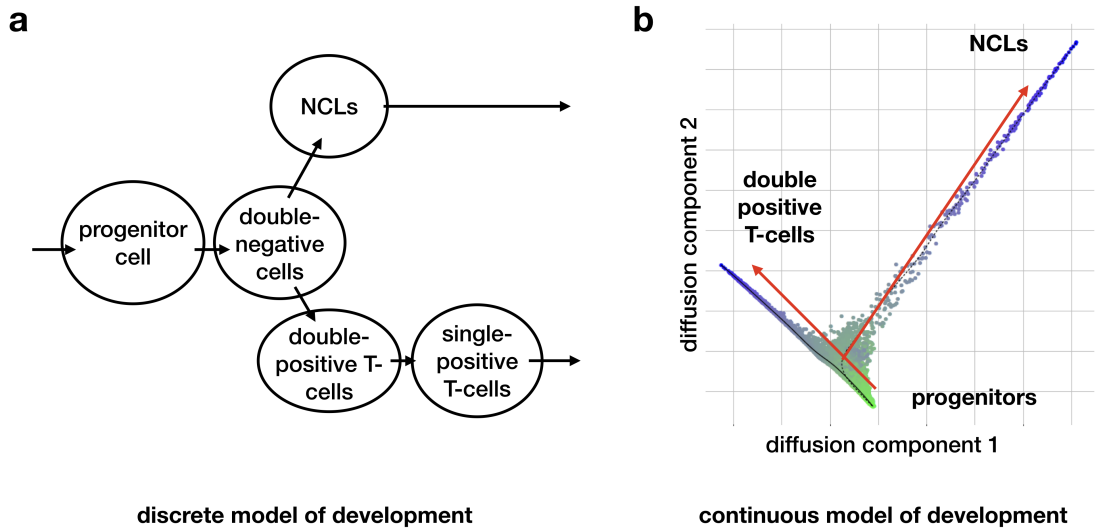
Supplementary Note 1: Pseudodynamics model

Contents



SN1 Figure 1: Illustration of the difference between a discrete and a continuous model of development. NCL: non-conventional lymphocyte.

SN1.1 Introduction to this Supplementary Note

This document is intended to give some background on parameter estimation for partial differential equation (PDE) models as well as an explanation of the input data, methods and results connected to the pseudodynamics models in more detail. We start with a motivation for continuous models of development (Sec. SN1.2). We go on to describe the pseudodynamics model in detail (Sec. SN1.3). We then describe forward simulation of the pseudodynamics model (Sec. SN1.5.1) which is necessary for parameter estimation (Sec. SN1.5). Subsequently, we describe how we integrated Rag2 knock-out Drop-seq samples into the pseudodynamics framework to map out beta-selection (Sec. SN1.6).

Cross-references to elements of the other Supplementary Notes are labels with "SNX" for Supplementary Note X.

SN1.2 Discrete versus continuous models of development

We stated in the main text that development has often been "modeled" based on discrete cell stages in cell biology, such as the progenitor, double-negative, double-positive and single-positive cell stages of T-cell maturation. We illustrated the advantages of continuous models with gene expression trajectories, heat maps and applications of the pseudodynamics model. Here, we provide another illustration of the conceptual difference between the two models (SN1 Figure 1). PDE models can be viewed as a limiting case of an ordinary differential equation model with infinitely many states. This limit consideration has no meaning with respect to the idea of cell biological models as these are typically based on surface markers and could not resolve enough states to reach this limit.

SN1.3 Pseudodynamics model

SN1.3.1 The pseudodynamics model for a non-branching developmental process

We want to model the progression of a population of single cells along a developmental trajectory in time. In this work, we define developmental progress as transcriptomic progress (cell state) and quantify cell state as pseudotime. Pseudotime is a dimension reduction from the space of all genes to a single dimension. Individual gene expression trajectories can therefore be mapped to cell state and accordingly also to the cell state-specific parameters of the model discussed below. Note that the dimensionality of the model can be increased by merging multiple one-dimensional trajectories into a branching model. One could also look at such dynamic models in higher dimensions, for example marker gene coordinates or diffusion components. We consider the cell density $u(s,t)$ for cell state s at a time point t. The integral over an interval $[s_1, s_2]$, $\int_{s_1}^{s_2} u(s,t)ds$, provides the number of cells with $s \in [s_1, s_2]$ at time point t. Assuming directed and random movement of cells as well as cell division and cell death, the temporal evolution of $u(s,t)$ can be described using a PDE. The PDE model includes diffusion with diffusion rate $D(s,t)$, drift (advection) with parameter $v(s,t)$ and a population growth term $g(s,t)$. In general those rates can depend on both, state and time.

$$\frac{\partial}{\partial t}u(s,t) = \underbrace{\frac{\partial}{\partial s}\left(D(s,t)\frac{\partial}{\partial s}u(s,t)\right)}_{\text{diffusion}} - \underbrace{\frac{\partial}{\partial s}\left(v(s,t)u(s,t)\right)}_{\text{drift}} + \underbrace{g(s,t)u(s,t)}_{\text{population growth}}. \tag{SN1.1}$$

The population size $N(t)$ is given by the integral of $u(s,t)$ with respect to the cell state s over the whole state domain $[0, s_{max}]$,

$$N(t) = \int_{s=0}^{s=s_{max}} u(s,t)ds. \tag{SN1.2}$$

The cell state-dependent parameters drift, diffusion and the cell state- or time-dependent growth rate are individually parameterized as natural cubic spline functions (ncs) of the cell state s with parameters α (Luzyanina, Roose, and Bocharov 2009). One can also add a parameter dependence on time t which we show here using the example of the birth-death rate.

$$\begin{aligned} v(s) &= \exp\left(\text{ncs}\left(s|\vec{\alpha}_v\right)\right) \\ D(s) &= \exp\left(\text{ncs}\left(s|\vec{\alpha}_D\right)\right) \\ g(s) &= \text{ncs}\left(s|\vec{\alpha}_{g,s}\right) \\ g(t) &= \text{ncs}\left(t|\vec{\alpha}_{g,t}\right) \\ g(s,t) &= g(s) * g(t). \end{aligned} \tag{SN1.3}$$

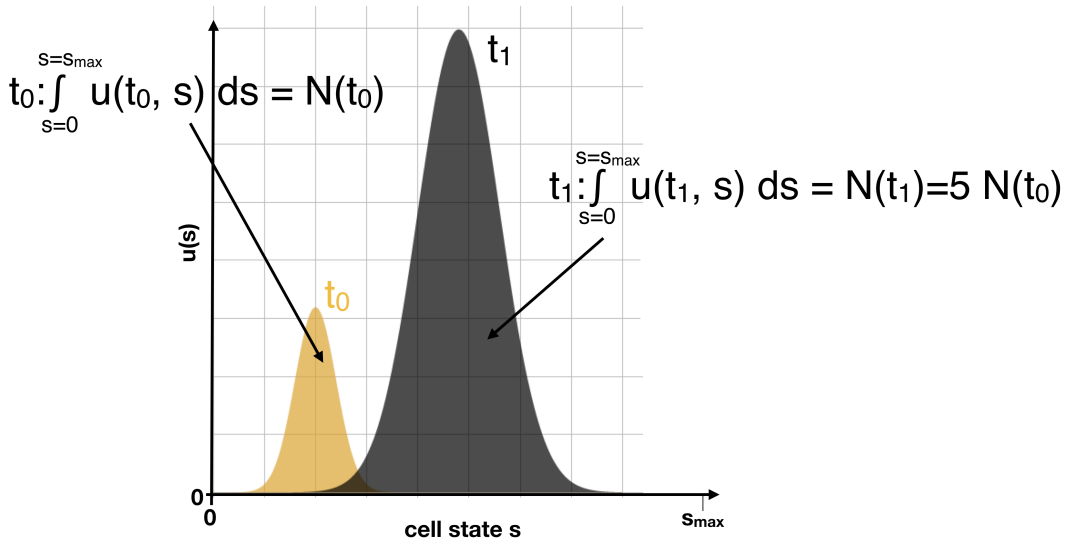
The vectors $\vec{\alpha}_v$, $\vec{\alpha}_D$, $\vec{\alpha}_{g,t}$ and $\vec{\alpha}_{g,s}$ parameterize the values of the spline at predefined nodes. The spline parameters of v and D are chosen in log-space and the exponential of the spline is computed to guarantee that drift and diffusion rates are positive.

In summary, the full state-dependent parameter set with a time-dependence of the birth-death rate of a non-branching pseudodynamics model that has to be estimated during fitting is:

$$\theta_{\text{non-branching}} = \{\vec{\alpha}_v, \vec{\alpha}_D, \vec{\alpha}_{g,s}, \vec{\alpha}_{g,t}\}. \tag{SN1.4}$$

SN1.3.1.1 Population size and growth

The initial condition $u(s,0)$ is the point-wise mean of the normalized kernel density estimates (kde) of the set of cell states (cells) S_0^r. S_0^r is observed in each sample r at the initial time point $t = 0$ and the kde is evaluated on a predefined grid and then normalized making it an initial



SN1 Figure 2: Pseudodynamics models the total population size at a certain time point via the integral of the simulated 1D density at this time point with respect to cell state. In the example shown here, the population grows by factor 5 from $N(t_0)$ at time $t_0 = 0$ to $5 \times N(t_0)$ at time $t_1 = 1$. Note that the population both increases in size and moves to a higher cell state (progresses in development).

probability distribution. The initial probability distribution is scaled by the mean of population size measurements $\bar{N}_0 = \frac{1}{|Z|}\sum_{z\in Z} N_0^z$ of all replicates z at time point zero to yield a number density. Note that the samples $r \in R$ are replicates of the cell state measurement (e.g. single-cell RNA-seq) and the samples $z \in Z$ are replicates of the population size measurements, which are independent experimental procedures.

$$u(s,0) = \frac{1}{|R|}\sum_{r\in R}\left(\mathrm{kde}\left(s|S_0^r\right)\right) * \bar{N}_0 \tag{SN1.5}$$

We provide an example in which the population is predicted to grow by factor 5 from $t = 0$ to $t = 1$ in SN1 Figure 2.

SN1.3.1.2 Boundary conditions

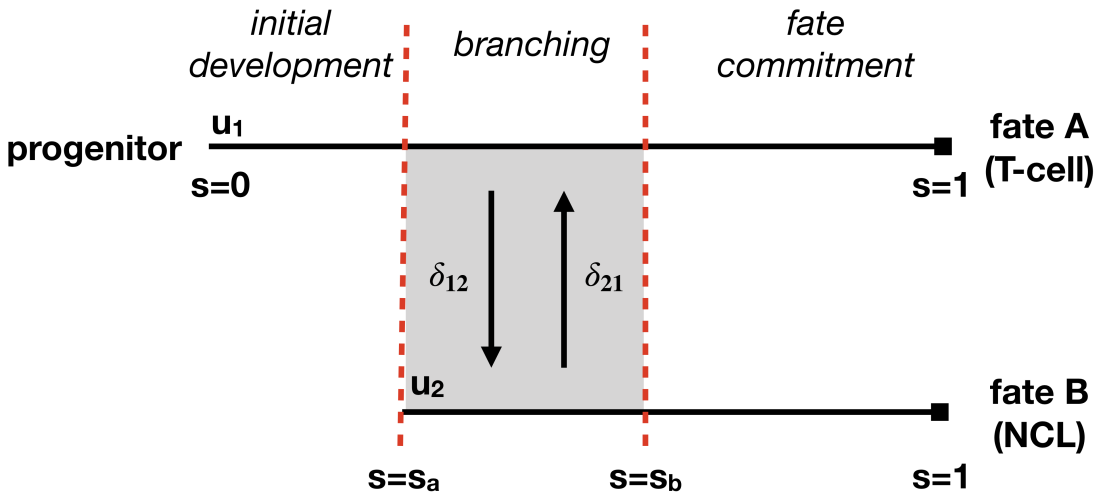
Depending on the biological hypothesis and data generation several boundary conditions could be possible. It seems reasonable to assume a no-flux boundary condition at $s = 0$ and $s = s_{max}$ for most applications. This corresponds to a Robin boundary condition at $s = 0$:

$$\left(D(s)\frac{\partial u}{\partial s}(s,t) - v(s)u(s,t)\right)\Big|_{s=0} = 0 \tag{SN1.6}$$

At the right boundary, $s = s_{max}$, we assume that the cells are differentiated and do not advance further in the cell state leading to zero drift and the Robin boundary condition simplifies to:

$$\frac{\partial}{\partial s}u(s,t)\Big|_{s=s_{max}} = 0 \tag{SN1.7}$$

The assumption of zero drift at the right hand side improves numerical properties of the system. To ensure no drift at the right-hand boundary without restraining variability in the model, we linearly



SN1 Figure 3: Illustration of the pseudodynamics model for a branching process from progenitor to fate A or fate B. NCL: non-conventional lymphocyte. Variables are named as in (SN1.8).

scaled the observations to the interval $[0, 0.9]$ and ran simulations on the interval $[0, 1]$. On the interval [0.9,1] the drift can be reduced to 0 by a smooth spline. Values in this interval correspond to developmental states which were not observed. Note that pseudotime is on an arbitrary scale and the ordering is invariant to scaling so that linear scaling does not change the biological meaning of the coordinate.

SN1.3.2 The pseudodynamics model for a branching developmental process

Branching is the split of a developmental trajectory into two trajectories which correspond to lineages with separate cellular fates and a common precursor cell state (SN1 Figure 1). Therefore, a branching developmental process is characterized by an initial state (a root cell), two final states (terminal fates, tip cells), a branching point or branching region in which the split occurs and three regions (root to branching region, branching region to fate 1, branching region to fate 2) in which the system is governed by (SN1.1). We split such a branching process into a main branch (root via branching region to fate 1) and a side branch (branching region to fate 2) which overlap at a branching region (SN1 Figure 3).

As before, the progress coordinate (cell state) is pseudotime computed on the full data set. This pseudotime coordinate is still one dimensional, cells are divided into branches based on other criteria (such as using diffusion component coordinates) but still carry these pseudotime coordinates which are globally assigned to all cells from all branches (Haghverdi et al. 2016).

To account for branching trajectories we couple two PDEs which describe progression along the main branch and the side branch of the developmental process (SN1 Figure 3). Note that those two PDEs represent two processes in one state dimension, $s \in \mathbb{R}$. We define a branching region on each branch to couple those two processes. The branching region is an interval $[s_a, s_b]$ in cell state in which cells can switch from branch i to branch j with a propensity δ_{ij}. Denoting the main

branch with u_1 and the side branch with u_2 we can extend the non-branching model (SN1.1),

$$\begin{aligned}\frac{\partial}{\partial t}u_1(s,t) &= \frac{\partial}{\partial s}\left(D_{b1}(s)\frac{\partial}{\partial s}u_1(s,t)\right) - \frac{\partial}{\partial s}\left(v_{b1}(s)u_1(s,t)\right) + g_{b1}(s)u_1(s,t) \\ &\quad - T(s)\Big(\delta_{12}u_1(s,t) - \delta_{21}u_2(s,t)\Big), \\ \frac{\partial}{\partial t}u_2(s,t) &= \frac{\partial}{\partial s}\cdot\left(D_{b2}(s)\frac{\partial}{\partial s}u_2(s,t)\right) - \frac{\partial}{\partial s}\left(v_{b2}(s)u_2(s,t)\right) + g_{b2}(s)u_2(s,t) \\ &\quad + T(s)\Big(\delta_{12}u_1(s,t) - \delta_{21}u_2(s,t)\Big),\end{aligned} \tag{SN1.8}$$

in which the function $T(s)$ defines the branching interval. T(s) is one in the interval in cell state space (pseudotime coordinate) between s_a (start of the branching interval in pseudotime) and s_b (end of the branching interval in pseudotime) and zero otherwise:

$$T(s) = \begin{cases}1, & \text{if } s \in [s_a, s_b] \\ 0, & \text{otherwise.}\end{cases} \tag{SN1.9}$$

Diffusion, drift and growth rates can differ between branches and are hence indexed. Further, we dropped the dependence of the rates on time as in this manuscript only state-dependent rates are considered for the model with branching. We again use boundary conditions as for the non-branching system:

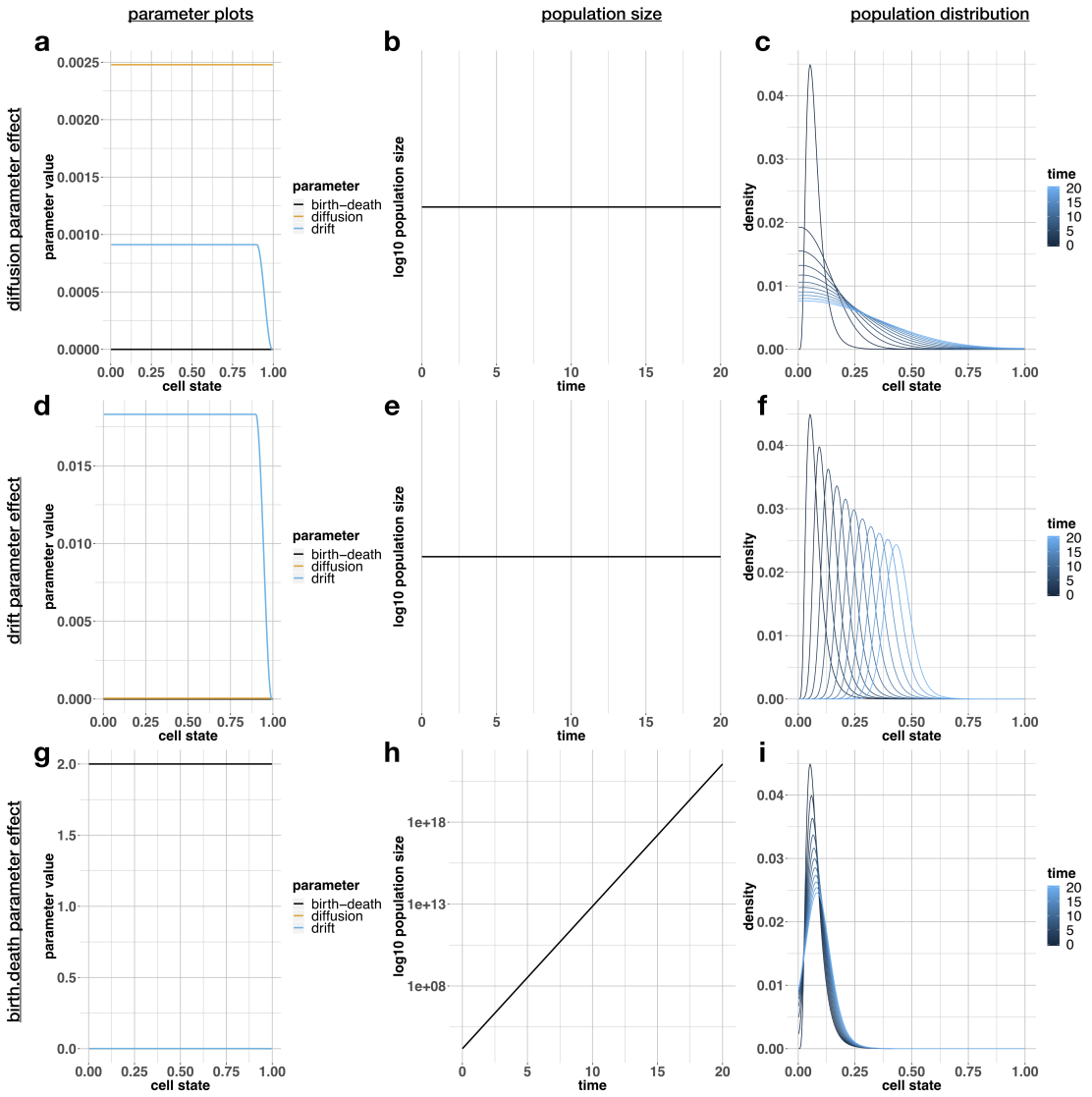
$$\begin{aligned}&\left(D_{b1}(s)\frac{\partial u_1}{\partial s}(s,t) - v_{b1}(s)u_1(s,t)\right)\Big|_{s=0} = 0 \\ &\left(D_{b2}(s)\frac{\partial u_2}{\partial s}(s,t) - v_{b2}(s)u_2(s,t)\right)\Big|_{s=s_a} = 0 \\ &\frac{\partial}{\partial s}u_1(s,t)\Big|_{s=s_{max}} = 0 \\ &\frac{\partial}{\partial s}u_2(s,t)\Big|_{s=s_{max}} = 0.\end{aligned} \tag{SN1.10}$$

These branching models can be designed with arbitrarily many branches and branching regions. The initial condition on each branch is computed as in Sec. SN1.3.1.1 but is additionally scaled with the fraction of cells measured on the respective branch in the single cell experiment,

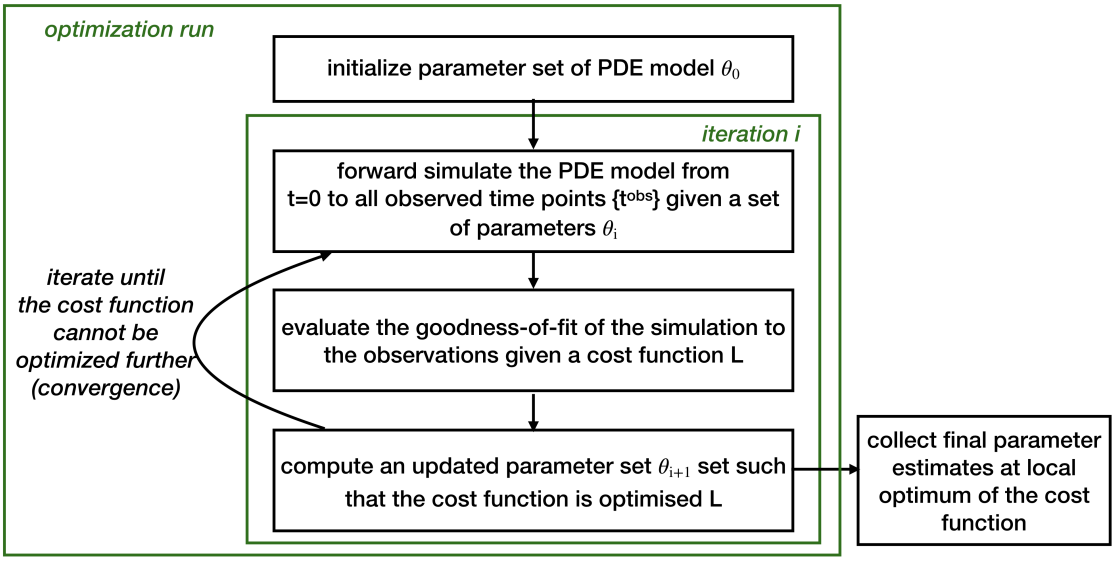
$$u_b(s,0) = \frac{1}{|R|}\sum_{r\in R}\left(\mathrm{kde}\left(s|S^r_{b,0}\right)\right) * w_{b,0} * \bar{N}_0. \tag{SN1.11}$$

SN1.3.3 Illustration of the effects of diffusion, drift and birth-death rate on the time-dependent population distribution

We designed three example parameterizations to illustrate the effects of the three pseudodynamics model parameters diffusion (SN1 Figure 4a-c), drift (SN1 Figure 4d-f) and birth-death rate (SN1 Figure 4g-i). The diffusion parameter models population heterogeneity in development. A high diffusion parameter therefore causes the variance of the population distribution in cell state space to increase over time (SN1 Figure 4c) while it does not affect the total population size (SN1 Figure 4b). The drift parameter models directed components in development. A high drift parameter therefore causes the mean of the population distribution in cell state space to change over time (SN1 Figure 4f) while it does not affect the total population size (SN1 Figure 4e). The birth-death parameter models changes in the population size during development. A constant birth-death parameter in cell state therefore does not influence the normalized population distribution in cell state space over time (SN1 Figure 4i) but it does affects the total population size (SN1 Figure 4h).



SN1 Figure 4: Influence of the pseudo dynamics model parameters on the population distribution across cell state and population size as functions of time. (**a,b,c**) Example to illustrate the effect of the diffusion parameter. (**d,e,f**) Example to illustrate the effect of the drift parameter. (**g,h,i**) Example to illustrate the effect of the birth-death parameter. (**a,c,g**) Parameters which are used to forward simulate the system. The parameters are functions of cell state. (**b,d,h**) Simulated population size as function of time. (**c,e,i**) Simulated population distribution across cell state as function of time.



SN1 Figure 5: Algorithm for parameter estimation on a PDE model. Green boxes: loops. Continue to SN1 Figure 6 for the methods of evaluation of the parameter estimates.

SN1.4 Forward simulation of the pseudodynamics model: Finite differences and finite volumes

The pseudodynamics model is a system of PDEs and can be simulated using different numerical methods. A first implementation using finite differences failed to conserve mass in the absence of population growth terms ($g(s) = g(t) = 0$). A finite volumes implementation proved to be numerically suitable for the pseudodynamics model both with and without branching. However, finite volumes only simulate average densities in a grid volume and not density values at a grid point like finite differences. Accordingly, we introduced the likelihood function presented in Sec. SN1.5.2 to allow parameter estimation.

SN1.5 Parameter estimation

SN1.5.1 Parameter estimation on PDE models

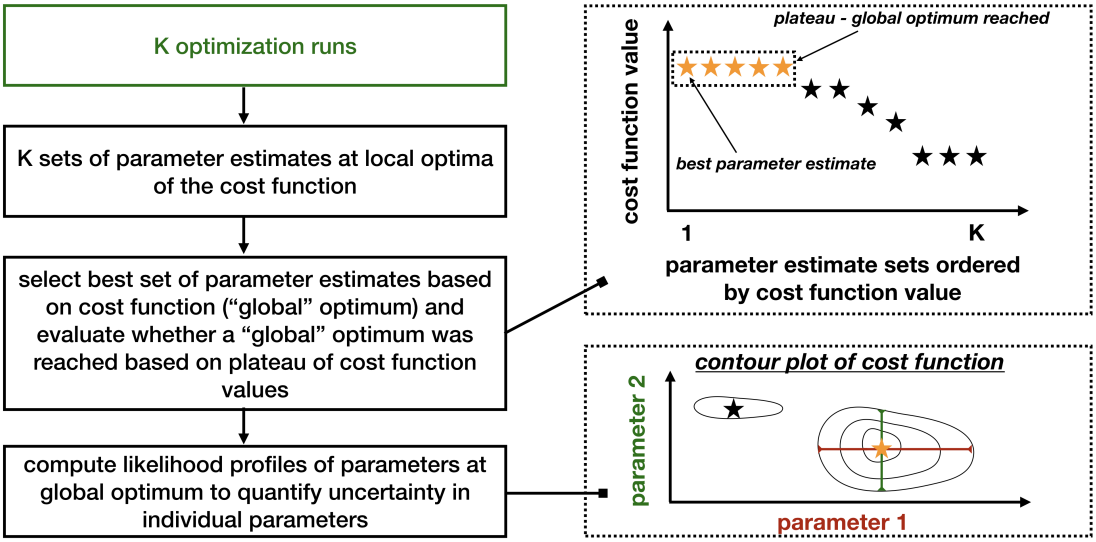
In this section we outline the parameter estimation for the considered system of PDEs. Note that the described workflows have already been established in the field of parameter estimation of PDE models (Stapor et al. 2018; Tarantola 2005; Hross and Hasenauer 2016).

There is no closed-form parameter estimate available for the pseudodynamics model. Therefore, parameter estimation has to be done iteratively. Parameter estimation for PDEs requires four core components:

The first component is a method to simulate the PDE for a proposed parameterization θ (SN1 Figure 5) (see Sec.).

The second component is a likelihood that evaluates the goodness-of-fit of the simulation to the measurements (SN1 Figure 5). The likelihood might include regularization terms if necessary.

The third component is a method to propose a better parameter set (update) given a current set in the iteration (SN1 Figure 5). Such parameter update methods are usually based on derivatives of the likelihood with respect to the parameters and thereby yield parameter updates that optimize the likelihood. We chose the Matlab optimizer fmincon with either the interior point or the trust



SN1 Figure 6: Method to evaluate convergence of estimation to global optimum of the regularized likelihood and to quantify uncertainty in the parameter estimates at the global optimum. Yellow stars: parameter estimates at global optimum of regularized likelihood. Black stars: parameter estimate at local optima of regularized likelihood. Red (green) lines in contour plot: Projection of confidence interval of parameter 1 (parameter 2) onto corresponding axis. Refer to SN1 Figure 5 for the details on the individual optimization runs in the green box.

region algorithm implemented in fmincon. The choice of algorithm by application is recorded in the respective application sections in Sec. SN3.2 and SN2.2. To ensure the identification of the global optimum we performed multi-start local optimization.

The fourth component is a method to evaluate the confidence in the parameter estimates (SN1 Figure 6). Here, we computed likelihood profiles or local approximations of the likelihood at the best parameter estimates to quantify the uncertainty of the parameter estimates.

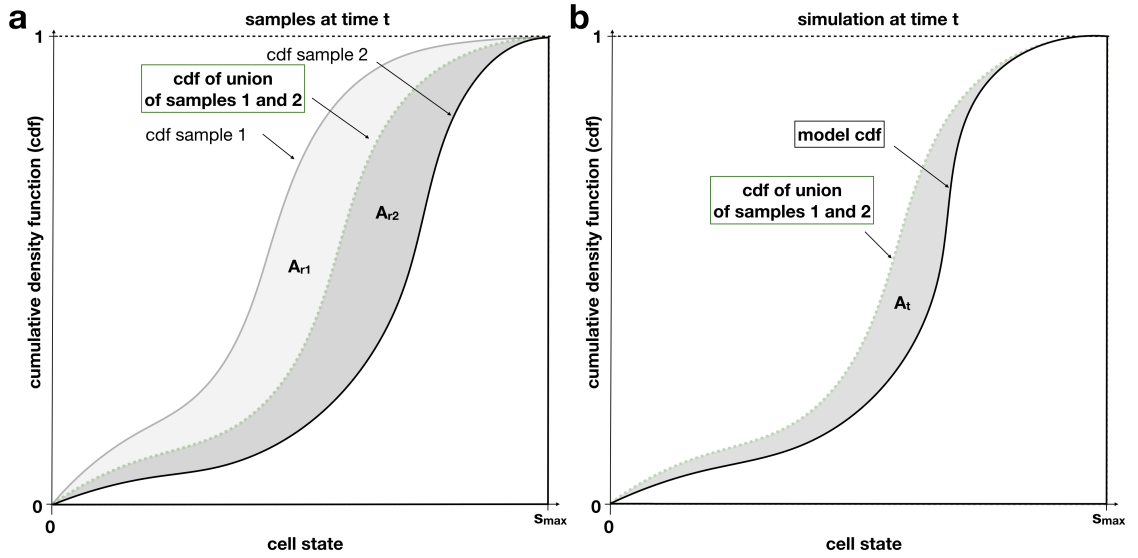
We give a detailed description of each of these steps in this section.

SN1.5.2 Likelihood function

We used a likelihood function to evaluate the fit of our model to the data. To construct an appropriate likelihood we considered three contributes. The first describes the fit of the simulated cell densities to the measured cell densities. The second describes the fit of the population size. The fraction of cells on the individual branches is included in the third part. For numerical reasons we minimized the negative log-likelihood. As we assumed independence of the contributing measurement types, the log-likelihood consists of a sum of the three terms. We describe in the following sections how to compute these individual terms. The combined log-likelihood function is given by

$$\begin{aligned}\log L(\theta) = &\left(\sum_{b\in B} \sum_{t\in T^{cdf}} \log L\left(\mathrm{ecdf}_{S_{b,t}}(s)|\theta\right) \right) + \left(\sum_{t\in T^{N}} \log L\left(\bar{N}_t|\theta, \sigma_t^N\right) \right) \\ &+ \left(\sum_{b\in B\setminus b_{max}} \sum_{t\in T^{cdf}} \log L\left(w_{b,t}|\theta, \sigma_{b,t}^w\right) \right)\end{aligned} \tag{SN1.12}$$

where B is the set of branches modeled, $w_{b,t}$ is the mean observed proportion of cells on branch



SN1 Figure 7: Likelihood for the fit of the simulated population density across cell state to the observed densities. (**a**) The parameters of the normal likelihood at a given time point, mean and variance of the area between the curves, are estimated based on the area between the empirical cumulative density functions (ecdf) of the samples and the ecdf of the union of all samples (dotted green line) on a single branch b. The case corresponding to two replicates is illustrated. The ecdfs of two samples, $\mathrm{ecdf}_{S_b^1}$ and $\mathrm{ecdf}_{S_b^2}$ and the ecdf of the union of the samples, ecdf_{S_b}, are indicated. The area between the ecdf of sample i and the ecdf of the union is denoted by $A_{ri} = \mathcal{A}(\mathrm{ecdf}_{S_b^i}, \mathrm{ecdf}_{S_b})$. The mean and variance of values, A_{r1} and A_{r2} (shaded areas). (**b**) The population density fit terms of the likelihood of a parameter set of the pseudodynamics model given a sample is evaluated as the probability of observing an area between the simulated cdf and the ecdf of the union of all samples (dotted green line) given the parameterization of the normal likelihood inferred in (**a**).

b compared to the total population size, $\sigma_{b,t}^w$ is the standard error of the mean of the proportion of cells in branch b at time t, T^{cdf} is the set of time points at which the population density was observed, $S_{b,t} = \bigcup_{r \in R} S_{b,t}^r$ is the vector of cell state observations at time point t in branch b (note that $S_{b,t}$ is the union of the sets of each replicate $r \in R$ of the cell state observations) and $\mathrm{ecdf}_{S_{b,t}}(s)$ its empirical cumulative density function, T^N is the sets of time points at which the population size was observed, $\bar{N}_t = \frac{1}{|Z|} \sum_{z \in Z} N_t^z$ is the mean observed population size at time point t (across replicates Z of the population size observations) and θ is the set of parameters of the pseudodynamics model and σ_t^N is the standard error of the mean of the observed population size at time t. Note that the likelihood term on the fraction of cells per branch was evaluated on all branches except one as the proportions across all branches sum up to one.

SN1.5.2.1 Population distribution across cell state terms of the likelihood

In the numerical implementation using finite volumes, the average concentration in a grid interval, $\frac{1}{s_{i+1}-s_i} \int_{s_i}^{s_{i+1}} u(s,t)ds$, is simulated rather than the concentration itself. Therefore, it is not possible to assess the probability of finding a cell at that cell state, $p(s,t)$, directly, but only the probability of finding a cell in the considered interval $[s_i, s_{i+1})$. Hence, we formulated the log-likelihood function based on the cumulative density function (cdf). This log-likelihood also incorporates replicates. Firstly, all replicates were pooled and the area between the pooled cdf and the

cdf of individual replicates was computed (SN1 Figure 7a). One can now determine the mean and the standard deviation of the area between the curves. We assumed that the area between the cdf of the pooled sample and the simulated distribution (SN1 Figure 7b) is normally distributed with the computed mean and standard error of mean. The following likelihood was computed for each branch separately. The deviation of the simulated density from the observed density is quantified by the area between the cumulative density function of the simulated density ($\mathrm{cdf}_{u_b}(s,t)$):

$$\mathrm{cdf}_{u_b}(s,t) = \frac{\int_0^s u_b(\bar{s},t)d\bar{s}}{\int_0^{s_{max}} u_b(\bar{s},t)d\bar{s}} \tag{SN1.13}$$

and the empirical cumulative density function of the union of all samples $S_{b,t}$ on branch b at time point t ($\mathrm{ecdf}_{S_{b,t}}(s)$) (SN1 Figure 7b):

$$\mathrm{ecdf}_{S_{b,t}}(s) = \frac{1}{|S_{b,t}|}\sum_{s'\in S_{b,t}} \mathbb{1}_{s'\leq s} \tag{SN1.14}$$

We assumed that the area between the curves ($\mathcal{A}(\mathrm{cdf}_1(s),\mathrm{cdf}_2(s)) = \int_0^{s_{max}} |\mathrm{cdf}_1(s) - \mathrm{cdf}_2(s)|ds$, the L1-norm of the two cdfs) is normally distributed with mean ($\mu^{\mathcal{A}}(t)$) and standard deviation ($\sigma^{\mathcal{A}}(t)$) computed based on the area between the ecdf of the replicate r of a given time point t ($\mathrm{ecdf}_{S^r_{b,t}}$) and the ecdf of the union of all samples at that time point ($\mathrm{ecdf}_{S_{b,t}}$) (SN1 Figure 7a).

$$\log L(\mathrm{ecdf}_{S_{b,t}}|\theta) = \mathcal{N}\left(\frac{1}{|R_t|}\sum_{r\in R_t}\mathcal{A}\left(\mathrm{ecdf}_{S^r_{b,t}}(s),\mathrm{ecdf}_{S_{b,t}}(s)\right)\,\middle|\,\mu=\mu_b^{\mathcal{A}}(t),\sigma^2=(\sigma_b^{\mathcal{A}}(t))^2\right) \tag{SN1.15}$$

$$\begin{aligned}\mu_b^{\mathcal{A}}(t) &= \mathcal{A}\left(\mathrm{cdf}_{u_b}(s,t),\mathrm{ecdf}_{S_{b,t}}(s)\right)\\ &= \int_0^{s_{max}} \left|\mathrm{cdf}_{u_b}(\bar{s},t) - \mathrm{ecdf}_{S_{b,t}}(\bar{s})\right|\,d\bar{s}\end{aligned} \tag{SN1.16}$$

$$(\sigma_b^{\mathcal{A}}(t))^2 = \frac{1}{|R_t|}\sum_{r\in R_t}\left(\mathcal{A}\left(\mathrm{ecdf}_{S^r_{b,t}}(s),\mathrm{ecdf}_{S_{b,t}}(s)\right) - \frac{1}{|R_t|}\sum_{r\in R_t}\mathcal{A}\left(\mathrm{ecdf}_{S^r_{b,t}}(s),\mathrm{ecdf}_{S_{b,t}}(s)\right)\right)^2 \tag{SN1.17}$$

where R_t is the set of replicates at time point t.

However, in Sec. SN3.2.3 we used a log-likelihood (SN1.15) by estimating the standard deviation across measurement time points together with the other parameters as there was only one replicate per time point. The estimate for the variance is independent of time and given by,

$$(\sigma_b^A)^2 = \frac{1}{|T^{\mathrm{cdf}}|}\sum_{t\in T^{\mathrm{cdf}}}\mathcal{A}\left(\mathrm{ecdf}_{S_{b,t}}(s),\mathrm{cdf}_{u_b}(s,t)\right)^2. \tag{SN1.18}$$

If no replicates were measured, we employed a least-squares objective function instead of eq . (SN1.15),

$$\log L(\mathrm{ecdf}_{S_{b,t}}|\theta) = \mathcal{A}\left(\mathrm{ecdf}_{S_{b,t}}(s),\mathrm{cdf}_{u_b}(s,t)\right)^2. \tag{SN1.19}$$

SN1.5.2.2 Population size terms of the likelihood

Firstly, note that the population size measurements and the population density in cell state estimates derive from independent samples in our experimental set up. Also note that the population size observations are not linked to branches as we only observe the total number of cells across all branches.

We assumed normally distributed measurement noise in the population size samples with standard deviation σ_t^N and mean zero. We approximated σ_t^N with the standard error of the mean of the observations at time point t:

$$\sigma_t^N = \sigma_t^{N,obs} / \sqrt{n_t^N} \tag{SN1.20}$$

where $\sigma_t^{N,obs}$ is the observed standard deviation of the samples at time point t and n_t^N the number of population size samples at time point t. Hence, the likelihood of the population size observations given the model is given by

$$\log L(\bar{N}_t|\theta, \sigma_t^N) = \mathcal{N}\left(\bar{N}_t \middle| N\left(t, \theta\right), \sigma^2 = \left(\sigma_t^N\right)^2\right) \tag{SN1.21}$$

and $N(t, \theta)$ is the population size predicted by the model for the parameters θ at time point t across all branches:

$$N(t, \theta) = \sum_{b \in B} \int_0^{s_{\max}} u_b(s, t) ds \tag{SN1.22}$$

SN1.5.2.3 Branch weight terms of the likelihood

We calculated the relative number of cells $w_{b,t}$ at each time point t on each branch b using the cell state distribution observations. To include this proportionality information in the estimation, we assumed normally distributed measurement noise on the proportion observations with standard deviation $\sigma_{b,t}^w$ and mean zero. $\sigma_{b,t}^w$ is approximated by the standard error of the mean of the observations at time point t,

$$\sigma_{b,t}^w = \sigma_{b,t}^{w,obs} / \sqrt{R_t} \tag{SN1.23}$$

where $\sigma_{b,t}^{w,obs}$ is the observed standard deviation of the samples at time point t and R_t the number of replicates of cell state distribution observations at time point t and branch b. Hence, the likelihood of the proportion of cells on branch 1 given the model is given by

$$\log L(w_{b,t}|\theta, \sigma_{b,t}^w) = \mathcal{N}\left(w_{t,r} \middle| \mu = \mu_b^w(t, \theta), \sigma^2 = (\sigma_{b,t}^w)^2)\right) \tag{SN1.24}$$

$\mu_b^w(t, \theta)$ is the proportion predicted by the model for the parameters θ at time point t on branch b:

$$\mu_b^w(t, \theta) = \frac{\int_0^{s_{\max}} u_b(s, t) ds}{\sum_{\bar{b} \in B} \int_0^{s_{\max}} u_{\bar{b}}(s, t) ds} \tag{SN1.25}$$

The above calculations were performed for all branches $b \in B \setminus b_{max}$. If we consider a non-branching developmental process SN1.1 this part is obviously not considered in the likelihood.

SN1.5.2.4 Notes regarding the likelihood functions

Error model for the fraction of cells on each branch If the number of cells is binomial distributed between the two branches, the resulting fraction on a branch is approximately normally distributed. This follows from the central limit theorem (e.g. from the central limit theorem for

Bernoulli-sequencesGeorgii 2009, p. 138). In general for each replicate the fraction of cells on a branch would be a sample from a normal distribution with the true success probability as expectation. Depending of the sample size, the standard deviation would vary. Here, we decided to estimate the effective standard deviation from the data. This ensures consistency with the other parts of the likelihood and allow us to capture other sources of measurement error than the sampling error modeled by the binomial distribution. In the case of more than two branches it remains to be investigated what likelihood is preferable.

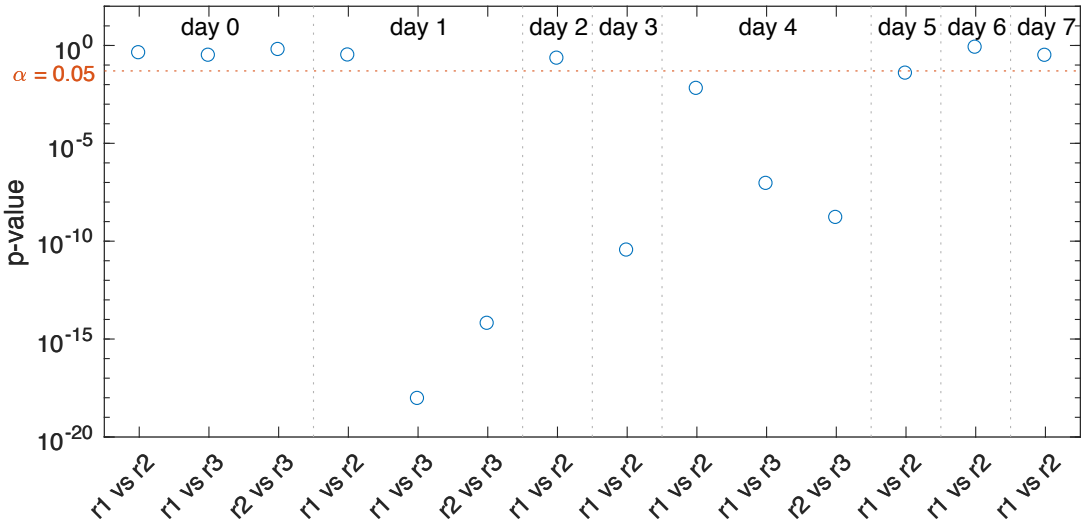
Error model for the distribution of cells We used error models that are based on a normally distributed area between an observed ecdf and a predicted cdf (Sec. SN1.5.2.1). One could also model the distance between two distributions as the Kolmogorov-Smirnov (K-S) distance. Here, we discuss usage of the K-S distance as an error model for the population distribution.

K-S tests on the samples of the T-cell development data suggest the replicates do not all come from the same distribution (Figure SN1 Figure 8). More precisely, we tested all replicates from the same time point using a two sample K-S test. Of the eleven combinations, seven tests rejected the hypothesis that the replicates follow the same distribution at a significance level of $\alpha = 0.05$. However, the K-S test is too sensitive in many cases: In the context of immunofluorescence histogram comparison the K-S test was shown to be overly sensitive (Lampariello 2000). To confirm that a likelihood based on the K-S distance can not capture the true variability in the data, we generated a distribution of K-S distances that would be expected if the measured data was generated by sampling from the same underlying distribution. To this end we generated a subsample from the distribution of the pooled data corresponding to a day and evaluated the K-S distance and the area between the sampled and the pooled cdf for 1000 subsamples. As the number of measured single cells varied between replicates, we tried to recreate this in the generated data by sampling the sizes of the 1000 subsamples from a distribution with the same first two moments as the measured samples. For the days one to eight, we sampled from a normal distribution with the same mean and variance as the single cell sample sizes. For day zero, the subsample size was sampled from a log-normal distribution with the same mean and variance as the single cell sample sizes. We compared a the distance of the measured samples to the pooled sample with b) the distance of the subsamples and to the pooled sample. We found that the distances of the measured samples are outliers with respect to the distribution of distances generated by subsampling in several cases. This indicates, that the measurements are probably not samples from the pooled distribution and that there is additional confounding variation in the samples. Hence, the K-S distance or the area between curves do not capture the full variability between replicates (SN1 Figure 9). Thus, we decided to use a normal assumption, as more fitting statistics are not yet available for the distribution of single cells a long the cell state trajectory. This error model does not stem from an generative statistical model. The underlying statistics might need to be studied more in further experiments to develop such a generative model.

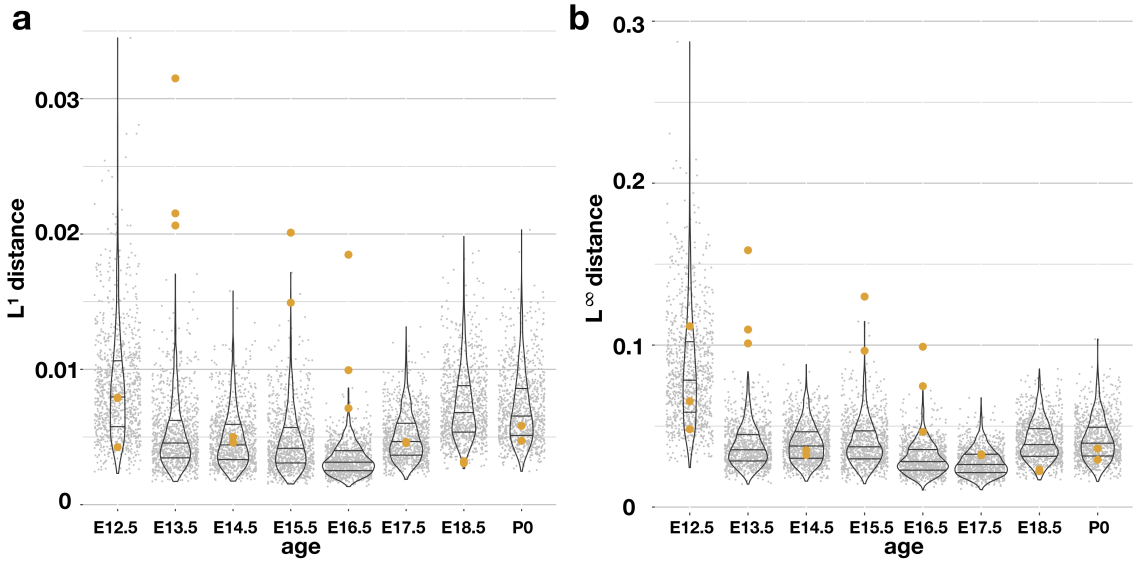
SN1.5.3 Regularization

In addition to the likelihood terms, we used a quadratic regularization on the difference between two neighboring spline sampling points (nodes), e.g. $(\alpha_D)_1 - (\alpha_D)_2$. This results in the objective function:

$$J_\rho(\theta) = -\log(L(\theta)) + \rho\Bigg(\sum_{b=1}^{B}\Bigg[\sum_{i=1}^{n_b^D-1}((\vec{\alpha}_{D^b})_{i+1} - (\vec{\alpha}_{D^b})_i)^2 + \sum_{i=1}^{n_b^v-1}((\vec{\alpha}_{v^b})_{i+1} - (\vec{\alpha}_{v^b})_i)^2 + \sum_{i=1}^{n_b^g-1}((\vec{\alpha}_{g^b,s})_{i+1} - (\vec{\alpha}_{g^b,s})_i)^2 + \sum_{i=1}^{n_t^g-1}((\vec{\alpha}_{g,t})_{i+1} - (\vec{\alpha}_{g,t})_i)^2\Bigg]\Bigg) \tag{SN1.26}$$



SN1 Figure 8: For each day the replicates of the T-cell development data were compared in pairs using a two sample two-sided Kolmogorv-Smirnov test. The resulting p-values are presented together with the significance level $\alpha = 0.05$. Of the eleven possible combinations of the data used for parameter estimation (i.e. all time points without day 0), seven combinations were judged by the K-S test to come from different distributions at a significance level of $\alpha = 0.05$.



SN1 Figure 9: For each measurement day the distribution of distances (L_1 top and L_∞ bottom) between $n = 1000$ subsamples and the pooled cdf is depicted as dots and violin plots with the 25%, 50% and 75% quantiles as horizontal lines. The distances between the measured and the pooled cdf are plotted in yellow. For several days these distances lie clearly outside the subsample-distribution.

where ρ is the regularization hyper-parameter, and n_b^D, n_b^v and n_b^g are the number of spline sampling points (nodes) on branch b out of B branches for the splines $D^b(s)$, $v^b(s)$ and $g_s^b(s)$ on each branch, respectively, and n_g^t is the number of spline sampling points (nodes) for the spline $g_t(t)$ which is a function of time. Note that we did not use time-dependent birth-death rates in all of our models so that $g_t(t)$ and its associated term in eq. SN1.26 were not always included.

This regularization can also be interpreted as a prior in a Bayesian sense,

$$p(\theta) = \rho\Bigg(\prod_{b=1}^{B}\Bigg[\prod_{i=1}^{n_b^D-1}((\vec{\alpha}_{D^b})_{i+1}-(\vec{\alpha}_{D^b})_i)^2\prod_{i=1}^{n_b^v-1}((\vec{\alpha}_{v^b})_{i+1}-(\vec{\alpha}_{v^b})_i)^2 \prod_{i=1}^{n_b^g-1}((\vec{\alpha}_{g^b,s})_{i+1}-(\vec{\alpha}_{g^b,s})_i)^2+\prod_{i=1}^{n_t^g-1}((\vec{\alpha}_{g,t})_{i+1}-(\vec{\alpha}_{g,t})_i)^2\Bigg]\Bigg). \tag{SN1.27}$$

The normalization is included in ρ. Hence, the log-likelihood with regularization can be interpreted as a log-posterior, $J_\rho(\theta) = \log(p(\theta|D))$, with $p(\theta|D) = L(\theta)p(\theta)$.

SN1.5.3.1 Selection of the regularization hyper-parameter for spline smoothness

We performed leave-one-out cross-validation on a time point basis to select the regularization hyper-parameter for the mouse embryonic stem cell differentiation data (Sec. SN3.2.2) and the T-cell development data (Sec. SN3.2.2). In short, we removed the data corresponding to one time point and fitted the model on the reduced data set and evaluated the unregularized likelihood on the withheld time point. We performed this leave-one-out fitting for each time point and added the unregularized log-likelihood value for each withheld time point to receive a score based on which the hyper-parameter can be selected as the maximum likelihood estimator.

SN1.5.4 Uncertainty analysis

We performed multi-start optimization and chose the parameter estimate with the optimal regularized likelihood value as the final parameter estimate θ^*. To assess uncertainty, often a local approximation of the likelihood around θ^* using the Hessian H of the negative log-likelihood is used to calculate approximate confidence intervals:

$$CI_{\alpha,i,\text{approx}} = (\theta_i^*|\exists\theta : (\theta-\theta^*)^T H(\theta-\theta^*) < \Delta_\alpha). \tag{SN1.28}$$

Confidence intervals for the individual parameters can also be computed from profile likelihoods (SN1 Figure 6) (Raue et al. 2009). We computed the likelihood profile $PL_i(c)$ for the i-th parameter, θ_i, by re-optimizing the regularized likelihood for a fixed value $\theta_i = c$, over all other parameters $\theta_{j\neq i}$. The re-optimization is performed for a range of values c in an interval around the optimal value θ_i^*.

$$PL_i(c) = max_{\theta_{j\neq i},\theta_i=c}L(\theta) \tag{SN1.29}$$

Analogously, we consider posterior profiles (Hug et al. 2013) for the likelihood with regularization in the sense that we interpret the regularization as a prior and the likelihood with regularization as a posterior,

$$PP_i(c) = max_{\theta_{j\neq i},\theta_i=c}\exp(J_\rho(\theta)). \tag{SN1.30}$$

The confidence intervals for individual parameters are then given by

$$CI_{\alpha,i} = \left\{c\middle|\frac{PL_i(c)}{L(\theta^*)} \geq \exp\left(-\frac{\Delta_\alpha}{2}\right)\right\} \tag{SN1.31}$$

and

$$CI_{\alpha,i} = \left\{ c \middle| \frac{PP_i(c)}{\exp(J_\rho(\theta^*))} \geq \exp\left(-\frac{\Delta_\alpha}{2}\right) \right\} \tag{SN1.32}$$

where Δ_α is the αth-percentile of the χ^2 distribution with one degree of freedom.

SN1.5.5 Implementation

The numerical implementation is based on the method of lines. The model is discretized in space using finite volumes. We used 300 equally spaced grid points for the cell state linearly scaled into the interval $[0, 1]$. This yields 299 finite volume centers on the main branch at which the mean number density is simulated. The cell state observations in side branches, such as the non-conventional lymphoid cell side branch in the T-cell maturation system, were scaled into the interval $[s_a, 1]$ (SN1 Figure 3) and were discretized at the same resolution in cell state as the main branch. Here, s_a is the earliest cell state observed on the side branch. Accordingly, the number of volume centers on the side branch is $(1 - s_a) \times 299$.

For the solution of the resulting system of ODEs we employ the Sundials CVODE suite and AMICI (Fröhlich et al. 2017) as Matlab interface. We used a multi-start approach (with gradient information) for the optimization and profile likelihood computation which are implemented in the PESTO-Toolbox (Stapor et al. 2018).

Initial parameters for multi-start local optimization were sampled using latin-hypercube sampling. The initial distribution of cells over cell state was obtained from the first sampling point by kernel density estimation using the Matlab function `ksdensity.m`.

SN1.6 Mapping a developmental check-point on a trajectory with mutant data and pseudodynamics: Mapping beta-selection in T-cell maturation using Rag1 and Rag2 knock-out mice

SN1.6.1 Estimation of the checkpoint location

To predict the point of beta-selection across the transcriptomic state coordinate s, we trained the pseudodynamics model on the wild type data subset of the combined wild type and knockout pseudotime data. Note, that we did not use the model parameters obtained from fitting the pseudodynamics model to the diffusion pseudotime coordinates of the wild-type only data because the scaling of the diffusion pseudotime values and the local structure of the overall ordering may slightly change as the neighborhood graph (the diffusion map) changes after the addition of the mutant cells to the wild-type data. We adjusted this model for a developmental arrest at some cell state s'. We then estimated the point of beta-selection as a transcriptomic coordinate as the least-squares estimator s^* of s' on the mutant data (SN1.33).

To imitate the beta-selection process in the forward simulation, we set the drift term to zero for $s > s'$, i.e. cells do not differentiate beyond that point. However, the diffusion rate is not changed, as stochasticity of gene expression still effects the cell state. As cells that did not pass the beta selection will undergo apoptosis if they progress in cell state, we decreased birth-death parameter after the assumed check-point to -3, a lower bound of the previously estimated birth-death parameter.

We computed the value of the least-squares function for s' between the smallest cell state grid point not observed at the initial time point and the highest cell state observed on the T-cell lineage. In a best case scenario, one would also have measurements of the mutant population at the initial time point and choose the maximal cell state observation in this initial mutant sample as the start

of the interval on which the likelihood profile is estimated.

$$s^* = \arg\min_{s'} \log J_{mut}(\mathrm{ecdf}_{S^{mut}_{b,t}} | v_{mut}(s|s'), D_{WT}(s), g_{mut}(s|s')) \tag{SN1.33}$$

$$v_{mut}(s|s') = \begin{cases} v_{WT}(s), & \text{if } s \leq s' \\ 0, & \text{otherwise} \end{cases} \tag{SN1.34}$$

$$g_{mut}(s|s') = \begin{cases} g_{WT}(s), & \text{if } s \leq s' \\ -3, & \text{otherwise .} \end{cases} \tag{SN1.35}$$

As we only have two replicates of the Rag2 knock-out measurement time points at E16.5 ($S^{mut,1}_{b,t=16.5}$, $S^{mut,2}_{b,t=16.5}$) and one Rag1 knock-out sample of E14.5 ($S^{mut}_{b,t=14.5}$), we used a least-squares objective function of the area between the measured ecdfs and computed cdfs (SN1.19). To incorporate the two replicates for the second time point (E16.5), we evaluated the ecdfs for the two replicates on the grid that was used for simulation and compared the simulated distribution to the mean distribution on this grid,

$$\begin{aligned} J_{mut}(\mathrm{ecdf}_{S^{mut}_{b,t}} | \theta_{mut}) = {} & \mathcal{A}\left(\mathrm{ecdf}_{S^{mut}_{b,t=14.5}}(s), \mathrm{cdf}_{u_b}(s, t=14.5|\theta_{mut})\right)^2 \\ & + \mathcal{A}\left(\mathrm{mean}\left(\mathrm{ecdf}_{S^{mut,1}_{b,t=16.5}}(s), \mathrm{ecdf}_{S^{mut,2}_{b,t=16.5}}(s)\right), \mathrm{cdf}_{u_b}(s, t=16.5|\theta_{mut})\right)^2 \end{aligned} \tag{SN1.36}$$

with $\theta_{mut} = (v_{mut}(s|s'), D_{WT}(s), g_{mut}(s|s'))$.

SN1.6.2 Incorporation of mutant data into a pseudotemporal ordering of wild type data

It is necessary to have a common coordinate system of wild-type and knock-out development to integrate both in the pseudodynamics analysis.

There are multiple ways to include mutant data into a pseudotemporal ordering. Firstly, one could attempt a mapping if the cells in the mutant samples do not fall into the developmental manifold defined on the wild-type samples. We used a second approach and computed the developmental manifold directly on the union of wild-type and knock-out samples, which is only possible if the resulting manifold is not dominated by genotype effects: The pseudotemporal ordering of the combined wild-type and Rag1/2 knock-out samples was not dominated by the difference between wild type and knock-out and we could therefore directly use the diffusion pseudotime coordinates for model training and beta-selection point prediction (Supp. Fig. 13). We validated the diffusion map model for the combined wild-type and mutant data by comparing it to the wild-type only model:

Firstly, we checked that the mutant cells fall into the wild-type manifold in the diffusion map. We observed the distribution of mutant and wild-type cells in 2D cross-sections of the diffusion map in the first few diffusion components. In each projection, the mutant cells were within the manifold cross-section defined by the mutant cells (Supp. Fig. 13-c).

Secondly, we checked that the overall structure of the developmental manifold in transcriptome space is not changed. We plotted the diffusion pseudotime coordinates of the wild-type cells of the wild-type-only and the combined wild-type and knock-out data set against each other (Supp. Fig. 13d). The comparison of these two sets of pseudotime coordinates is not a comparison of scalar values but one of orderings as the diffusion pseudotime coordinates change if mutant cells are added to the graph.

We compared these two orderings with a rank-based correlation metric: Kendell's tau coefficient (Kendell's rank correlation coefficient). Kendell's tau coefficient was 0.967 on the progenitor

and T-cell subset of wild-type cells and 0.916 on the progenitor and non-conventional lymphocyte subset suggesting that the overall structure of the neighborhood graph was not changed by the addition of the mutant cells.

Supplementary Note 2: Pseudodynamics simulation studies

Contents

SN2.1 Introduction to this Supplementary Note

This document details the set-up and results of the simulation studies which we used to characterise the performance of pseudodynamics.

Cross-references to elements of the other Supplementary Notes are labels with "SNX" for Supplementary Note X.

SN2.2 Simulation studies

We performed forward simulations of the pseudodynamics model with given parameter values and subsequently sampled population density and size observations from the the simulated densities at various time points. We then fit the pseudodynamics model to these simulated observations to show identifiability of the model in different scenarios. We show that the model parameters are practically identifiable, in the sense of structural identifiability, given sufficient data and that one can distinguish a drift- and a birth-death rate induced steady state based on the model fits.

SN2.2.1 Identifiability

SN2.2.1.1 Aim

We show that the parameters of the pseudodynamics model are practically identifiable based on population density and population size observations for two sets of simulated data.

SN2.2.1.2 Data Generation

To generate the data, we simulated the model (eq. (SN1.1)) for the time points $t = (1, 2, 3, 4, 5)$ for two different parameter sets. A log-normal distribution with parameters $\mu = -2.7$ and $\sigma = 0.8$ multiplied by $N_0 = 1.5 \times 10^4$ was chosen as initial distribution. From the simulated cell density over cell state we computed the total number of cells and the probability distribution of the cell states. Using the pdf, we drew three samples of cell state measurements for 10000 cells each. We further generated a sample of twenty population sizes for each simulated time point using the simulated population size as mean of a normal distribution with variance for each time point given by $\sigma_N = (1 \times 10^6, 2 \times 10^6, 3 \times 10^6, 4 \times 10^6, 5 \times 10^6, 6 \times 10^6)$ for the first example and $\sigma_N = (1 \times 10^7, 5 \times 10^7, 2 \times 10^8, 2 \times 10^9, 2 \times 10^{10}, 2 \times 10^{11})$ for the second example.

SN2.2.1.3 Implementation

The diffusion and drift parameters, $D(s)$ and $v(s)$, were parametrized using cubic splines in log space and with nine equally spaced sampling points in s. The rates $D(s)$ and $v(s)$ were computed as the exponential of the spline. The birth-death parameter, $g(s)$, was parametrized using cubic splines in linear space with nine equally spaced sampling points in s. This left the values $D_1, \ldots, D_9$, $v_1, \ldots, v_9$ and $g_1, \ldots, g_9$ for estimation, which are the nodes of the splines in log space.

The drift rate is decreased to zero in the interval $[0.9, 1]$ of cell states by using a hermite c-spline. We performed 80 multi-starts using an interior point method. We investigated the regularizations $\rho = \{0, 1\}$. The confidence intervals were computed for $\rho = 0$ using likelihood profiles eq. (SN1.29).

SN2.2.1.4 Results

For the first example, the density (SN2 Figure 1 a) and population size (SN2 Figure 1 b) fits agree well with the data to which the model was fitted to. The parameter splines used for the generation of the simulated data lie mostly in or close to the 99% confidence intervals of the the estimated parameters (SN2 Figure 1 c-d). Also in the second example, the density (SN2 Figure 2 a) and population size (SN2 Figure 2 b) fits agree well with the simulated data. The parameter splines used for the generation of the simulated data lie mostly in or close to the 99% confidence intervals of the estimated parameters (SN2 Figure 2 c-d). The diffusion rate in this case has a comparably

high uncertainty. Also confidence intervals get broader in the region with fewer samples and parameters become unidentifiable in this region.

We conclude that our parameter estimation can recover a pseudodynamics model parameterization that still allows for qualitative and quantitative statements about the underlying biology.

SN2.2.2 Distinguishing a drift- and a birth-death rate-induced steady state

SN2.2.2.1 Aim

Cell populations that develop according to the pseudodynamics model can generate similar normalized density samples in cell state space in subsequent time points (steady state).

Such steady state observations can correspond to a local minimum (attractor) on the underlying developmental potential which corresponds to zero drift at the minimum (as the drift can be thought of as the derivative of this developmental potential). A drift close to zero at and after the attractor in cell state space corresponds to a lack of directed development leading away from this attractor.

However, such steady state observations can also be a result of a non-uniform birth-death rate around the steady-state if source-sink-like effects are present: The normalized density may be stationary if the birth death rate is positive before the observed steady-state and negative after. Note that such a steady-state does not correspond to a local minimum on the developmental potential (quasi-steady state).

We show that one can distinguish a drift- and a birth-death rate-induced steady state based on the pseudodynamics fits and that the model parameters are practically identifiable.

SN2.2.2.2 Data Generation

To generate the data, we simulated the model eq. (SN1.1) for the time points $t = (1, 2, 3, 4, 5)$. A log-normal distribution with parameters $\mu = -2.7$ and $\sigma = 0.8$ multiplied by $N_0 = 1.5 \times 10^4$ was chosen as initial distribution. From the simulated cell density over cell state we computed the total number of cells and the probability distribution of the cell states. Using the pdf, we drew three samples of cell state measurement for 10^4 cells each. We further generated a sample of twenty population sizes for each simulated time point using the simulated population size as mean of a normal distribution with variance for each time point given by $\sigma_N = (1 \times 10^7, 5 \times 10^7, 2 \times 10^8, 2 \times 10^9, 2 \times 10^{10}, 2 \times 10^{11})$.

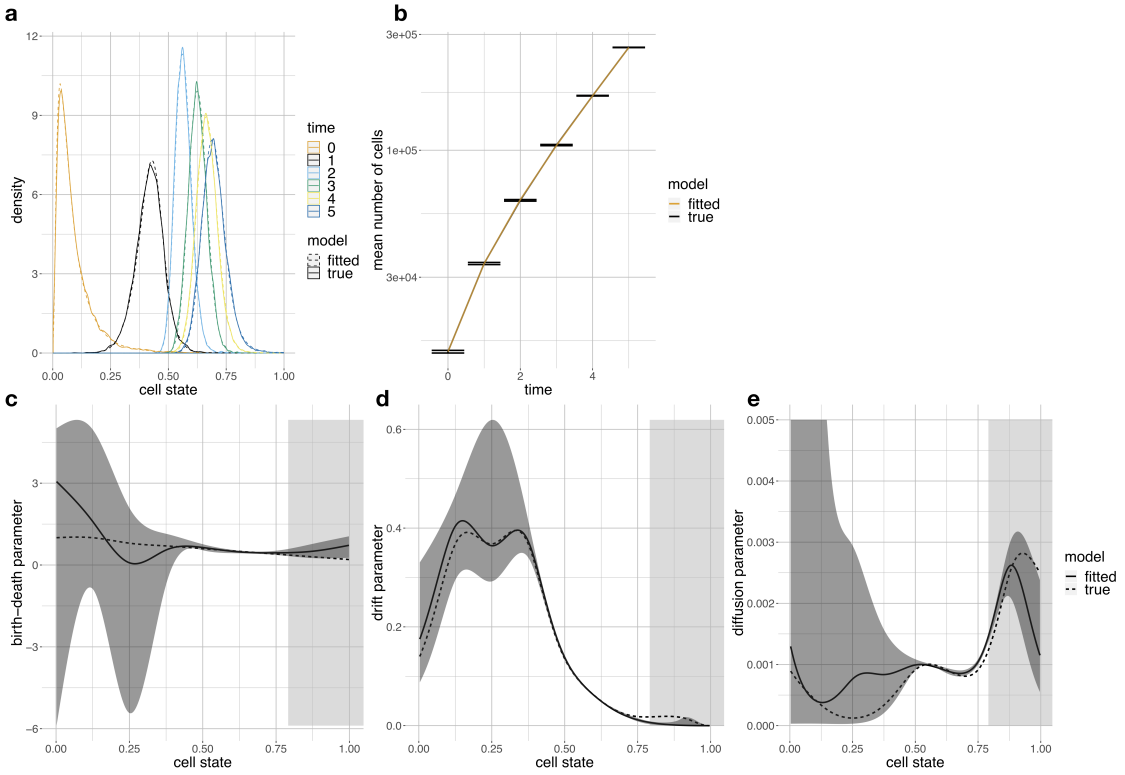
Using the setting described above, we simulated data for two test cases. In the first case, we assume that the system is governed by an attractor, i.e. at a certain cell state the drift decreases and cells are not differentiating further. In the second case we consider a quasi steady state behavior. Cells are still differentiating however the cell death increases at a certain cell state.

SN2.2.2.3 Implementation

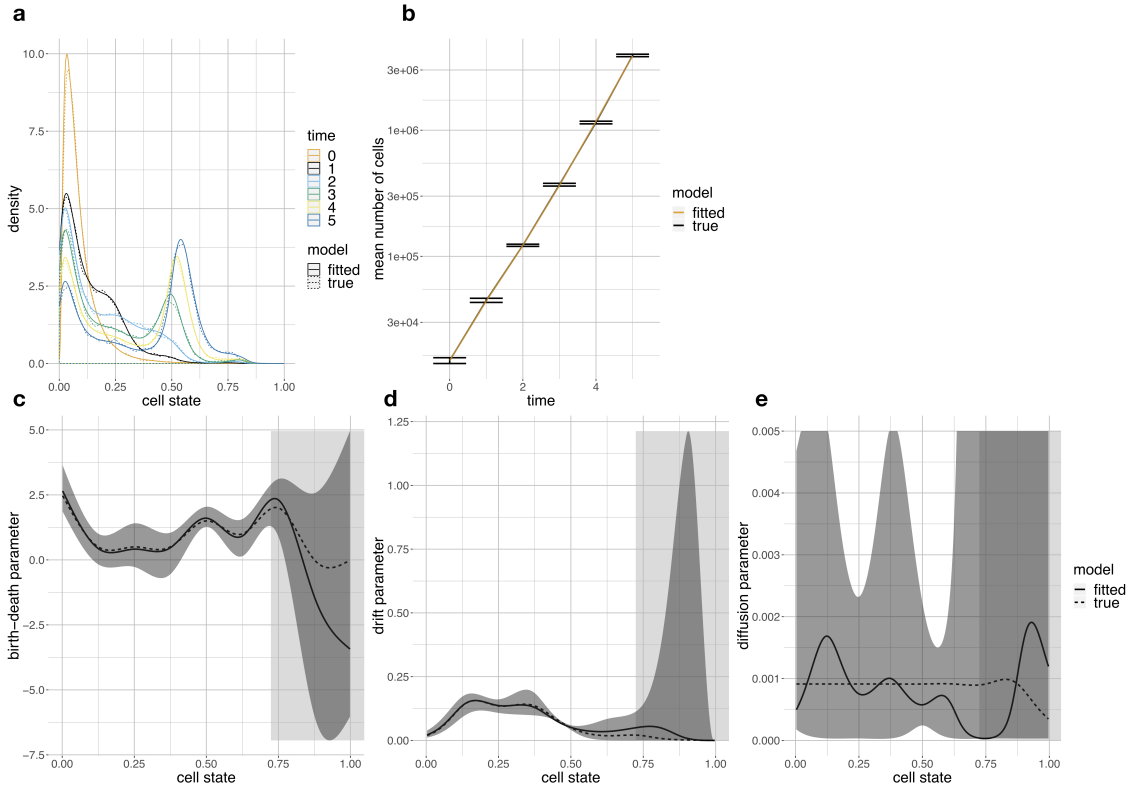
For this analysis, we employed the numerical implementation described in Sec. SN2.2.1.3.

SN2.2.2.4 Results

Firstly, we showed that an attractor and a quasi-steady state pseudodynamics parameterization can induce steady-state like observations in cell state space (SN2 Figure 3 a,b). The density (SN2 Figure 3 a,b) and population size (SN2 Figure 3 c) fits agree well with the simulated data for the respective case. The parameters of both models are not fully identifiable for $s > 0.75$ (SN2 Figure 3 d-f, right of the vertical lines) in the cell state space in which 1% of the observed cells lie. However, the parameter used for the data generation lie within the 99% confidence interval



SN2 Figure 1: The pseudodynamics model parameters are practically identifiable on simulated data (I). **(a)** Population density fits and raw densities to which the model is fit to (true). **(b)** Population size fits and raw sizes to which the model is fit to (true) with one standard deviation around the mean as error bars with $n = 20$ samples at each time point. **(c-e)** Birth-death rate **(c)**, drift **(d)** and diffusion parameter **(e)** fits and true values (used for simulation). The shaded area represents the approximate 99% confidence interval (CI) which was approximated as a spline through the 99% confidence interval upper and lower bounds on the spline nodes. The confidence intervals on the spline nodes were calculated using profile likelihoods eq. (SN1.29). The rectangular shaded areas represent the cell state interval at which the cumulative density of the set of all cells from all time points used for model fitting is above 99%. Accordingly, only 1% of the simulated cells fall into this interval.



SN2 Figure 2: The pseudodynamics model parameters are practically dentifiable on simulated data (II). **(a)** Population density fits and raw densities to which the model is fit to (true). **(b)** Population size fits and raw sizes to which the model is fit to (true) with one standard deviation around the mean as error bars with $n = 20$ samples at each time point. **(c-e)** Birth-death rate **(c)**, drift **(d)** and diffusion parameter **(e)** fits and true values (used for simulation). The shaded area represents the approximate 99% confidence interval (CI) which was approximated as a spline through the 99% confidence interval upper and lower bounds on the spline nodes. The confidence intervals on the spline nodes were were calculated using profile likelihoods eq. (SN1.29). The rectangular shaded areas represent the cell state interval at which the cumulative density of the set of all cells from all time points used for model fitting is above 99%. Accordingly, only 1% of the simulated cells fall into this interval.

of the estimated parameters. The attractor is characterized by a drop of the drift parameter to zero around the observed steady state. The quasi-steady state is characterized by a transition of a positive to a negative birth-death rate around the observed steady state. For the considered scenario, the statistical inference using the pseudodynamics framework allows a clear distinction of both scenarios based on birth-death rate and drift parameter (SN2 Figure 3 d,e) even when accounting for parameter uncertainties. One does not need model selection in this scenario but one can simply identify the steady-state generating mechanism based on the parameter values.

SN2.2.3 Investigation of Influence of Single Cell Sampling Bias

SN2.2.3.1 Aim

There is evidence of sampling bias on the single cell level (Segerstolpe et al. 2016) (Haber et al. 2017). In this section we want to investigate how a bias affects parameter estimation and how it can be included in parameter estimation using the likelihood. We simulated data for an unbiased and biased case and performed parameter estimation using the likelihood (eq. (SN1.12)) on both data sets. Additionally, we included a bias correction in the likelihood and repeated the parameter estimation on the biased data set.

SN2.2.3.2 Measuring sampling bias

Conceptually, sampling bias could be assessed by acquiring unbiased population distribution measurements independently from single-cell RNA-seq, for example by measuring the distribution of a population across surface marker-gated sub-populations in flow-cytometry or in stained tissue sections. One could then map cell states between the unbiased surface-marker-based measurements and single-cell RNA-seq, e.g. clusters in single-cell RNA-seq that correspond to cell types that can be identified via surface markers. A linear sampling bias in single-cell RNA-seq could then be identified as a shift in the population distribution between both measurements.

SN2.2.3.3 Data Generation

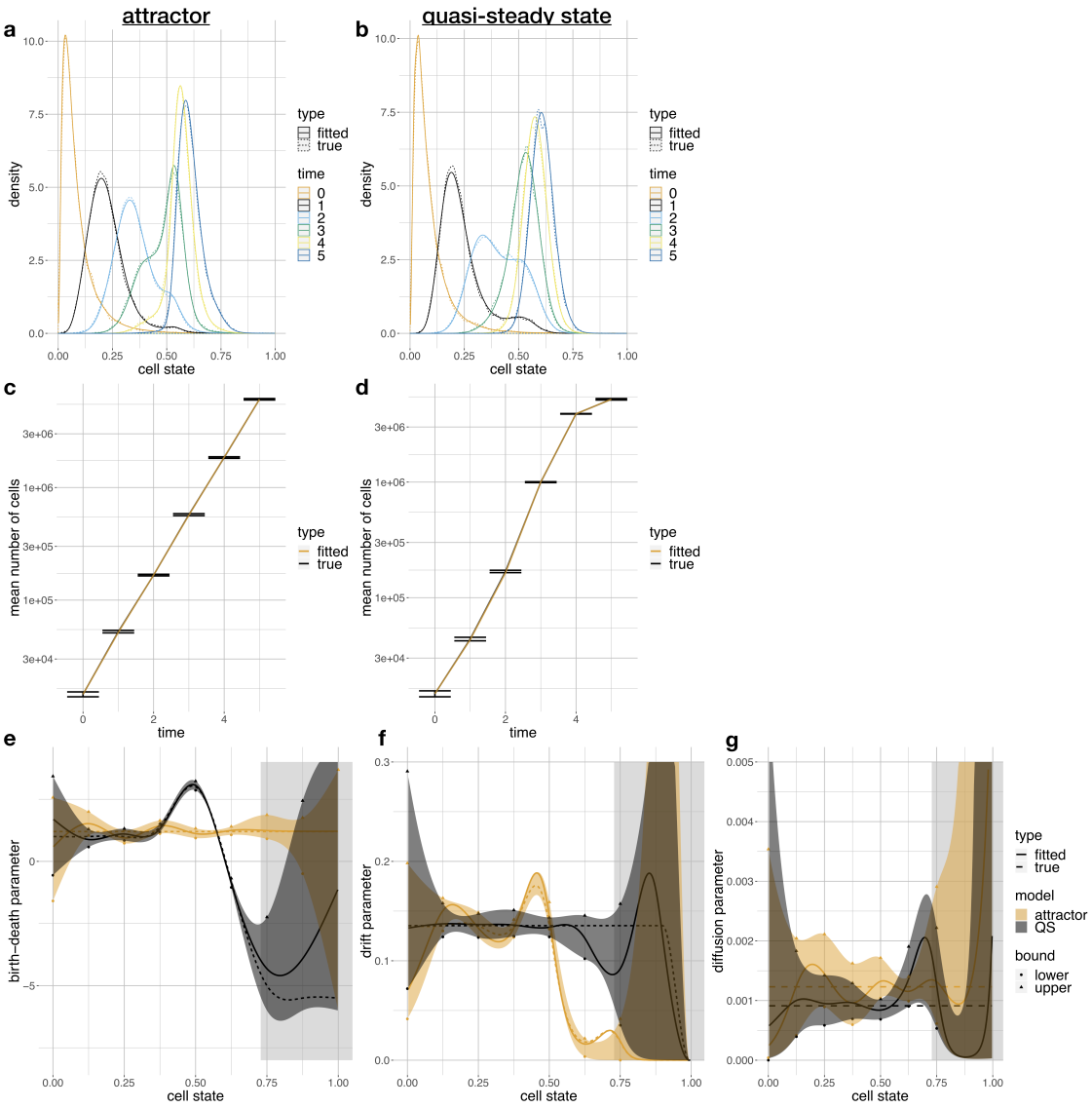
To generate the data, we simulated the model (eq. (SN1.1)) for the time points $t = (1, 2, 3, 4, 5)$. A log-normal distribution with parameters $\mu = -2.7$ and $\sigma = 0.8$ multiplied by $N_0 = 1.5 \times 10^4$ was chosen as initial distribution. From the simulated cell density over cell state we computed the total number of cells and the probability distribution of the cell states. For the unbiased data set, we drew three samples of cell state measurement for 10000 cells each from the simulated pdf. For the biased case, we assumed a linear scaling of the pdf towards later time points, $bias = as + b$. From the simulation only the probability to find a cell in an interval $[s_i, s_{i+1})$,

$$p_i(t) := \frac{1}{\int_0^1 u(s,t)ds} \int_{s_i}^{s_{i+1}} u(s,t)ds = \frac{1}{N(t)} \int_{s_i}^{s_{i+1}} u(s,t)ds \text{ for } i = 1, \ldots, n_v,$$

can be computed. We computed the biased probability to sample a cell in this interval as

$$p_i^{bias}(t) = p_i(t) \left(a\frac{s_i + s_{i+1}}{2} + b \right) \frac{1}{\sum_{i=1}^{n_v} p_i(t) \left(a\frac{s_i+s_{i+1}}{2} + b \right)}. \tag{SN2.1}$$

We created a biased cell state sample, by sampling 10000 cells from p_i^{bias}. This was repeated to create three replicates. We further generated a sample of twenty population size measurements for each simulated time point using the simulated population size as mean of a normal distribution



SN2 Figure 3: Pseudodynamics fits on simulated steady-state data. All plots shown are fits with a regularization parameter of zero. **(a,b)** Population density fits and raw densities to which the model is fit to (true) for the attractor **(a)** and the quasi-steady state **(b)** model. **(c,d)** Population size fits and raw sizes to which the model is fit to (true) for the attractor **(c)** and the quasi-steady **(d)** state model with one standard deviation around the mean as error bars with $n = 20$ samples at each time point. **(e-g)** Birth-death rate **(e)**, drift **(f)** and diffusion parameter **(g)** fits and true values (used for simulation) for both the attractor and the quasi-steady state (QSS) model. The shaded area represents the approximate 99% confidence interval (CI) which was approximated as a spline through the 99% confidence interval upper and lower bounds on the spline nodes. The confidence intervals on the spline nodes were calculated using profile likelihoods eq. (SN1.29). The rectangular shaded area represent the cell state interval at which the cumulative density of the set of all cells from all time points used for model fitting is above 99%. Accordingly, only 1% of the simulated cells fall into this interval.

with variance for each time point given by $\sigma_N = (3\times10^6, 7\times10^6, 1\times10^7, 7\times10^6, 7\times10^6, 5\times10^6)$ for both, the biased and the unbiased data set. The population size is not affected by the sampling bias.

SN2.2.3.4 Implementation

For this analysis, we employed the numerical implementation described in Sec. SN2.2.1.3.

Bias correction To correct for the bias in the statistical inference, firstly the PDE is simulated and the unbiased pdf, $p_i(t)$, is calculated. Secondly, the simulated pdf is biased according to the bias in the measurement, i.e., in our case eq. (SN2.1). The cdf in eq. (SN1.12) is computed from the biased pdf. In this simulation example, we assume that the bias was experimentally measured correctly and supply the likelihood with the correct scaling parameters, a and b. However, the effect of the bias can alternatively also be parametrized and fitted.

SN2.2.3.5 Results

The parameter estimation on the unbiased data provided a good fit to the simulated data (SN2 Figure 4 a)-f)) and recovered the true parameters. Using the biased data and the standard likelihood a good fit to the biased data was achieved (SN2 Figure 4 a)-f)). However, biased parameter estimates were recovered (SN2 Figure 4 g)-i)). Fitting to the biased data using the likelihood with bias correction yielded a good fit to the unbiased data (SN2 Figure 4 a)-f)) and the true parameter lay in the 99% CI of the estimated parameter (SN2 Figure 4 g)-i)).

SN2.2.4 Simulated data for a branching model

SN2.2.4.1 Aim

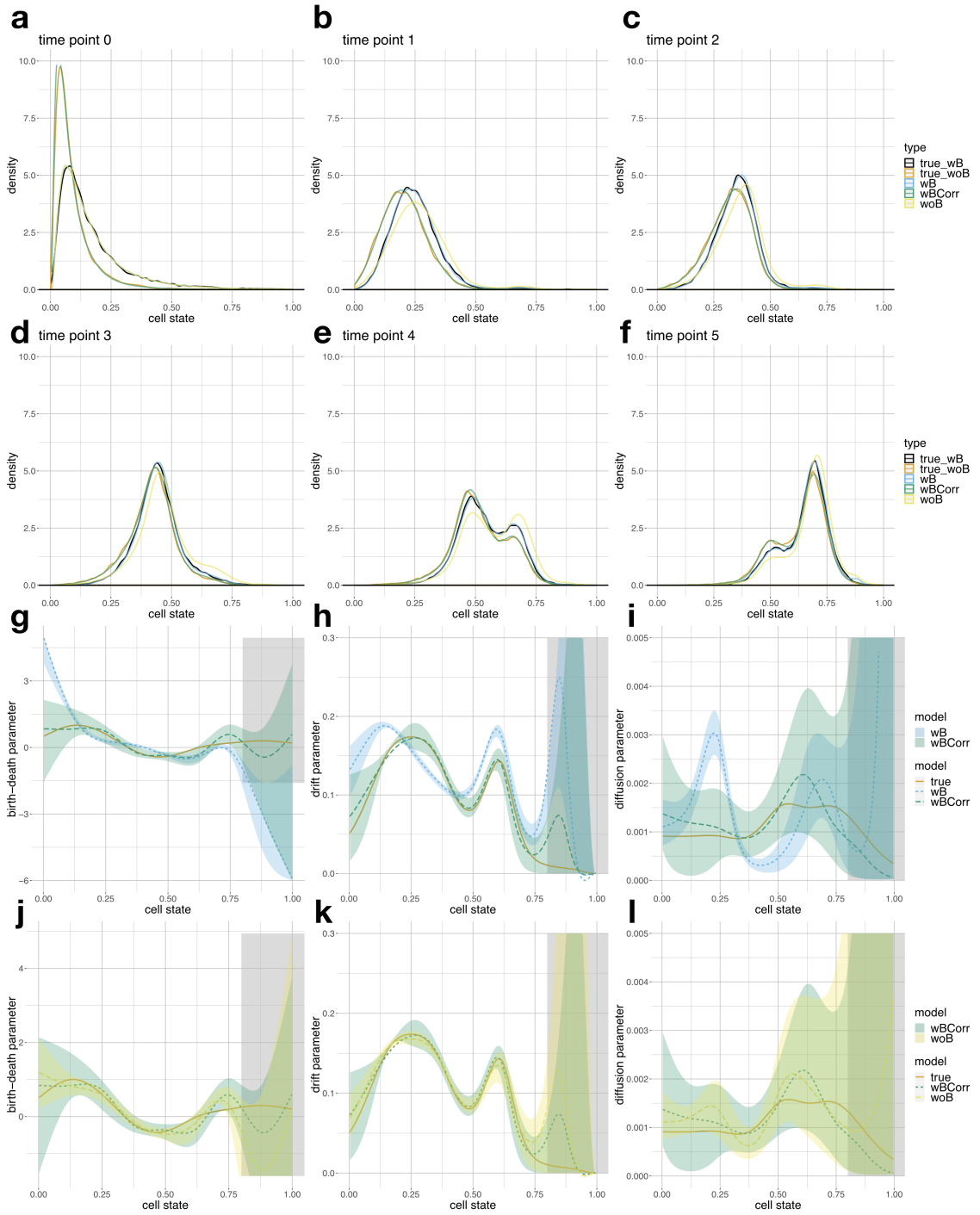
We consider parameter estimation for a model with branching eq. (SN1.8) and show that parameters are identifiable based on the considered data.

SN2.2.4.2 Data Generation

We simulated model eq. (SN1.8) for the time points $t = (0, 1, 2, 3, 4, 5)$, with branching region $[s_a, s_b] = [69/299, 130/299]$. A log-normal distribution with parameters $\mu = -2.7$ and $\sigma = 0.8$ multiplied by $N_0 = 1.5 \times 10^4$ was chosen as initial distribution on branch 1 and a log-normal distribution with parameters $\mu = -2.4$ and $\sigma = 0.9$ multiplied by $N_0 = 2 \times 10^3$ was chosen as initial distribution on branch 2. We drew the number of single cells measured on each branch at each time point from a binomial distribution with success probability given as the simulated proportion on branch 1 and number of trials given by the (preset) total number of samples of 10000. This was repeated three times to generate replicates. Using the sampled numbers of cells on each branch, we sampled from the pdf of the simulated cell state distribution on each branch for each measurement time point. We also replicated the cell state samples three times. From the simulated total number of cells, we generated 20 samples of normally distributed population sizes with variance for each time point given by $\sigma_N = (1 \times 10^7, 5 \times 10^7, 2 \times 10^8, 2 \times 10^9, 2 \times 10^{10}, 2 \times 10^{11})$ for each time point.

SN2.2.4.3 Implementation

The diffusion and drift parameters, $D_b(s)$ and $v_b(s)$, were parametrized using cubic splines in log space and with nine equally spaced sampling points in s for branch 1 and three equally spaced



SN2 Figure 4: The figure caption is on the next page.

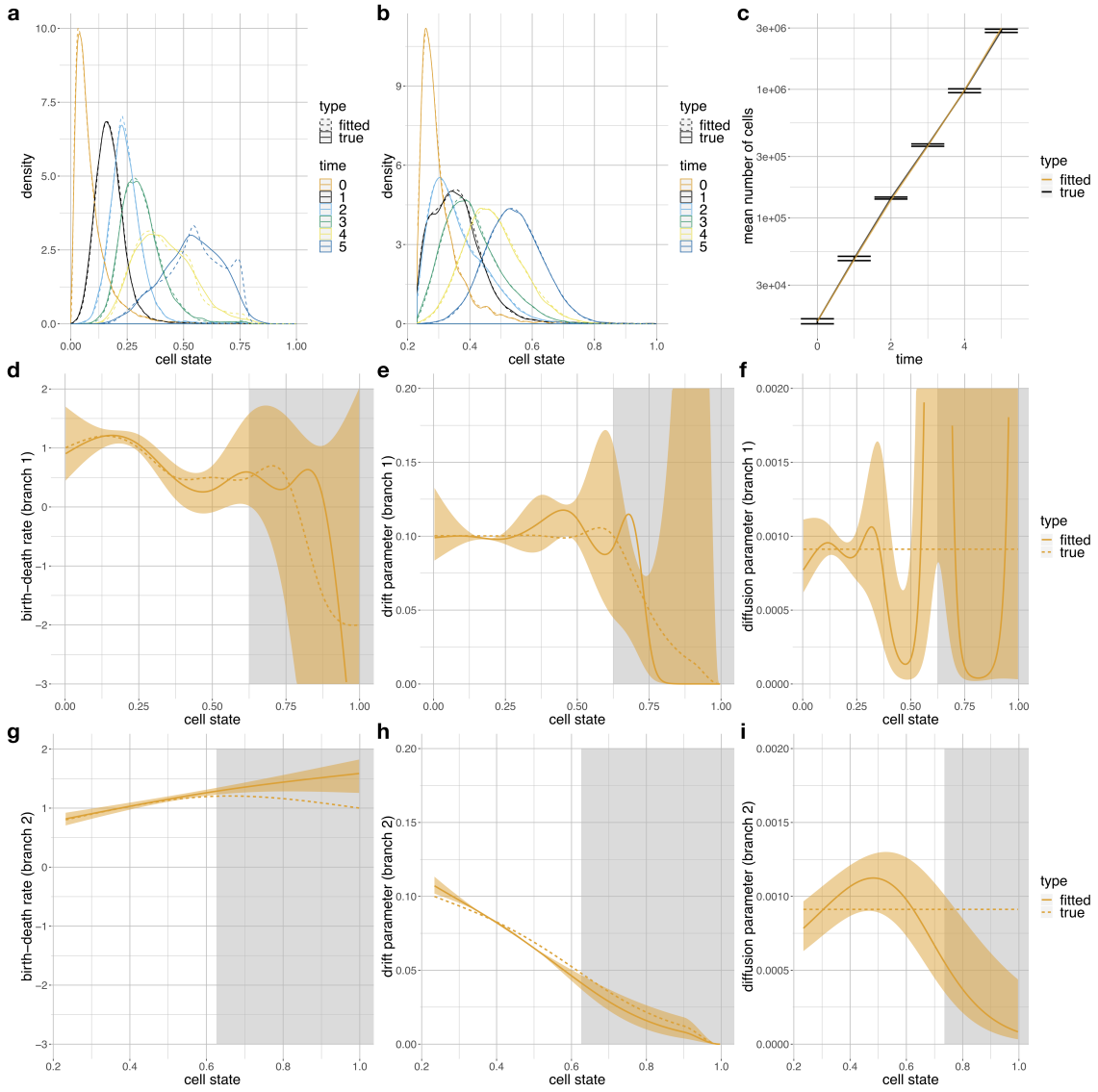
SN2 Figure 4: The pseudodynamics model parameters are practically identifiable on simulated data with biased sampling. true_wbias: biased population density measurements which are the input to model fits, true_wobias: latent unbiased population density measurements which we hope to recover in model fits, wB: model fit based on biased data without bias correction, wBCorr: model fit based on biased data with bias correction, woB: model fit based on unbiased data (and without bias correction). **(a-f)** Population density fits and measured densities by time point. data biased: density after biased subsampling, data unbiased: density before biased subsampling. **(g-l)** Birth-death rate **(g,j)**, drift **(h,k)** and diffusion parameter **(i,l)** fits with bias correction (wBCorr) and true parameter values which were used for simulation (true: unbiased). **(g-i)** includes the parameter fits without bias correction (wB) as reference, **(j-l)** includes the fits to unbiased data (woB) as reference. The shaded area represents the approximate 99% confidence interval (CI) which was approximated as a spline through the 99% confidence interval upper and lower bounds on the spline nodes. The confidence intervals on the spline nodes were calculated using profile likelihoods eq. (SN1.29). The rectangular shaded areas represent the cell state interval at which the cumulative density of the set of all cells from all time points used for model fitting is above 99%. Accordingly, only 1% of the simulated cells fall into this interval.

sampling points in s for branch 2. The rates $D_b(s)$ and $v_b(s)$ were computed as the exponential of the spline. The birth-death parameter, $g_b(s)$, was parametrized using cubic splines in linear space with nine equally spaced sampling points in s for branch 1 and three equally spaced sampling points in s for branch 2. This left the values $D_{1,1}, \dots, D_{1,9}$,$D_{2,1}, \dots, D_{2,3}$, $v_{1,1}, \dots, v_{1,9}$, $v_{2,1}, \dots, v_{2,3}$, $g_{1,1}, \dots, g_{1,9}$, $g_{2,1}, \dots, g_{2,3}$ and the propensities for switching the branch $\delta_{1,2}$ and $\delta_{2,1}$ for estimation. The drift rates on are decreased to 0 in the interval $s \in [0.9, 1]$ of cell states by using a hermite c-spline on both branches. We performed 120 multi-starts using an interior point method and investigated the regularizations $\rho = \{0, 1\}$. The confidence intervals were computed for $\rho = 0$ using profile likelihoods eq. (SN1.29).

SN2.2.4.4 Results

The parameter estimation on the branching data provided a good fit for both branches and the total population (SN2 Figure 5 a)-c)). The parameter splines used for the generation of the simulated data lie mostly in or close to the 99% confidence intervals of the estimated parameters (SN2 Figure 5 d)-i)).

We conclude that the parameter estimation can recover a pseudodynamics model parameterization that still allows for qualitative and quantitative statements about the underlying biology also for the branching model.



SN2 Figure 5: The pseudodynamics model parameters are practically identifiable on simulated data of a branching process (II). **(a,b)** Population density fits and raw densities to which the model is fit to (true) by branch: main branch **(a)** and side **(b)**. **(c)** Population size fits and raw sizes to which the model is fit to (true) with one standard deviation around the mean as error bars with $n = 20$ replicates per time point. **(d-i)** Birth-death rate **(d,g)**, drift **(e,h)** and diffusion parameter **(f,i)** fits and true values (used for simulation) by branch. Parameters on branch 1 in **(d-f)** and parameters on branch 2 in **(g-i)**. The shaded area represents the approximate 99% confidence interval (CI) which was approximated as a spline through the 99% confidence interval upper and lower bounds on the spline nodes. The confidence intervals on the spline nodes were calculated using profile likelihoods eq. (SN1.29). The rectangular shaded areas represent the cell state interval at which the cumulative density of the set of all cells on this branch from all time points used for model fitting is above 99%. Accordingly, only 1% of the simulated cells on this branch fall into this interval.

Supplementary Note 3: Pseudodynamics application details

Contents

SN3.1 Introduction to this Supplementary Note

This document contains detailed descriptions of the processing of and the results on the data sets presented in the main text.

Cross-references to elements of the other Supplementary Notes are labels with "SNX" for Supplementary Note X.

SN3.2 Application of the pseudodynamics model to data

SN3.2.1 Mouse embryonic stem cell differentiation in vitro

SN3.2.1.1 Aim

We used pseudodynamics on this previously published data set to show the ability of our framework to generalize to other data sets and to demonstrate the core idea/concept of the framework: resolving the dynamics (time-dependency) of a distribution of cells in cell state space during development. The analysis of this data set does not include population growth dynamics.

SN3.2.1.2 Model

We chose the non-branching pseudodynamics model for this system (Sec. SN1.3.1): The linear trajectory corresponds to the trajectory from embryonic stem cell to a differentiated state to which the cells move after the removal of differentiation inhibitor in vitro.

SN3.2.1.3 Implementation

The diffusion and drift parameters, $D(s)$ and $v(s)$, were parametrized using natural cubic splines based on nine equally spaced sampling points in s.

Accordingly, only the values $D_1, \ldots, D_9$ and $v_1, \ldots, v_9$ had to be estimated. Population size estimates are difficult to interpret and were not available for this in vitro experiment. Therefore, the population growth parameter $g(s)$ was set to zero. For numerical reasons we employed a log-transformation of the parameters. We computed the exponential of the spline to ensure that $D(s)$ and $v(s)$ are positive. The drift rate is decreased to 0 in the interval $[0.9, 1]$ of cell states by using a hermite c-spline. We performed 40 multi-starts using an interior point algorithm for each of the cross-validation folds and investigated the regularizations $\rho = \{0, 1, 10\}$.

SN3.2.1.4 Raw data

The data consists of one Drop-seq sample of each of the four time points (0d, 2d, 4d, 7d) after LIF removal (Klein et al. 2015). We based our analysis on the processed data set nmeth.3971-S2.mat (Haghverdi et al. 2016) on which a diffusion map and diffusion pseudotime have already been successfully fit before (Haghverdi et al. 2016).

SN3.2.1.5 Data processing

The steps described in this section are also explained with code and plots in the jupyter notebooks that were used to perform this analysis (Supp. data 1.1). We fitted a diffusion map to all cells

(scanpy: k=20, knn=False)(Wolf, Angerer, and Theis 2018). We fitted diffusion pseudotime to this data set with the extremal cell in diffusion component 1 on the side of the manifold dominated by the time point 0d sample as a root cell.

SN3.2.1.6 Results

The results of this analysis are presented in the main text in section "Pseudodynamics interpolates time series samples from stem cell differentiation". Here, we also discuss the fit quality:

This data set presents three challenges for inference: a) few sampled time points, b) no replicates and c) no total population size estimators available. Accordingly, one risks over-fitting if one includes weak penalties for model complexity and non-identifiability if one include non-zero birth-death rates. The fits in the main text in Fig. 1e are the fits of a conservative model that is not likely to over-fit: A model that was fit without birth-death rate parameters (all set to zero) and with a model complexity penalty. Accordingly, this parameterization performs well in cross-validation. We added less constrained fits with no model complexity penalty to convince the readers that the under-fitting is not a model flexibility issue (main text Fig. 1f): The fits to the data improve but the model over-fits, which we show via cross-validation.

SN3.2.2 T-cell development in the embryonic thymus

SN3.2.2.1 Aim

We used pseudodynamics on this data set to elucidate the dynamics of T-cell maturation including population growth phenomenons which link to selection as described in the main text. This analysis makes use of the full capabilities of pseudodynamics, including branching processes, integration of replicates, integration of knock-outs and population growth dynamics.

SN3.2.2.2 Model

We chose the branching pseudodynamics model with two branches for this system (Sec. SN1.3.2) for the diffusion pseudotime-based analysis: The main branch corresponds to the trajectory from progenitor to mature $\alpha\beta$-T-cells and the side branch to the non-conventional lymphocyte lineage. We chose the non-branching pseudodynamics model (Sec. SN1.3.1) for the monocle2-based analysis: The branch corresponds to the trajectory from progenitor to mature $\alpha\beta$-T-cells.

SN3.2.2.3 Implementation

The diffusion, drift and population growth parameters ($D(s)$, $v(s)$ and $g(s)$) were parametrized with natural cubic splines based on nine equally spaced sampling points in s on the main branch and with three sampling points on the side branch. This left the values $D_1, \dots, D_{12}$, $v_1, \dots, v_{12}$ and $g_1, \dots, g_{12}$ and the transition propensities $\delta_{1,2}$ and $\delta_{2,1}$ for estimation. For numerical reasons we employed a log-transformation of the parameters, except for $g_1, \dots, g_{12}$. To ensure positivity of $D(s)$ and $v(s)$ we computed the exponential of the spline. The drift rate is decreased to zero in the cell state interval, $[0.9, 1]$, by using a hermite c-spline. We performed up to 120 multi-starts using an interior point algorithm. If a plateau of 6 starts was achieved at 40 or 80 multi-starts we did not perform the full 120 starts. Even though convergence was not achieved for all cross-validation folds, the results seem to be robust in the evaluated interval of regularization parameters and we investigated the regularizations $\rho = \{0, 1, 10, 30, 100, 300, 1000\}$.

To choose the regularization we performed leave-one-out cross validation. For each regularization parameter, we also estimated the parameter on the data excluding all data corresponding to one time point. This was repeated for all time points. Using the parameter estimates from the

reduced data sets, the log-likelihood on the data from the left out time point can be computed. We compared the prediction error, i.e. the sum of the log-likelihood at all withheld time points (Supp. Fig. 9a), and chose the regularization with the smallest prediction error (in this case $\rho = 10$).

The confidence intervals were computed using profile posteriors (SN1.30) for $\rho = 10$. The confidence interval of the spline node at cell state $s = 0$ of the drift parameter estimate along the T-cell lineage was not upper-bounded by the bound on the parameter in log space (0) and was therefore set to this upper bound for the computation of the spline through the upper bounds of the confidence intervals of the spline nodes of the drift parameter in Fig. 2d. For the fitting of the data ordered by monocle2, the diffusion, drift and birth-death parameters, $D(s)$, $v(s)$ and $g(s)$, were parameterized with natural cubic splines based on nine equally spaced sampling points in s. This left the values $D_1, \ldots, D_9$, $v_1, \ldots, v_9$ and $g_1, \ldots, g_9$ for estimation. For numerical reasons we employed a log-transformation of the parameters D_i and v_i, $i = 1, \ldots, 9$. We computed the exponential of the spline to ensure that $D(s)$ and $v(s)$ are positive. The drift rate is decreased to 0 in the interval $[0.9, 1]$ of cell states by using a hermite c-spline. We performed 80 multi-starts using an interior point method. We investigated the regularizations $\rho = \{10, 30, 100\}$.

SN3.2.2.4 Raw data

The Drop-seq samples were whole thymus extracts from the following days of embryonic development (number of replicates in brackets): wild-type E12.5 (3), wild-type E13.5 (3), wild-type E14.5 (2), wild-type E15.5 (2), wild-type E16.5 (3), wild-type E17.5 (2), wild-type E18.5 (2), wild-type P0 (2), Rag2 knock-out E14.5 (1), Rag1 knock-out E16.5 (2). Here, replicates are thymus samples from separate animals. STAR aligner (v2.4.2), Picard tools (v1.96) and Drop-seq tools (v1.0) were used to convert raw FASTQ files into digital gene expression matrices. The commands and parameters that were used are listed in table SN3 Table 1.

The population size was measured at each time point that was also sampled in the Drop-seq experiment in wild-type mice. Note that population size determination and Drop-seq were performed on separate data sets. We used previously published population size observations at E12.5-E17.5 (Cook 2010). We obtained population size estimates of mice at the time points E18.5 and P0. At E18.5, we counted 4.9 million and 5.8 million cells per thymus lobe in lobes from separate animals. At P0, we counted 16.26 million and 15.0 million cells cells per thymus lobe in two lobes from the same animal, which were processed separately.

SN3.2.2.5 Data procession

The steps described in this section are also explained with code and plots in the jupyter notebooks that were used to perform this analysis (Supp. data 1.2, 4.3). We extracted thymic hematopoietic cells in silico from the whole thymus single-cell RNA-seq samples as described in the online methods. We performed the remainder of the workflow separately for the wild-type-only and the combined wild-type and Rag1/Rag2 knock-out data sets. We fit a diffusion map to all thymic hematopoietic cells (scanpy: k=100, knn=False)(Wolf, Angerer, and Theis 2018). We fit diffusion pseudotime with the tip of the progenitor branch as a root cell with one branching event. The resulting grouping has a group of putative myeloid and dendritic cells which we identified based on various markers (Supp. Fig. 2,3). We discarded this group to receive a subset of cells that could be allocated to the T-cell or the non-conventional lymphocyte lineages. We fit a new diffusion map and diffusion pseudotime on this "gated" data set and observed that the overall structure was not drastically changed (Supp. Fig. 6a,b), as would be expected after the removal of a small branch. Next, we adjusted the branch allocation: We kept the assignments of cells to the progenitor group (Supp. Fig. 6b,c). We split the remaining cells into T-cells and non-conventional lymphocytes

Purpose	Software Package	Command name	Free Parameters
Convert to SAM	Picard tools	FastqToSam.jar	–
Tag cell barcode	DropSeq tools	TagBamWithReadSequenceExtended	BASE_QUALITY=10, BARCODED_READ=1, NUM_BASES_BELOW_QUALITY=1, DISCARD_READ=False
Tag molecular barcode	DropSeq tools	TagBamWithReadSequenceExtended	BASE_QUALITY=10, BARCODED_READ=1, NUM_BASES_BELOW_QUALITY=1, DISCARD_READ=True
Remove low-quality barcodes	DropSeq tools	FilterBAM	–
Remove polyA tail	DropSeq tools	PolyATrimmer	MISMATCHES=0, NUM_BASES=6
Convert to FASTQ	Picard tools	SamToFastq.jar	–
Align reads to exome	STAR	–	–
Sort BAM to speed merging	Picard tools	SortSam.jar	–
Merge with tagging and alignment info	Picard tools	MergeBamAlignment.jar	INCLUDE_SECONDARY_ALIGNMENTS=False
Label read with exon	DropSeq tools	TagReadWithGeneExon	–
Screen for bead errors	DropSeq tools	DetectBeadSynthesisErrors	NUM_BARCODES=2000
Form DGE matrix	DropSeq tools	DigitalExpression	MIN_NUM_GENES_PER_CELL= 1000

SN3 Table 1: RNA-sequencing read alignment pipeline.

(NCL) based on a linear partition in the diffusion component (DC) 1 - DC2 plane (Supp. Fig. 6c). The resulting grouping contained progenitor cells, T-cells and non-conventional lymphocytes. We chose the progenitor and T-cell sets as the main branch in the branching model and chose the non-conventional lymphocytes as the side branch. Next, we had do define a branching region. Based on the DC1-DC2 plane and the diffusion pseudotime values on the main branch, we allocated all cells which were not in the progenitor group and which had a diffusion pseudotime value below $s_{max}^{branching}$ (0.25 for wild-type-only data and or as 0.2 combined wild-type and mutant data) to the branching region. Note that the different choice of $s_{max}^{branching}$ is just a result for the overall lower diffusion pseudotime values in the combined wild-type and mutant data. Lastly, we scaled the diffusion pseudotime coordinates so that both T-cell and NCL branches have the same maximum coordinate value: We performed a linear scaling of the pseudotime values in the NCL group (which lay in the interval $[s_{max}^{branching}, \max(s)]$) into the interval $[s_{max}^{branching}, \max(s_{TC})]$ where s_{TC} is the set of pseudotime observations in the T-cell group and s is the set of all pseudotime observation at this stage. Note that such a scaling does not change the ordering of cells on the side branch and does not change the ordering of non-conventional lymphocytes with respect to the branching region and the progenitor group. The scaling does therefore not change the developmental meaning of the model. We called these scaled diffusion pseudotime values "cell state" and used them throughout the manuscript. For numeric reasons, we linearly scaled the pseudotime values into the interval $[0, 0.9]$ for pseudodynamics model fitting.

We also used pseudodynamics on a linear trajectory based on a monocle2 pseudotemporal ordering. We first gated the diffusion pseudotime-based branches corresponding to non-conventional lymphocytes and putative myeloid and dendritic cells out to receive a data set that mostly contains the progenitor and T-cells (Supp. Fig. 6a). We then ran monocle2 on this data set.

SN3.2.2.6 Results

The results of this analysis are presented in the main text in section "Pseudodynamics extends previous models of T-cell maturation". We inferred the cell state space both with diffusion pseudotime (dpt) and with monocle2. The cell state space inferred with monocle2 is roughly a monotonous transform of the cell state space inferred with dpt, as shown in Supp. Fig. 12c. One can consider this monotonous dependency a sanity check that both projections quantify progress along the same transcriptional manifold. Here, we provide a detailed side-by-side comparison of the parameter fits based on dpt (shown in Fig. 3c) and monocle2 (shown in Supp. Fig. 12d-f) cell state space. In the main text, we discuss two major characteristics of the birth-rate functions which are population expansion (large positive) after beta-selection and selection (negative) after population expansion. Both are reflected in fits on dpt (Fig. 3c) and fits based on monocle2 (Supp. Fig. 12d). One can approximate the position of beta-selection in monocle2 pseudotime cell state space based on Rorc, Bcl-xL and Bcl2 expression (Supp. Fig. 11c): The resulting estimate around cell state 15 is consistent with the location of the proliferative burst identified by pseudodynamics (Supp. Fig. 12d, around cell state 15).

Both pseudodynamics fits also exhibit positive drift rates in the cell state range with negative birth-death rates that we claim to correspond to selection, which is motivates our claim that there is directed development in this region (Fig. 3c, Supp. Fig. 12b). We did not find the drift plateau before beta-selection in the the monocle2-based fits. However, monocle2 did overall not capture the details of this data set very well (Supp. Fig. 11d-k): The sequence of DN2a, Phase 2, Phase 3 and DP T-cells is not reflected in the 2D projection produced by monocle2 (Supp. Fig. 11d-k) and some DN2a and Phase 2 cells receive very high pseudotime values (Supp. Fig. 11e,f). Therefore, details such as the drift plateau may be lost in the pseudodynamics fits based on monocle2. For this reason, we chose the diffusion pseudotime ordering as the main underlying cell state space in the main text.

SN3.2.2.7 Alluvial plots

Introduction We computed alluvial plots (main text Fig. 3a) for this data set to visualize the quantitative description of cellular transitions in development provided by pseudodynamics. Alluvial plots show how the binned distribution of a population at a time point $t+1$ is related to the binned distribution at time point t. In the case of pseudodynamics, the binning is done in the cell state space. In particular, bin i at time point t is connected with bin j at time point $t+1$ via a "flow" which shows what fraction of cells in bin i at t excited bin i towards bin j and what the how much this flux contributes to the total population at bin j at $t+1$. In the context of pseudodynamics, one can visualize normalized or non-normalized flows between cell state bins (intervals).

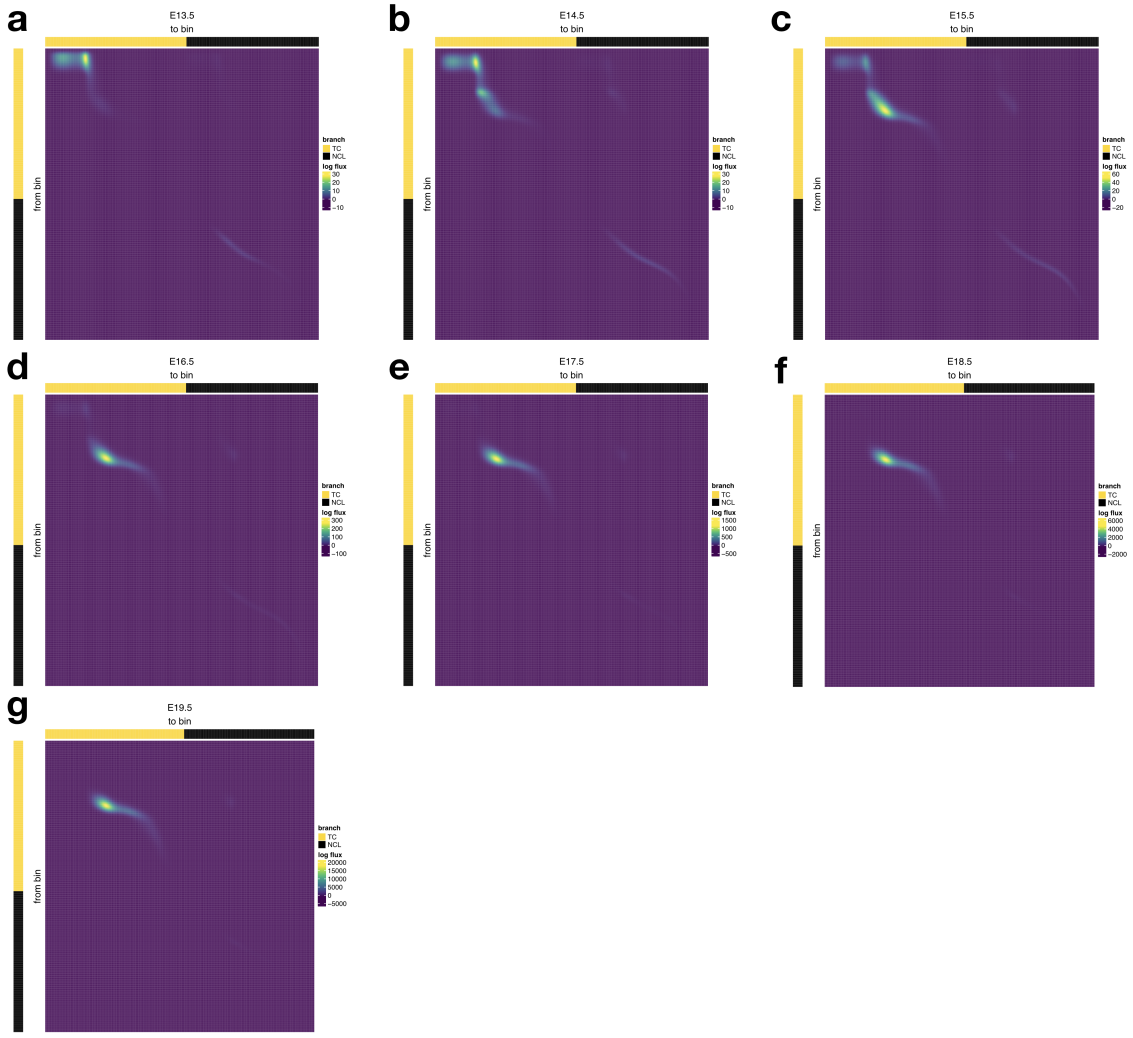
Methods To compute the flux (flow) from bin i at time point t to all bins at time point $t+1$, we simulated the PDE system forward from t to $t+1$ with initial condition, $u_{t,i}$, that has as entries the average number density of the solution at bin i in bin i, $\bar{u}_i(t) = \frac{1}{s_{i+1}-s_i}\int_{s_i}^{s_{i+1}} u(s,t)ds$ and 0 for all other bins. Here, the bins represent volumes in the finite volume approximation. As the finite volume approximation results in a linear ODE (of the form $\dot{x} = Ax$ with x the average number density in the finite volume bins and some forward matrix A), the simulated distribution across bins at $t+1$ starting from $u_{t,i}$ represents the flux from bin i. we repeated this forward simulation across bins and across time points (bins x time points simulations) to compute all fluxes at all time points

The binning has to have a large granularity for the finite volume approximation to hold but such a high resolution in the cell state space is difficult to visualize in alluvial plots. Accordingly, we pooled fluxes across sets of adjacent bins of the same size for visualization. We provided a heat map visualization of the fluxes per time point in the cell state grid used for the finite volume approximation in SN3 Figure 1.

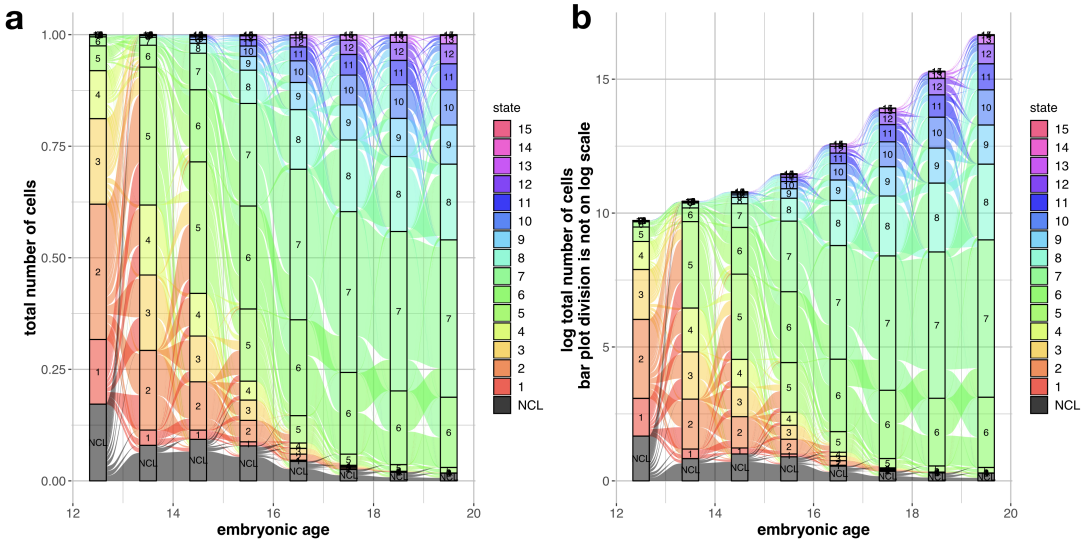
Results We visualized the normalized flows (SN3 Figure 2a) and overall population size side by side in Fig. 3a,b in the main to text to link the state-transition overview of an alluvial plot with the pseudodynamics concept of non-normalized distributions that carry total population size information. One could show a distribution across cell state bins which is scaled with the total number of cells in this system so that the height of flows and bins can directly be interpreted as a number of cells. We provided an animated visualization of such a linearly scaled alluvial plot in Supp. data 3. The linear scaling only yields interpretable visualizations if the alluvial plot is animated and the y-axis is re-scaled as a function of time as the population size grows exponentially. Here, we provide an alternative visualization in which we scale the alluvial plot with the log of the total number of cells (SN3 Figure 2b).

SN3 Figure 2a is an alluvial plot of the normalized distribution across bins. Note that here, a reduction in the height of a bin over time does not necessarily imply an absolute reduction in the number of cells in that bin, as the total growth in population size is not accounted for in this plot. SN3 Figure 2b is simply the normalized distribution and set of flows scaled with the log of the population size at that time point and mitigates this disadvantage of the normalized alluvial plot.

The heat maps in SN3 Figure 1 contain the same information as the alluvial plot, each heat maps corresponds to the flows between one pair of adjacent time points. However, the heat maps show the flows between the cell state bins in the discretization used for the finite volume approximation of the PDE. Both the heat maps and the alluvial plots show that most of the population distribution in the last three time points stems from bin 7 at the previous time point. Bin 7 contains the proliferative burst after beta-selection which dominates the population distribution. Birth-death rates become negative after the proliferative burst while the drift is non-zero (main text Fig. 3b). Bins which correspond to this interval of selection pressure can therefore only maintain their normalized distribution mass if they have influx from earlier bins. The alluvial plot therefore sum-



SN3 Figure 1: Heat map visualization of bin to bin flux by time point. The branches are color coded: black ("TC") is the T-cell lineage main branch and gold ("NCL", non-conventional lymphoid cells) is the NCL side branch. Each row and each column represent a bin the the finite volume approximation of this system, where there is one trajectory for each TC and NCL. Each heat map represents the flux from time point $tdays$ to $t + 1days$ where $t + 1$ is indicated in the panel title: (**a**): flux from E12.5 to E13.5, (**a**): flux from E13.5 to E14.5, (**c**): flux from E14.5 to E15.5, (**d**): flux from E15.5 to E16.5, (**e**): flux from E16.5 to E17.5, (**f**): flux from E17.5 to E18.5, (**g**): flux from E18.5 to P0 (19.5 days after E0).



SN3 Figure 2: Alluvial plot of flows of cells between bins on the T-cell maturation lineage. This figure is an alternative visualization of the alluvial plot presented in the main text in Fig. 3a. Each bar plot corresponds to one time point and is partitioned linearly by the fraction of cells of all cells at that time point in that bin. Accordingly, the composition of each bar plot represents the normalized distribution of cells across bins at that time point. The T-cell trajectory was divided into 15 equidistant bins in cell state (labeled 1-15), the non-conventional lymphoid cell branch was summarized to one bin (labeled NCL). The resulting 16 bins and their outflows are color coded. Outflow width represents the fraction of surviving cells transitioning into each bin at the old time point. Inflow width represents the contribution of each flow to the population size in a bin at the new time point. (**a**) The bar plots by time point were not scaled at all. The height of a bin in a bar plot is therefore the fraction of the cells in that bin at that time point (normalized distribution). (**b**) The height of the bar plots by time point were scaled to the log of simulated population size. Each bar plot is still partitioned linearly according to the normalized distribution across bins. The height of a bin in a bar plot is therefore the log of simulated overall population size multiplied with the fraction of the cells in that bin at that time point (normalized distribution).

marizes the observation of an apparent steady-state of the normalized distribution of cells across cell state at time points E17.5, E18.5, P0 (main text Fig. 2d): This apparent steady-state is due to a "source-sink"-like effect and is not due to an attractor cell state: The cell state interval around the proliferative burst is the source and population size decay (cell death) as well as efflux to higher bins create a sink phenomenon.

SN3.2.3 Proliferation rates of β-cells in the adult pancreas

SN3.2.3.1 Aim

We used pseudodynamics on this data set to compare a state- and a state- and time- dependent proliferation model.

SN3.2.3.2 Model

We chose the non-branching pseudodynamics model for this system (Sec. SN1.3.1): The linear trajectory corresponds to in vivo pancreatic β-cell maturation.

SN3.2.3.3 Implementation

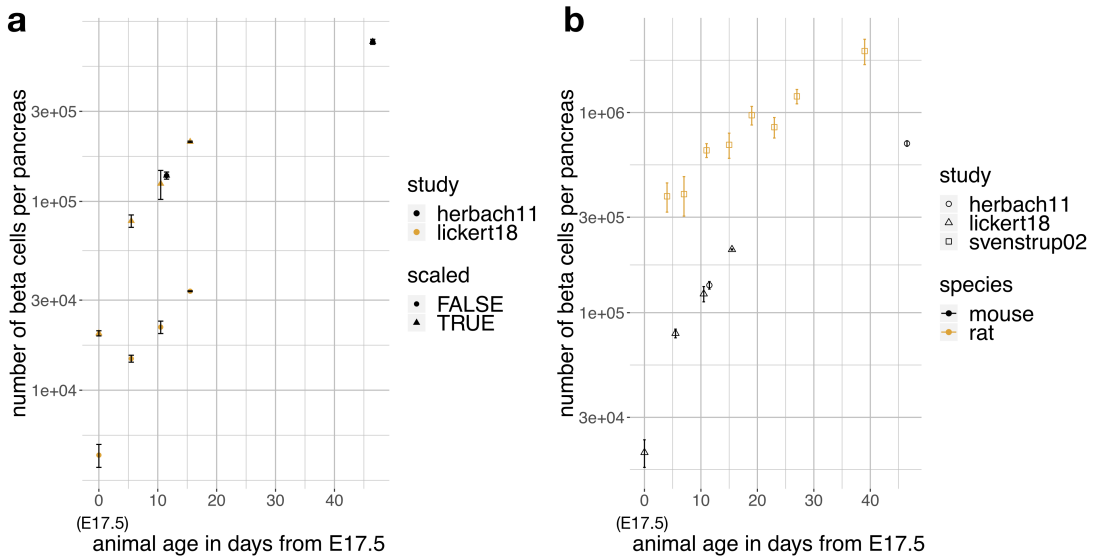
The diffusion and drift parameters, $D(s)$ and $v(s)$, are parametrized with natural cubic splines based on nine equally spaced sampling points in s. For the state-dependent growth rate, $g(s)$, we used nine equally spaced nodes in $[0, 1]$. The state and time-dependent growth rate was implemented as a product of a state- and a time-dependent growth rate, $g(s, t) = g_s(s)g_t(t)$, we chose nine equally spaced sampling points in $[0, 1]$ for $g_s(s)$ and three equally spaced sampling points in $[0, 60]$ for $g_t(t)$. For numerical reasons we employed a log-transformation of the parameters D_i and v_i, $i = 1, \ldots, 9$. To ensure positivity of $D(s)$ and $v(s)$ we computed the exponential of the spline. The drift rate is decreased to 0 in the interval $[0.9, 1]$ of cell states by using a hermite c-spline. Since we had no replicates for the cell densities, we estimated the variance together with the parameters (SN1.18) in the log-likelihood. We performed 100 multi-starts using the trust-region method implemented in `fmincon.m` and investigated the regularizations $\rho = \{0, 0.1, 1\}$. To compute a time-dependent growth rate from the estimated state-dependent growth rate, we simulated the population distribution u on a grid in time with 300 grid points equally spaced in $[0, 61.5]$. The mean growth rate at a time point t, $g_s(t)$, was then computed as

$$g_s(t) = \int_{s=0}^{s_{max}} g(s)u(s, t)ds. \tag{SN3.1}$$

SN3.2.3.4 Raw data

The single-cell RNA-seq data consists of one sample of each of the seven time points (E17.5, P0, P3, P9, P15, P18, P60) (Qiu et al. 2017) (SN3 Figure 4a).

We collected total β-cell population estimates from immunostained pancreatic sections of embryonic and neonatal mice from E17.5, P4, P9, P10 and P14 (SN3 Figure 4b) by extrapolating the mean number of β-cells per section to the total number of cell per pancreas based on the section volume and the total pancreas volume. We counted β-cells in 6 (E17.5), 4 (P4), 4 (P9) and $\geq$3 (P14) sections per animal in three animals each. We counted the following number of β-cells per animal across all sections: 703, 284 and 644 in E17.5, 1119, 941 and 879 in P4, 1205, 797 and 1236 in P9 and 1100, 1134 and 1110 in P14. We scaled the these numbers of β-cells to the expected number of β-cells per pancreas (see below). Secondly, we used previously published measurements for P11 and P45 (Herbach et al. 2011).



SN3 Figure 3: Merging pancreatic β-cell total population size data in neonatal mice from two studies. **(a)** Total number of β-cells by age in mice. Shown are the mean total number of β-cell per pancreas in mice from age E17.5 onwards at days E17.5 (n=3 replicates), P4 (n=3 replicates), P9 (n=3 replicates), P10 (n=8 replicates), P14 (n=5 replicates) and P45 (n=3 replicates) with one standard error around the mean as error bars. We collected data on E17.5, P4, P9, P10 and P14 within this study (lickert18). We used previous estimates(Herbach et al. 2011) (herbach11) for days P10 and P45. Replicates were separately measured in one unique animal per replicate. The lickert18 data are shown as raw data (scaled: FALSE) and scaled to the herbach11 data (scaled: TRUE). **(b)** Total number of β-cells by age in mice and rats. Shown are mouse (as detailed in panel a) and rat time course measurement of the total number of $beta$-cells per pancreas. The rat measurements are based on $n = 6$ replicates per time point. The mean number of $beta$-cells is shown with one standard error around the mean as error bars.

We computed a scaling factor between a linear extrapolation of the log of the population size mean at P10 of our data and our data and a previously published log of the population size mean at P10(Herbach et al. 2011). This scaling factor was $\frac{\log \mu^{N,herbach11}(t=P10)}{\log \mu^{N,lickert18}(t=P10)} = 1.175$. We scaled all our measurements in log space with this scaling factor to receive one homogeneous data set with observations at E17.5, P4, P9, P10, P14 and P45 (SN3 Figure 3a). We believe that our measurements capture the correct relative number of β-cells per pancreas across time. However, Herbach *et al.*(Herbach et al. 2011) performed more extensive cell counting so that we believe that their measurements capture the correct absolute numbers and accordingly scaled our data to this reference. The scaling in log space assumes that our β-cell measurement deviate from the data measured by Herbach *et al.* by a linear scaling factor in log space.

We fit the population size observations in log space with pseudodynamics. To transform the observed distribution to the log space, we computed the mean (N_t^{obs}) and standard error ($\sigma_t^{N,obs}$) of the untransformed data and computed the mean (LN_t) and and standard deviation (σ_t^{LN}) of a log-normal distribution based on these statistics:

$$\begin{aligned} \sigma_t^{LN} &= \sqrt{\log\left(1 + \frac{(\sigma_t^{N,obs})^2}{(N_t^{obs})^2}\right)} \\ LN_t &= \log\left(\frac{N_t^{obs}}{\sqrt{1 + \frac{(\sigma_t^{N,obs})^2}{(N_t^{obs})^2}}}\right) \end{aligned} \tag{SN3.2}$$

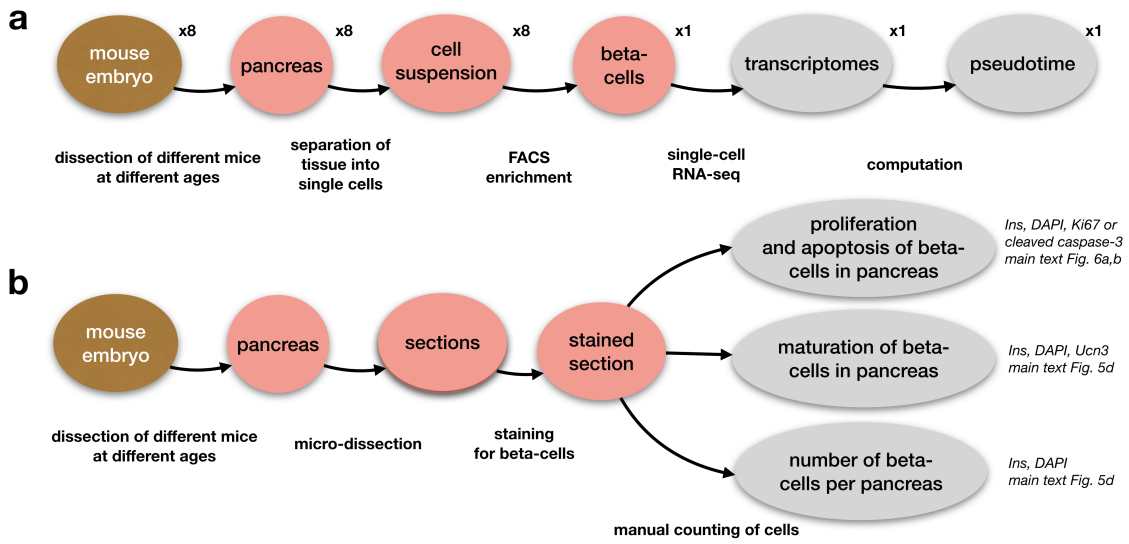
We note that the population size was estimated at different time points to the ones at which the population density was observed with single-cell RNA-seq. Both measurements do however lie in the same interval and can therefore be used in pseudodynamics.

We added β-cell counts per pancreas from an independent study in rats as a reference. Svenstrup *et al.*(Svenstrup et al. 2002) estimated the total number β-cells in the pancreas of rats at multiple developmental time points based on manual counting of β-cells in immunostained pancreatic sections (SN3 Figure 3b). The total number of β-cells per rat was estimated at day 1, 4, 8, 12, 16, 20, 24, 36, 150 and 300 after birth with six replicates each. Replicates were separately measured in one unique animal per replicate. The β-cell population size measurements are presented in Fig. 1 E,F in(Svenstrup et al. 2002) but the raw data was not supplied. Accordingly, we extracted the mean and standard error estimates of time points 1, 4, 8, 12, 16, 20, 24 and 36 directly from the plots in Fig. 1E in(Svenstrup et al. 2002). According to the Carnegie stage comparison(*Carnegie Stage Comparison - Embryology* n.d.), rats are 1.5 days delayed in development compared to mice around birth. Therefore, we added 1.5 days to the time coordinate of all population size estimators from rat to map them to the corresponding stage in mouse development. The rat data presented in SN3 Figure 3b only serves as reference to validate the global population size trend that we recorded in mice. We only use the mice data for model fitting.

SN3.2.3.5 Data processing

The steps described in this section are also explained with code and plots in the jupyter notebooks that were used to perform this analysis (Supp. data 1.4). We fit a diffusion map to all cells (scanpy: k=30, knn=False). We fit diffusion pseudotime to this data set with the cell extremal cell in diffusion component 1 on the side of the manifold dominated by the E17.5 sample as a root cell.

Details on the preprocessing of the data can be found in the scripts associated with this analysis.



SN3 Figure 4: Experimental design: β-cell development. **(a)** Experimental design to measure cell state distribution with single-cell RNA-seq. **(b)** Experimental design to estimate population size, maturation status, apoptosis rates and proliferation rates from stained pancreatic tissue sections.

SN3.2.3.6 Results

The results of this analysis are presented in the main text in section "Pseudodynamics identifies time-dependent proliferation of pancreatic β-cells as a state dependent effect".

SN3.2.4 A two-compartment model for proliferation rates of β-cells in the adult pancreas

SN3.2.4.1 Aim

We used pseudodynamics on this data set to compare a time- and state-dependent and an only state-dependent proliferation model based on a ordinary differential equation model with only two compartments. Here, state-dependence means that the parameters are compartment specific.

SN3.2.4.2 Model

The two-compartments represent immature and mature pancreatic β-cells, defined based on the presence of Ucn3 in immunostainings of tissue sections. The number of cells in the immature compartment is modeled by m_1 and the number of cells in the mature compartment by m_2,

$$\begin{aligned} \dot{m}_1 &= -d_1(t)m_1 + d_2(t)m_2 + g_1(t)m_1 \\ \dot{m}_2 &= d_1(t)m_1 - d_2(t)m_2 + g_2(t)m_2, \end{aligned} \tag{SN3.3}$$

with transition rate from m_1 to m_2, $d_1(t)$, transition rate from m_2 to m_1, $d_2(t)$, and growth rate $g_i(t)$ for compartment i. We compared three different parameterizations.

P1 Constant rates (woTime): $d_1(t) = d_{1,const}$, $d_2(t) = d_{2,const}$, $g_1(t) = g_{1,const}$ and $g_2(t) = g_{2,const}$;

P2 Constant transition and time-dependent growth rates parameterized using natural cubic splines (wTime1): $d_1(t) = d_{1,const}$, $d_2(t) = d_{2,const}$, $g_1(t) = \text{ncs}\,(t|\vec{\alpha}_{g,1})$ and $g_2(t) = \text{ncs}\,(t|\vec{\alpha}_{g,2})$

P3 Time dependent growth and transition rates parameterized using natural cubic splines (wTime2): $d_1(t) = \text{ncs}\left(t|\vec{\alpha}_{d,1}\right)$, $d_2(t) = \text{ncs}\left(t|\vec{\alpha}_{d,2}\right)$, $g_1(t) = \text{ncs}\left(t|\vec{\alpha}_{g,1}\right)$ and $g_2(t) = \text{ncs}\left(t|\vec{\alpha}_{g,2}\right)$.

All parameters are state-dependent in all three parameterizations as we allow one parameter per compartment.

We allowed a parameter for backwards motion of cells from the mature to the progenitor compartment in the discrete model. We did so for two reasons:

Firstly, diffusion in a continuous model corresponds to backwards and forwards rates in a chain discretization of the continuous coordinate (such as in finite volumes or finite differences). Accordingly, one could not account for diffusion in a discrete model if one did not allow backwards rates.

Secondly, there is conflicting evidence from lineage tracing experiments about the existence of a developmental transition from mature β-cells (*Ins*$^+$*Ucn3*$^+$) to immature β-cells (*Ins*$^+$*Ucn3*$^-$):

Van der Meulen *et al.* studied β-cell maturation based on the marker Ucn3 via lineage tracing *in vivo* in mice(Meulen et al. 2017). In particular, they found that there are small fluxes of transdifferentiation of α-cells via "virgin" β-cells (*Ins*$^+$*Ucn3*$^-$) to mature β-cells (*Ins*$^+$*Ucn3*$^+$). The number of β-cells with α-cell lineage label is less than 0.5% of the total β-cell number at two days and four month after the lineage labeling in mice at two months of age (van der Meulen *et al.*(Meulen et al. 2017), Fig. 6F). Therefore, we argue that the transdifferentiation flux of α-cells to β-cells can be neglected without great loss of accuracy for the population size model. Van der Meulen *et al.* also traced mature β-cells and did not observe mature (*Ins*$^+$*Ucn3*$^+$) lineage-labeled "virgin" β-cells (*Ins*$^+$*Ucn3*$^-$) (van der Meulen *et al.*(Meulen et al. 2017), Fig. 1K): They did not find evidence for a transition from *Ins*$^+$*Ucn3*$^+$ to *Ins*$^+$*Ucn3*$^-$ cells.

Talchai *et al.* studied β-cell dedifferentiation in mice subjected to physiological stress(Talchai et al. 2012). Talchai *et al.* reported evidence which suggests that mature β-cells can dedifferentiate to progenitor β-cells (*Ins*$^-$*Ngn3*$^+$) and other cell types present in the pancreas. These *Ins*$^-$*Ngn3*$^+$ progenitor β-cells are thought to be progenitors of *Ins*$^+$*Ucn3*$^-$ immature β-cell(Qiu et al. 2017)(Petersen et al. 2017). Accordingly, the results from Talchai *et al.* suggest that β-cell dedifferentiation might cause a transition of mature β-cells (*Ins*$^+$*Ucn3*$^+$) to immature β-cells (*Ins*$^+$*Ucn3*$^-$).

The backward rate parameter has the interpretation of diffusion in the limit of many compartments and cannot be clearly rejected based on evidence from lineage tracing experiments. Therefore, we included this parameter in the model. We would like to note that the addition of a dedifferentiation rate parameter to the discrete compartment model does not force a dedifferentiation flux in the model, it gives the model the freedom to fit such a flux during parameter estimation if this helps explaining the data.

SN3.2.4.3 Implementation

The likelihood consists of two parts, a part describing the fit to the overall log population size and a part describing the fit to the distribution of cells between the two states,

$$\log L(\theta) = \left(\sum_{t \in T^N} \log L\left(LN_t|\theta, \sigma_t^{LN)}\right) \right) + \left(\sum_{t \in T^N} \log L\left(p_{2,t}|\theta, \sigma_t^p\right) \right). \tag{SN3.4}$$

The likelihood for the overall log population size is a normal distribution

$$L(\log(N_t)|\theta, \sigma_t^{LN}) = \mathcal{N}\left(LN_t \middle| \mu = \mu^{LN}\left(t, \theta\right), \sigma^2 = \left(\sigma_t^{LN}\right)^2\right) \tag{SN3.5}$$

with LN_t and σ_t^{LN} as described in eq. SN3.2 and

$$\mu^{LN}(t,\theta) = \log(m_1(t) + m_2(t)) \tag{SN3.6}$$

where $\mu^{LN}(t,\theta)$ is the logarithm of the overall population size predicted by the model for the parameters θ at time point t.

For the fit to the distribution of cells, we consider the fraction of cells in the mature compartment, $p_{2,t}$. Similar to the fraction of cells on each branch above, we assume a normally distributed measurement noise and the likelihood

$$L(p_{2,t}|\theta,\sigma_t^p) = \mathcal{N}\left(p_{2,t}\middle|\mu = \mu^p(t,\theta), \sigma^2 = (\sigma_t^p)^2\right) \tag{SN3.7}$$

with

$$\sigma_t^p = \sigma_t^{p,obs}/\sqrt{R_t} \tag{SN3.8}$$

where $\sigma_t^{p,obs}$ is the observed standard deviation of the fraction of mature cells at time point t and R_t the number of replicates of the fraction measurements at time point t and

$$\mu^p(t,\theta) = \frac{m_2(t)}{m_1(t)+m_2(t)} = \frac{m_2(t)}{N(t)} \tag{SN3.9}$$

where $\mu^p(t,\theta)$ the fraction of mature cells predicted by the model for the parameters θ at time point t.

We performed 100 multi-starts using the interior point algorithm implemented in `fmincon.m` and investigated the regularizations $\rho = \{0, 1, 10\}$. The confidence intervals were computed using likelihood profiles (SN1.29).

SN3.2.4.4 Raw data

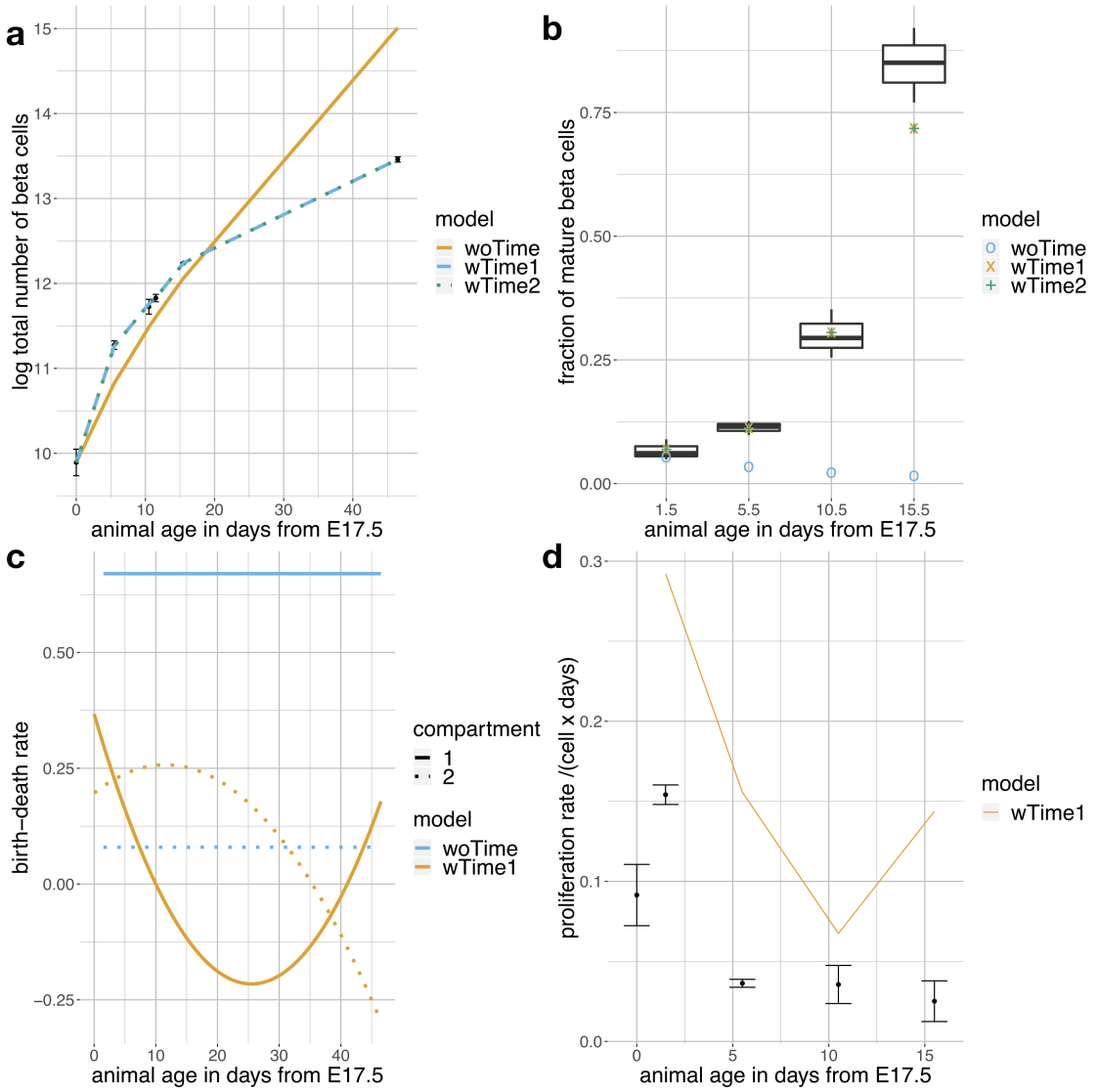
The single-cell data are number of *Ins*$^+$*Ucn3*$^-$ and *Ins*$^+$*Ucn3*$^+$ cells in tissue section at days P0, P4, P9 and P14. We defined the set of Ins+ cells as pancreatic β-cells and grouped these into two compartments: Immature β-cells (*Ins*$^+$*Ucn3*$^-$) and mature β-cells (*Ins*$^+$*Ucn3*$^+$) (main text Fig. 5a,b). Based on the total number of these cells across multiple sections per animal, we computed the fraction of mature and immature β-cells by time point. We performed this procedure in three animals per time point, counting a total of 1917 cells, 2674 cells, 2198 cells and 3638 cells respectively for the time points P0, P4, P14 and P14. We computed the mean and standard deviation of the fraction of mature β-cell per compartment by time point across animals.

We used the same population size data as described for the continuous model (Sec. SN3.2.3.4).

The first population size measurement (E17.5) and the first single cell maturation quantification (P0) were recorded at 1.5 days apart. We made the simplifying assumption that the single cell maturation is the same at E17.5 as at P0 so that we did not have to co-estimate the initial condition during pseudodynamic model fitting: We set the initial condition at E17.5 to the measurement at P0 and removed the measurement at P0 from the set of measurements on which the likelihood is evaluated.

SN3.2.4.5 Results

We compared the three models presented in sec. SN3.2.4.2 with likelihood ratio tests and by qualitatively evaluating their fits to the data. The constant model without time dependence (woTime) fits the population size and the population distribution data much worse than both other models with time-dependent parameters even if it is completely unconstrained with a regularization parameter of 0 (SN3 Figure 5a,b). The model woTime is nested in both wTime1 and wTime2. We



SN3 Figure 5: A two-compartment ordinary differential equation model for β-cell maturation requires time-dependent parameters to fit the measurements. The model acronyms woTime, wTime1 and wTime2 are introduced in sec. SN3.2.4.2. The model fits shown are based on a regularization parameter of 1 for wTime1 and wTime2, and based on a regularization parameter of 0 for woTime, to allow woTime to be unconstrained. **(a)** Population size fits by model. The observed proliferation rates are shown in black with one standard deviation around the mean as error bars. **(b)** Population distribution fits by model, presented is the fraction of mature cells in the two-compartment model. The observations are shown as boxplots with $n = 3$ ($t = 1.5$), $n = 3$ ($t = 5.5$), $n = 3$ ($t = 10.5$) and $n = 3$ ($t = 15.5$) replicates. Replicates were separately measured in one unique animal per replicate. The centre of each boxplots is the sample median, the whiskers extend from the upper (lower) hinge to the largest (smallest) data point no further than 1.5 times the interquantile range from the upper (lower) hinge. **(c)** Birth-death rate fits by model and compartment. **(d)** Average birth-death rate per time point by model and observed proliferation rates by time point with one standard deviation around the mean as error bars. The birth-death rates at a given time point are computed as the convolution of the simulated population distribution at that time point with the parameter fit, both functions of cell state.

performed model selection with a likelihood-ratio test between woTime and wTime1 and between woTime and wTime2. In both cases, the difference in likelihood between the model without time-dependent parameters (woTime) and the model with time-dependent parameters (wTime1, wTime2) is significant, which is in line with the bad fits of the model without time-dependence (SN3 Figure 5a,b). These results suggest that it is necessary to include time-dependent parameters in the model to fit β-cell maturation if a discrete two-compartment model is used to describe the biological process.

Both models with time-dependent parameters are very similar in terms of their fits (SN3 Figure 5a,b). The more complex model wTime2 is indeed not significantly better than the model wTime1 when compared with a likelihood ratio test. This result suggests that time-dependent transition rates are not necessary to fit β-cell maturation data if the birth-death rates are allowed to vary with time.

As also presented in the main text in Fig. 6c for the continuous model, we can also compare the computed average proliferation rates per time point with the observed rates. The observed and predicted rates are globally similar (SN3 Figure 5a,b) suggesting that this two-compartment model can indeed explain the global behavior of the system.

SN3.2.4.6 Discussion

We showed here that it is necessary to assume time-dependent birth-death rates in a two-compartment ordinary differential equation model for β-cell maturation to fit the data presented here. We showed in the main text in sec. "Pseudodynamics identifies time-dependent proliferation of pancreatic β-cells as a state dependent effect" that it is not necessary to assume time-dependent birth-death rates in a continuous PDE model for β-cell maturation. These two results highlight our conclusion that the necessity to assume time-dependent parameters to explain the observed time-dependence of proliferation rates in β-cell maturation depends on the description of the underlying cell state: The time-dependent birth-death was that was found to be necessary in the two-state maturation system was not necessary in the continuous system which suggests that the time-dependence that was found here might be a discretization artifact. This artefact results from the information content reduction inherent in the discretization of the continuous distribution in cell state space. The discretization lies in the usage two-stage surface marker-based discretization of a continuous 1D pseudotime reconstruction. Accordingly, one has to be careful to assume biological regulation mechanisms that support time-dependent parameters if the developmental system can be described as a continuous trajectory.

Bibliography

1. Luzyanina, T., Roose, D. & Bocharov, G. Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data. *J. Math. Biol.* **59**, 581–603 (2009).
2. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
3. Stapor, P. *et al.* PESTO: Parameter EStimation TOolbox. *Bioinformatics* **34**, 705–707 (2018).
4. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. (SIAM, 2005).
5. Hross, S. & Hasenauer, J. Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinformatics* **32**, 2321–2329 (2016).
6. Georgii, H.-O. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. (Walter de Gruyter GmbH & Co KG, 2015).
7. Lampariello, F. On the use of the Kolmogorov–Smirnov statistical test for immunofluorescence histogram comparison. *Cytometry A* (2000).
8. Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).
9. Hug, S. *et al.* High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. *Math. Biosci.* **246**, 293–304 (2013).
10. Fröhlich, F., Theis, F. J., Rädler, J. O. & Hasenauer, J. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics* **33**, 1049–1056 (2017).
11. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).
12. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

13. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
14. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
15. Cook, A. M. Proliferation and lineage potential in fetal thymic epithelial progenitor cells. (2010).
16. Qiu, W.-L. *et al.* Deciphering Pancreatic Islet β Cell and α Cell Maturation Pathways and Characteristic Features at the Single-Cell Level. *Cell Metab.* **27**, 702 (2018).
17. Herbach, N., Bergmayr, M., Göke, B., Wolf, E. & Wanke, R. Postnatal development of numbers and mean sizes of pancreatic islets and beta-cells in healthy mice and GIPR(dn) transgenic diabetic mice. *PLoS One* **6**, e22814 (2011).
18. Svenstrup, K., Skau, M., Pakkenberg, B., Buschard, K. & Bock, T. Postnatal development of beta-cells in rats. Proposed explanatory model. *APMIS* **110**, 372–378 (2002).
19. van der Meulen, T. *et al.* Virgin Beta Cells Persist throughout Life at a Neogenic Niche within Pancreatic Islets. *Cell Metab.* **25**, 911–926.e6 (2017).
20. Talchai, C., Xuan, S., Lin, H. V., Sussel, L. & Accili, D. Pancreatic β cell dedifferentiation as a mechanism of diabetic β cell failure. *Cell* **150**, 1223–1234 (2012).
21. Petersen, M. B. K. *et al.* Single-Cell Gene Expression Analysis of a Human ESC Model of Pancreatic Endocrine Development Reveals Different Paths to β-Cell Differentiation. *Stem Cell Reports* **9**, 1246–1261 (2017).
22. Carnegie Stage Comparison - Embryology. Available at: https://embryology.med.unsw.edu.au/embryology/index.php/Carnegie_Stage_Comparison. (Accessed: 12th January 2018)
23. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
24. Wolf, F. A. *et al.* Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* (2017).

25. Yui, M. A. & Rothenberg, E. V. Developmental gene networks: a triathlon on the course to T cell identity. *Nat. Rev. Immunol.* **14**, 529–545 (2014).
26. Kernfeld, E. M. *et al.* A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity* (2018). doi:10.1016/j.immuni.2018.04.015
27. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

22. Carnegie Stage Comparison - Embryology. Available at: https://embryology.med.unsw.edu.au/embryology/index.php/Carnegie_Stage_Comparison. (Accessed: 1st December 2018)