The evolving concept of cell identity in the single cell era *Development* **146** dev169748

[3] Saelens W, Cannoodt R, Todorov H and Saeys Y 2019 A comparison of single-cell trajectory inference methods *Nat. Biotechnol.* **37** 547–54

[4] Tritschler S, Büttner M, Fischer D S, Lange M, Bergen V, Lickert H and Theis F J 2019 Concepts and limitations for learning developmental trajectories from single cell genomics *Development* **146** dev170506

[5] Weinreb C, Wolock S, Tusi B K, Socolovsky M and Klein A M 2018 Fundamental limits on dynamic inference from single-cell snapshots *Proc. Natl Acad. Sci. USA* **115** E2467–76

[6] Greulich P and Simons B D 2016 Dynamic heterogeneity as a strategy of stem cell self-renewal *Proc. Natl. Acad. Sci. USA* **113** 7509–14

[7] Tsakiridis A *et al* 2014 Distinct Wnt-driven primitive streak-like populations reflect *in vivo* lineage precursors *Development* **141** 1209–21

[8] Jefferys W H and Berger J O 1991 Sharpening Ockham's razor on a Bayesian strop Purdue University *Technical Report*

[9] Kass R E and Raftery A E 1995 Bayes factors *J. Am. Stat. Assoc.* **90** 773–95

[10] Courant R and Hilbert D 1937 *Methoden der mathematischen Physik: Zweiter Band* (Berlin: Julius Springer)

[11] Schnoerr D, Sanguinetti G and Grima R 2017 Approximation and inference methods for stochastic biochemical kinetics-a tutorial review *J. Phys. A: Math. Theor.* **50** 093001

[12] Wilkinson D J 2011 *Stochastic Modelling for Systems Biology* (Boca Raton, FL: CRC Press)

[13] Moscuoroums P G 1985 The distribution of the sum of independent gamma random variables *Ann. Inst. Statist. Math* **37** 541–4

[14] Kotz S, Balakrishnan N and Johnson N L 2004 *Continuous Multivariate Distributions, Volume 1: Models and Applications* vol 1 (New York: Wiley)

[15] Brooks S, Gelman A, Jones G and Meng X-L 2011 *Handbook of Markov Chain Monte Carlo* (Boca Raton, FL: CRC Press)

[16] Scott D W 2015 *Multivariate Density Estimation: Theory, Practice, and Visualization* (New York: Wiley)

[17] Toni T, Welch D, Strelkowa N, Ipsen A and Stumpf M P H 2008 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems *J. R. Soc., Interface* **6** 187–202

[18] Sisson S A, Fan Y and Tanaka M M 2007 Sequential Monte Carlo without likelihoods *Proc. Natl Acad. Sci. USA* **104** 1760–5

[19] Klinger E, Rickert D and Hasenauer J 2018 pyABC: distributed, likelihood-free inference *Bioinformatics* **34** 3591–3

[20] Manning C S *et al* 2019 Quantitative single-cell live imaging links HES5 dynamics with cell-state and fate in murine neurogenesis *Nat. Commun.* **10** 2835

[21] Stumpf P S *et al* 2017 Stem cell differentiation as a non-Markov stochastic process *Cell Syst* **5** 268–282.e7

[22] Gavagnin E, Ford M J, Mort R L, Rogers T and Yates C A 2018 The invasion speed of cell migration models with realistic cell cycle time distributions *J. Theor. Biol.* **481** 91–99

[23] Cox D R 1955 The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables *Mathematical Proceedings of the Cambridge Philosophical Society* vol 51 (Cambridge: Cambridge University Press) pp 433–41

[24] Kursawe J, Brodskiy P A, Zartman J J, Baker R E and Fletcher A G 2015 Capabilities and limitations of tissue size control through passive mechanical forces *PLoS Comput. Biol.* **11** e1004679

[25] Ambrosini A, Gracia M, Proag A, Rayer M, Monier B and Suzanne M 2017 Apoptotic forces in tissue morphogenesis *Mech. Dev.* **144** 33–42

[26] Rompolas P, Mesa K R, Kawaguchi K, Park S, Gonzalez D, Brown S, Boucher J, Klein A M and Greco V 2016 Spatiotemporal coordination of stem cell commitment during epidermal homeostasis *Science* **352** 1471–4

[27] Mesa K R, Kawaguchi K, Cockburn K, Gonzalez D, Boucher J, Xin T, Klein A M and Greco V 2018 Homeostatic epidermal stem cell self-renewal is driven by local differentiation *Cell Stem Cell* **23** 677–86

[28] Grelaud A, Robert C P and Marin J-M 2009 ABC methods for model choice in Gibbs random fields *C. R. Acad. Sci.* **347** 205–10

[29] Marin J-M, Pudlo P, Estoup A and Robert C P 2015 Likelihood-free model choice *Handbook of Approximate Bayesian Computation* (Boca Raton, FL: Chapman and Hall/CRC Press) pp 153–78

[30] Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M and Robert C P 2015 Reliable ABC model choice via random forests *Bioinformatics* **32** 859–66

[31] Fischer D S *et al* 2019 Inferring population dynamics from single-cell RNA-sequencing time series data *Nat. Biotechnol.* **37** 461–8

[32] Kimmel J C, Hwang A B, Scaramozza A, Marshall W F and Brack A S 2020 Aging induces aberrant state transition kinetics in murine muscle stem cells *Development* **147** dev.183855

[33] Taylor-King J P, Riseth A N, Macnair W and Claassen M 2020 Dynamic distribution decomposition for single-cell snapshot time series identifies subpopulations and trajectories during iPSC reprogramming *PLoS Comput. Biol.* **16** e1007491

[34] Picco N, Hippenmeyer S, Rodarte J, Streicher C, Molnár Z, Maini P K and Woolley T E 2019 A mathematical insight into cell labelling experiments for clonal analysis *J. Anat.* **235** 686–96

[35] Lambert B, Maclean A L, Fletcher A G, Coombes A N, Little M H, Byrne H M, Combes A N, Little M H and Byrne H M 2018 Bayesian inference of agent-based models: a tool for studying kidney branching morphogenesis *J. Math. Biol.* **76** 1673–97

[36] Guerrero P, Perez-Carrasco R, Zagorski M, Page D, Kicheva A, Briscoe J and Page K M 2019 Neuronal differentiation influences progenitor arrangement in the vertebrate neuroepithelium *Development* **146** dev176297

[37] Simpson M J, Baker R E, Vittadello S T and Maclaren O J 2020 Practical parameter identifiability for spatio-temporal models of cell invasion *J. R. Soc. Interface* **17** 20200055