Scalable inference of cell differentiation networks in gene therapy clonal tracking studies of haematopoiesis

Del Core, Luca; Pellin, Danilo; Wit, Ernst C.; Grzegorczyk, Marco A.

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btad605

Publication date:
2023

Bioinformatics, 2023, **39(10)**, btad605
https://doi.org/10.1093/bioinformatics/btad605
Advance Access Publication Date: 29 September 2023
Original Paper

OXFORD

Systems biology

Scalable inference of cell differentiation networks in gene therapy clonal tracking studies of haematopoiesis

Luca Del Core [1,2,*], **Danilo Pellin** [3], **Ernst C. Wit** [1,4,†], **Marco A. Grzegorczyk** [1,†,*]

[1]University of Groningen – Bernoulli Institute, 9747AG Groningen, The Netherlands
[2]University of Nottingham – School of Mathematical Sciences, Nottingham NG72RD, United Kingdom
[3]Harvard Medical School, Boston, MA 02115, United States
[4]Università della Svizzera italiana – Institute of Computing, 6962 Lugano, Switzerland

*Corresponding authors. University of Nottingham – School of Mathematical Sciences, Nottingham NG72RD, United Kingdom.
E-mail: luca.delcore@nottingham.ac.uk (L.D.C.); University of Groningen – Bernoulli Institute, 9747AG Groningen, The Netherlands.
E-mail: m.a.grzegorczyk@rug.nl (M.A.G.)

†Equal contribution.

Associate Editor: Inanc Birol

Abstract

Motivation: Investigating cell differentiation under a genetic disorder offers the potential for improving current gene therapy strategies. Clonal tracking provides a basis for mathematical modelling of population stem cell dynamics that sustain the blood cell formation, a process known as haematopoiesis. However, many clonal tracking protocols rely on a subset of cell types for the characterization of the stem cell output, and the data generated are subject to measurement errors and noise.

Results: We propose a stochastic framework to infer dynamic models of cell differentiation from clonal tracking data. A state-space formulation combines a stochastic quasi-reaction network, describing cell differentiation, with a Gaussian measurement model accounting for data errors and noise. We developed an inference algorithm based on an extended Kalman filter, a nonlinear optimization, and a Rauch-Tung-Striebel smoother. Simulations show that our proposed method outperforms the state-of-the-art and scales to complex structures of cell differentiations in terms of nodes size and network depth. The application of our method to five *in vivo* gene therapy studies reveals different dynamics of cell differentiation. Our tool can provide statistical support to biologists and clinicians to better understand cell differentiation and haematopoietic reconstitution after a gene therapy treatment. The equations of the state-space model can be modified to infer other dynamics besides cell differentiation.

Availability and implementation: The stochastic framework is implemented in the R package Karen which is available for download at https://cran.r-project.org/package=Karen. The code that supports the findings of this study is openly available at https://github.com/delcore-luca/CellDifferentiationNetworks.

1 Introduction

Haematopoiesis is the process responsible for maintaining the number of circulating blood cells that are undergoing continuous turnover. This process has a tree-like structure with haematopoietic stem cells (HSCs) at the root node (Cooper and Adams 2023). Each cell division gives rise to progeny cells that can retain the properties of their parent cell (self-renewal) or differentiate, thereby "moving down" the haematopoietic tree. As the progeny cells move further away from the HSCs, their pluripotent ability is increasingly restricted. Clarifying how HSCs differentiate is essential for understanding how they attain specific functions and offers the potential for therapeutic manipulation (Kawamoto *et al.* 2010). Several mathematical models have been proposed to describe haematopoiesis *in vivo*. One of the first stochastic models of haematopoiesis was introduced in the early 1960s suggesting that it is the population as a whole that is regulated rather than individual cells that behave stochastically (Till *et al.* 1964).

Since then, many other works aimed at describing the dynamics of haematopoiesis as a stochastic process. For example, in 2002 a single cell-based stochastic model was proposed to describe the evolution of HSCs as a result of switching between a nonproliferating and a proliferating environment, with transition probabilities, leading to a microenvironment-dependent competition of cells in a stochastic sense (Roeder and Loeffler 2002). By simulating several growth scenarios, the authors were able to describe stem cell kinetics, individual clone tracking, fluctuating clonal contribution, and to compare the theoretical results with experimental findings (Roeder *et al.* 2005). More recently, various studies analysed data generated by advanced lineage tracing protocols to calibrate novel mathematical models of HSCs differentiation. For example, in 2001 an inference method of a two-compartment hidden Markov model of haematopoiesis was proposed (Catlin *et al.* 2001). The model parameters were calibrated on time-series of cellular binary markers from a hybrid cats study. Since data from one of the two compartments were not observed, the parameters were estimated by solving the Kolmogorov forward equations and using a nonlinear least squares inference approach. Later in

Received: 9 December 2022; **Revised:** 15 July 2023; **Editorial Decision:** 20 September 2023; **Accepted:** 28 September 2023

2010 a stochastic model of haematopoiesis was proposed to keep track of HSCs differentiation in multiple cell compartments (Dingli and Pacheco 2010): the cells move within a single-branch structure with unknown probability from the smallest bone marrow compartment, housing active HSCs, up to the largest compartment hosting cells that are leaving the bone marrow to enter the circulation. The parameters of this model were calibrated on granulopoiesis data. Subsequently, in 2019 a multidimensional continuous-time Markov model was developed to describe the rates of cell differentiation in a hierarchical fashion (Pellin *et al.* 2019). The likelihood of this model is derived from a local linear approximation of the biochemical Master equation, and the parameters are estimated via a penalized least squares method from clonal tracking data. In late 2019, another stochastic process of haematopoiesis was formulated as a continuous-time, multi-type branching processes, whose parameters were estimated using moment-based techniques from cellular barcoding data (Xu *et al.* 2019). More recently in 2020, a Bayesian networks framework has been proposed to describe cell differentiation (Di Serio *et al.* 2020): the process of haematopoiesis was described through different levels of cell differentiation and the corresponding parameters were estimated using clonal tracking data from gene therapy clinical trials. Some of these methods are able to take into account missing cell types, such as those that are difficult to collect from the bone marrow (Catlin *et al.* 2001, Xu *et al.* 2019), but none considers the bias provided by false-negative clonal tracking errors. The state-of-the-art methods assume that missing clone observations correspond to minimal clones and set the corresponding counts to zero. But this hypothesis is too restrictive, because it does not take into account other technical sources of false-negative errors, such as low-informative sample replicates and threshold detection failure (Kim *et al.* 2019). Besides, it has also been shown that false-negative errors strongly depend on calling pipeline parameters, as well as read coverage (Bobo *et al.* 2016).

To overcome the limitations of the existent approaches, we propose a novel stochastic framework aimed at investigating mechanistic models of cell differentiation from clonal tracking data while cautiously treating all the undetected values as non-measurable states. More precisely, we model cell differentiation using a continuous-discrete state-space formulation including a system of Itô-type stochastic differential equations (SDE) describing clonal dynamics, coupled with a measurement model that links the sparse and noisy corrupted measurements to the underlying process' states. In Section 2, we provide a formal definition of our modelling approach along with an expectation–maximization (E–M) algorithm, based on extended Kalman filtering (EKF) and Rauch–Tung–Striebel (RTS) smoothing, to infer the unknown parameters. In Section 3, we extensively test the method on several simulation studies including a direct comparison with the prior art and we apply our framework to five *in vivo* high dimensional clonal tracking datasets, comparing four biologically plausible models of cell differentiation. In Section 4, we discuss our results from both a methodological and biological perspective.

2 Materials and methods

A concise graphical representation of our proposed framework of Kalman reaction networks (Karen) is shown in Fig. 1. The input consists of a clonal tracking dataset, including clone-specific information on the number of cells

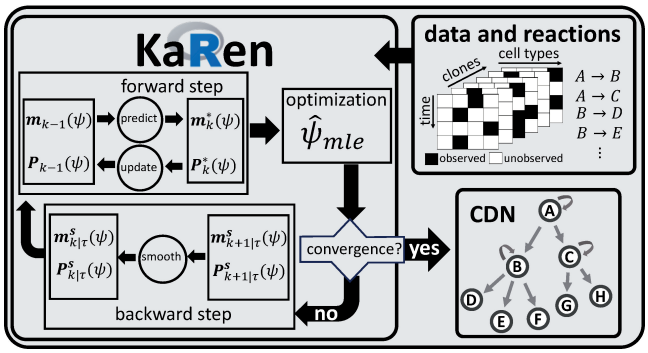

Figure 1. Analysis' flowchart: a clonal tracking dataset and the biochemical reactions (top-right) are the input of our framework Karen (left). It mainly consists of three parts: a filtering step, a maximum likelihood step, and a smoothing step. These steps are iterated until convergence is reached. The inferred cell differentiation network is returned (bottom-right).

generated for each lineage over time, and a set of biochemical reactions coding for cell duplication, cell death, and cell differentiation. Inference is done via an E–M algorithm. The E-step is based on a Kalman filter/smoother, aimed at estimating the state variables given the parameters inferred from the M-step. While in the M-step, a nonlinear optimization method updates the unknown parameters given the states estimated by the E-step. Both steps are iterated until convergence is reached. The inferred cell differentiation network is returned as the main output. More details on the input syntax can be found in the documentation of our R package Karen, which we published along with this article. The following subsections provide details on the state-space formulation of cell differentiation and the E–M algorithm.

2.1 A stochastic model for cell differentiation

Here, we assume that cell duplication, cell death and cell differentiation can take place from time t to time $t + \Delta t$ in a combinatorial number of ways directly proportional to the cell counts at time t and the corresponding rate parameters. This hypothesis is equivalent to the chemical law of mass action (Érdi and Tóth 1989) whereby cell duplication, cell death, and cell differentiation are biochemical reactions that can be properly described by stochastic quasi-reaction networks. Consistently with the definition of a chemical reaction network of Supplementary Section S1, we consider a Markov process

$$\boldsymbol{x}_t = (x_{1t}, \ldots, x_{nt}), \tag{1}$$

of one clone and n cell types evolving in a time interval $(t, t + \Delta t)$ according to a set of K distinct biochemical reactions whose net-effect vectors $\{\boldsymbol{\nu}_k\}_{k=1}^K$ and hazard functions $\{h_k(\boldsymbol{x}_t, \boldsymbol{\theta})\}_{k=1}^K$ are defined as

$$\boldsymbol{\nu}_k = \begin{cases} (\cdots 1 \cdots)' \\ \quad i(k) \\ (\cdots -1 \cdots)' \\ \quad i(k) \\ (\cdots -1 \cdots 2 \cdots)' \\ \quad i(k) \quad j(k) \end{cases} \qquad h_k(\boldsymbol{x}_t, \boldsymbol{\theta}) = \begin{cases} x_{i(k)t}\alpha_{i(k)} \\ x_{i(k)t}\delta_{i(k)} \\ x_{i(k)t}\lambda_{i(k)j(k)} \end{cases}, \tag{2}$$

where $i(k)$ and $j(k)$ are the cell types possibly involved in the kth reaction, and $j(k) \in \mathcal{O}(i(k))$, where

$$\mathcal{O}(i) = \{j|\lambda_{ij} > 0\} \tag{3}$$

is called the offspring set of cell type i. The hazard functions include a linear growth term $x_{i(k)t}\alpha_{i(k)}$ with duplication rate $\alpha_{i(k)} > 0$, a linear term $x_{i(k)t}\delta_{i(k)}$ for cell death with a death rate $\delta_{i(k)} > 0$, and a linear term $x_{i(k)t}\lambda_{i(k)j(k)}$ for cell differentiation from cell lineage $i(k)$ to cell lineage $j(k)$ with rate $\lambda_{i(k)j(k)} > 0$. The vector parameter

$$\boldsymbol{\theta} = \left(\alpha_1 \cdots \alpha_n, \delta_1 \cdots \delta_n, \boldsymbol{\lambda}'_{1\mathcal{O}(1)} \cdots \boldsymbol{\lambda}'_{n\mathcal{O}(n)}\right)', \tag{4}$$

appearing in the hazard functions, includes all the dynamic parameters, where $\boldsymbol{\lambda}'_{i\mathcal{O}(i)}$ is the vector of all the differentiation rates from cell lineage i to its offspring set $\mathcal{O}(i)$. Finally, we define the net-effect matrix and the hazard vector as

$$\begin{aligned} &\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_K] \in \mathbb{Z}^{n\times K}, \\ &\boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{\theta}) = \left(h_1(\boldsymbol{x}_t, \boldsymbol{\theta}), \ldots, h_K(\boldsymbol{x}_t, \boldsymbol{\theta})\right)', \end{aligned} \tag{5}$$

and, as probabilistic assumption, we use the Kolmogorov-forward ODEs

$$\begin{aligned} &\frac{\partial P(\boldsymbol{x}, t)}{\partial t} = -\nabla_x\{\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta})P(\boldsymbol{x};t)\} + \frac{1}{2}\nabla^2_x\{\boldsymbol{\beta}(\boldsymbol{x};\boldsymbol{\theta})P(\boldsymbol{x};t)\}, \\ &\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{V}\boldsymbol{h}(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{\beta}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{V}\begin{bmatrix} h_1(\boldsymbol{x}_t;\boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_K(\boldsymbol{x}_t;\boldsymbol{\theta}) \end{bmatrix}\boldsymbol{V}', \end{aligned} \tag{6}$$

obtained from a continuous approximation of the Master equation (see details in Supplementary Section S2).

2.2 State-space formulation

Since the aim of this work is to calibrate the parameters of the continuous-time stochastic model defined by Equations (1–6) on clonal tracking data that have been collected at discrete time points, we use a continuous-discrete state-space model (Del Core 2023). In this formulation the dynamic component is the system of Itô-type SDEs

$$\begin{aligned} &d\boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta})dt + \boldsymbol{\beta}(\boldsymbol{x};\boldsymbol{\theta})^{1/2}d\boldsymbol{W}, \\ &d\boldsymbol{W} \sim \mathcal{N}_n(\boldsymbol{0}, dt\boldsymbol{I}_n), \end{aligned} \tag{7}$$

where $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta})$ and $\boldsymbol{\beta}(\boldsymbol{x};\boldsymbol{\theta})$ are defined by Equations (1–6), combined with the measurement model

$$\begin{aligned} &\boldsymbol{y}_t = \boldsymbol{G}_t\boldsymbol{x}_t + \boldsymbol{r}_t, \quad \boldsymbol{r}_t \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{R}_t), \\ &\boldsymbol{R}_t = \rho_0\boldsymbol{I}_d + \rho_1\begin{bmatrix} (\boldsymbol{G}_k\boldsymbol{x}_t)_1 & & \\ & \ddots & \\ & & (\boldsymbol{G}_k\boldsymbol{x}_t)_d \end{bmatrix}, \quad d \le n, \end{aligned} \tag{8}$$

where $\boldsymbol{G}_t$ is a $d \times n$ matrix selecting only the measurable states of $\boldsymbol{x}_t$ subject to an additive noise $\boldsymbol{r}_t$, and $\boldsymbol{x}_t$ is a shorthand notation for $\boldsymbol{x}(t)$. The covariance matrix $\boldsymbol{R}_t$ models the measurement noise as a linear function $\boldsymbol{G}_t\boldsymbol{x}_t$ of the process states $\boldsymbol{x}_t$ via the vector parameter $\boldsymbol{\rho} = (\rho_0, \rho_1)'$, thus allowing to increase noise intensity with the magnitude of cell counts. Our proposed state-space formulation of Equations (7) and (8) can be interpreted as a hidden Markov model where all the states in $\boldsymbol{x}$ are latent, and some of these are measured as $\boldsymbol{y}$ through the measurement model of Equation (8).

2.3 Optimal filtering and smoothing

Consider the state-space model defined by Equations (7) and (8). Let $\boldsymbol{y}_{1:\tau}$ be the vector of measurements collected at time $t = t_1, t_2, \ldots, t_\tau$, and $\boldsymbol{x}_{1:k}$ the process' states from time t_1 to time t_k, where $k = 1, \ldots, \tau$. Assuming the Markov properties

$$\begin{aligned} &p(\boldsymbol{x}_k|\boldsymbol{x}_{1:k-1}, \boldsymbol{y}_{1:k-1}; \boldsymbol{\theta}) = p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}; \boldsymbol{\theta}) \\ &p(\boldsymbol{x}_{k-1}|\boldsymbol{x}_{k:\tau}, \boldsymbol{y}_{k:\tau}; \boldsymbol{\theta}) = p(\boldsymbol{x}_{k-1}|\boldsymbol{x}_k; \boldsymbol{\theta}) \\ &p(\boldsymbol{y}_k|\boldsymbol{x}_{1:k}, \boldsymbol{y}_{1:k-1}; \boldsymbol{\rho}) = p(\boldsymbol{y}_k|\boldsymbol{x}_k; \boldsymbol{\rho}), \end{aligned} \tag{9}$$

for the distributions involved in the dynamic and measurement models of Equations (7) and (8), the aim of optimal filtering and smoothing is to estimate

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:\tau}; \boldsymbol{\psi}), \quad \boldsymbol{\psi} = (\boldsymbol{\theta}', \boldsymbol{\rho}')', \tag{10}$$

called predictive ($k > \tau$), filtering ($k = \tau$) and smoothing ($k < \tau$) distributions, as a replacement of the (usually intractable) distribution $p(\boldsymbol{x}_{0:\tau}|\boldsymbol{y}_{1:\tau})$. Assuming a prior distribution $\boldsymbol{x}_0 \sim \mathcal{N}_n(\boldsymbol{x}_0|\boldsymbol{m}_0, \boldsymbol{P}_0)$ for $\boldsymbol{x}_t$ at $t = 0$, the distributions of Equation (10) are Gaussian, whose first two moments, and the underlying vector parameter $\boldsymbol{\psi}$, can be estimated by our proposed iterative algorithm which is summarized as follows (see Supplementary Section S6 for details):

1) **Prediction:** Solve the differential moment equations (DMEs)

$$\begin{cases} \dfrac{d\boldsymbol{m}^*(t)}{dt} = \boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{m}^*(t) \\ \boldsymbol{m}^*(t_{k-1}) = \boldsymbol{m}_{k-1} \end{cases} \tag{11a}$$

$$\begin{cases} \dfrac{d\boldsymbol{P}^*(t)}{dt} = \boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{P}^*(t) + \boldsymbol{P}^*(t)\boldsymbol{V}'_{\boldsymbol{\theta}} + \Delta t\boldsymbol{\beta}(\boldsymbol{m}^*(t), \boldsymbol{\theta}) \\ \boldsymbol{P}^*(t_{k-1}) = \boldsymbol{P}_{k-1} \end{cases} \tag{11b}$$

to obtain the first two moments $\boldsymbol{m}^*_k$ and $\boldsymbol{P}^*_k$ of the predictive distribution at time t_k ($k = 1, \ldots, \tau$), where $\boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{x}$ is a reformulation of $\boldsymbol{V}\boldsymbol{h}(\boldsymbol{x};\boldsymbol{\theta})$ as a linear function of $\boldsymbol{x}$.

2) **Update:** Compute the first two moments $\boldsymbol{m}_k$ and $\boldsymbol{P}_k$ of the filtering distribution at time t_k ($k = 1, \ldots, \tau$) via the following correction step

$$\begin{aligned} &\boldsymbol{\mu}_k = \boldsymbol{G}_k\boldsymbol{m}^*_k, \\ &\boldsymbol{S}_k = \boldsymbol{G}_k\boldsymbol{P}^*_k\boldsymbol{G}'_k + \boldsymbol{R}_k, \\ &\boldsymbol{K}_k = \boldsymbol{P}^*_k\boldsymbol{G}'_k\boldsymbol{S}_k^{-1}, \\ &\boldsymbol{m}_k = \boldsymbol{m}^*_k + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{\mu}_k), \\ &\boldsymbol{P}_k = \boldsymbol{P}^*_k - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}'_k, \end{aligned} \tag{12}$$

where $\boldsymbol{m}_k$, $\boldsymbol{P}_k$, $\boldsymbol{m}^*_k$, $\boldsymbol{P}^*_k$, $\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k$ depend on $\boldsymbol{\psi}$.

3) **Optimization:** Optimize the marginal likelihood of the measurements

$$\begin{aligned} &\boldsymbol{\psi} \leftarrow \underset{\boldsymbol{\psi}\ge 0}{\texttt{argmin}} -\ell(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\tau) \\ &\boldsymbol{y}_k \sim \mathcal{N}(\boldsymbol{\mu}_k(\boldsymbol{\psi}), \boldsymbol{S}_k(\boldsymbol{\psi})), \quad \forall k = 1, \ldots, \tau. \end{aligned} \tag{13}$$

4) **Smoothing:** Estimate $\boldsymbol{x}_k|\boldsymbol{y}_{1:\tau} \sim \mathcal{N}(\boldsymbol{m}^s_{k|\tau}, \boldsymbol{P}^s_{k|\tau})$, $k = 1, \ldots, \tau$, using the following backward step

$$\begin{cases} \boldsymbol{B}_{k+1} = \boldsymbol{P}_k e^{V'_\theta}(\boldsymbol{P}^*_{k+1})^{-1} \\ \boldsymbol{m}^s_{k|\tau} = \boldsymbol{m}_k + \boldsymbol{B}_{k+1}(\boldsymbol{m}^s_{k+1|\tau} - \boldsymbol{m}^*_{k+1}) \\ \boldsymbol{P}^s_{k|\tau} = \boldsymbol{P}_k + \boldsymbol{B}_{k+1}(\boldsymbol{P}^s_{k+1|\tau} - \boldsymbol{P}^*_{k+1})\boldsymbol{B}'_{k+1} \end{cases} \tag{14}$$

where $e^{(\cdot)}$ is the matrix exponential operator. We use a gradient-based method to solve the optimization problem of Equation (13). The gradient $\nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\psi}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_\tau)$ of the marginal log-likelihood of the measurements is estimated numerically with $p+q$ additional prediction and update steps, at time t_k, $k=1,\ldots,\tau$, to compute all the partial derivatives of $\boldsymbol{\mu}_k(\boldsymbol{\psi})$ and $\boldsymbol{S}_k(\boldsymbol{\psi})$ w.r.t. $\boldsymbol{\psi}$, where p and q are the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$. Steps 1–4 are iterated until convergence is reached. The inference procedure is summarized in Supplementary Algorithm S2.

2.4 Transition probabilities

The transition probability p_{ij} from cell type i to cell type j is defined as the multinomial probability

$$p_{ij} = \frac{\lambda_{ij}}{\alpha_i + \sum_{k\in\mathcal{O}(i)}\lambda_{ik}}, \tag{15}$$

where $\mathcal{O}(i)$ is the offspring set of cell type i, as defined by Equation (3).

2.5 Reaction constraints

To ensure identifiability of parameters in Equation (13) that involve only unobserved cell types, we use the following conservation laws

$$\lambda_{ab} = \sum_j \lambda_{bc_j}, \quad c_j \in \mathcal{O}(b), \quad j=1,\ldots,|\mathcal{O}(b)|, \tag{16}$$

where $\mathcal{O}(b)$ is the offspring set of cell type b, as defined by Equation (3), and the linear constraints

$$\begin{gathered} \lambda_{a^u b^u} = \frac{1}{m}\sum_{j=1}^{m}\lambda_{a^u b^o_j}, \alpha_{a^u} = \frac{1}{l}\sum_{j=1}^{l}\alpha_{a^o_j}, \delta_{a^u} = \frac{1}{l}\sum_{j=1}^{l}\delta_{a^o_j}, \\ a^u \in \chi^u, \quad b^o_j \in \mathcal{O}(a^u)\cap\chi^o, \quad a^o_j \in \chi^o \cap \mathcal{B}(a^u), \\ \mathcal{A}(a^u) = \mathcal{A}(a^o_j), \quad |\mathcal{O}(a^u)\cap\chi^o| = m, \quad |\chi^o \cap \mathcal{B}(a^u)| = l, \end{gathered} \tag{17}$$

where χ^u and χ^o are the sets of nonmeasured and measured cell types, $\mathcal{B}(x)$ represents the branch (Myeloid or Lymphoid) of cell lineage x, and finally $\mathcal{A}(x)$ is defined as the ancestor of cell type x. The constraint of Equation (16) can be viewed as a conservation law ensuring that all the cells differentiating from a cell lineage b to its offspring cells c_j, $j=1,\ldots,|\mathcal{O}(b)|$, are those that were previously produced by an ancestor a of b. The first constraint of Equation (17) assumes that an unobserved cell type a^u may differentiate into an unobserved cell type b^u with a rate that equals the average rate of differentiation of a^u into its observed offspring cells b^o_j, $j=1,\ldots,m$. The last two constraints of Equation (17) state that an unobserved cell type a^u may duplicate (or die) with a rate that equals the average duplication (death) rate of the observed cells a^o_j, $j=1,\ldots,l$, belonging to the branch $\mathcal{B}(a^u)$ and sharing an ancestor with a^u.

2.6 Haematopoietic models

The candidate models of cell differentiation are shown in Fig. 2. The simplest model (a) is a single-branch developmental tree where the HSCs produce all the blood cells through a single multipotent progenitor (MPP) that first differentiate into common myeloid-lymphoid progenitors (CMLP) giving rise to the mature blood lymphoid cells (NK, B, and T) without any further intermediate progenitor. Whereas the erythroid cells (platelets P and erythrocytes ERY) and myeloid cells (granulocytes G and monocytes M) are produced via the Megakaryocyte–erythroid progenitor (MEP) cells and the impaired adult myeloid progenitor (GMP) cells respectively. According to model (b) the lymphoid cells (T, B, NK) and the myeloid/erythroid cells (G, M, P, ERY) are generated through separate branches of differentiation. In particular, MPP cells first differentiate into either common myeloid progenitor (CMP) cells or common lymphoid progenitor (CLP) cells, before giving rise to MEP and GMP cells. This model is known as classical/dichotomic model of haematopoiesis (Kawamoto and Katsura 2009). In contrast, model (c) proposes the idea that myeloid progenitors represent a prototype of haematopoietic cells capable to produce both myeloid (G, M) cells, erythroid (P, ERY) cells, and lymphoid NK cells, whereas lymphoid T, B, and NK cells represent specialized types produced by the CLP cells. Indeed, CMP cells not only produce MEP and GMP cells, but also NK cells. This model is known as the myeloid-based model (Kawamoto and Katsura 2009). Finally, model (d) assumes that lymphoid T/B cells and lymphoid NK cells develop through different branches. While B and T cells are produced by the common progenitor CLP, the NK cells are generated by a separate progenitor NKP.

2.7 Model selection

The candidate models of cell differentiation are scored according to the Akaike information criterion (AIC) (Burnham *et al.* 2011), i.e.

$$AIC(\mathcal{M}) = 2p_{\mathcal{M}} - 2\ell_{\mathcal{M}}(\boldsymbol{\psi}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_\tau), \tag{18}$$

where $\ell_{\mathcal{M}}$ is the marginal log-likelihood of the measurements of model $\mathcal{M}$ and $p_{\mathcal{M}}$ its number of parameters.

3 Results

We tested our proposed method in several simulation studies based on the validation scenarios from Fig. 3. These networks are sufficiently complex and diverse, in terms of number of nodes and depth, to assess the performance of our method against recovery of the true data generative process and underlying parameters. After validating our method, we analysed data from two preclinical studies and three gene therapy clinical trials to shed light on cell differentiation and haematopoietic reconstitution. The specific results are reported in the next subsections.

3.1 Validation and comparison with the prior art

We tested our proposed method Karen and compared it with the state-of-the-art approaches, such as the generalized least squares (GLS) method (Pellin *et al.* 2019), the maximum likelihood method RestoreNet (Del Core *et al.* 2023), and the branchCorr method (Xu *et al.* 2019). The validation and comparisons were made in terms of robustness against (i) the sampling frequency τ, (ii) the fraction ζ of false-negatives, and

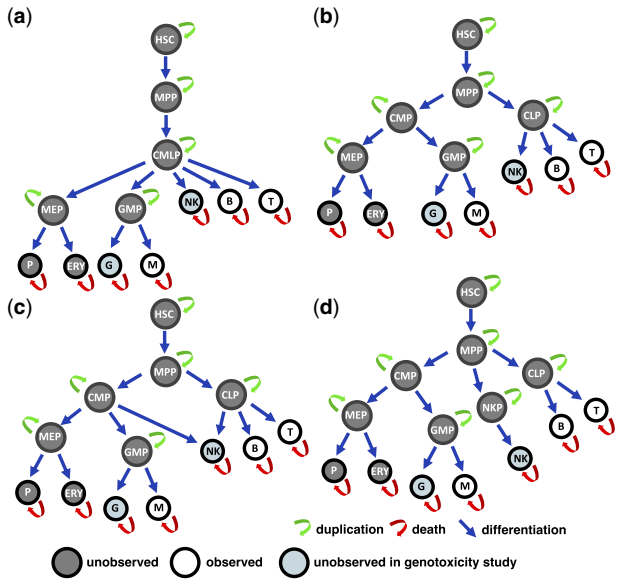

Figure 2. Graphical representation of the cell differentiation networks proposed as candidate models in the *in vivo* studies (a-d). Grey and white nodes represent unobserved and observed cell types. Light-grey nodes are the cell lineages whose data were collected in all studies except for the genotoxicity one. Arrows represent cell duplication (green), cell death (red), and cell differentiation (blue).

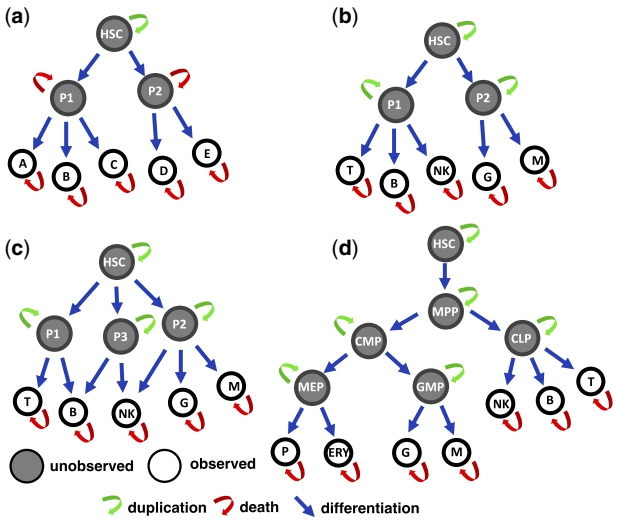

Figure 3. Graphical representation of the cell differentiation networks used in the in-silico studies (a-d). Grey and white nodes represent unobserved and observed cell types. Arrows represent cell duplication (green), cell death (red), and cell differentiation (blue).

(iii) the magnitude of the measurement noise parameters ρ_0 and ρ_1. To make the comparison of our method with the other candidates possible, we used the cell differentiation structure of Fig. 3a as the true generative model, whose type of biochemical reactions is the only one that can be handled by Xu *et al.* (2019). The corresponding biochemical reactions and the system of SDEs were defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). Two different comparative synthetic studies have been designed. In the first one all the cell types were measured, thus branchCorr was not included, since it does not allow for observed progenitors. In the second study the synthetic HSCs and progenitors P1–P2 were considered as unobserved states, and therefore GLS and RestoreNet were excluded from this comparison, since both methods do not allow for unobserved states. We used the Euler–Maruyama Algorithm S1 of the Supplementary Information to forward-simulate 100 independent stochastic trajectories of three clones from the true generative data process. Details on parameters and initial condition $\boldsymbol{x}_0$ used for the simulations are reported in Supplementary Tables S1 and S2. Then we used our proposed framework Karen and the other methods to infer the unknown parameters. Inference with Karen has been carried out using Supplementary Algorithm S2, with and without the conservation laws of Equation (16) that ensure identifiability of the parameters involving only unobserved cell types.

Results from Fig. 4 clearly indicate that our proposed method outperformed the prior art. In particular, Fig. 4a provides evidence that our proposed method is the most robust with respect to the false negative errors compared to the other methods, which provided more biased estimates for the parameters for a high percentage $\zeta = 90\%$ of missing data. Subsequently, Fig. 4b shows that a low sampling frequency ($\tau = 4$) of the simulated trajectories did not affect the estimates provided by our proposed method, whereas those obtained with any of the competitor approaches were biased. Finally, Fig. 4c suggests that after increasing the magnitude of the measurement noise parameters ρ_0 and ρ_1 up to 10, our proposed method still provided better estimates compared to the other methods. Further results for different values of ζ, τ, ρ_0, and ρ_1 can be found in Supplementary Section S7. In conclusion, the results from our synthetic studies show that overall our method outperformed the prior art in terms of false negative errors, sampling frequency and measurement noise. Details on the computational complexity are reported in Supplementary Section S10.

3.2 Model misspecification

We tested our method against model misspecification with an in-silico study. The cell differentiation networks of Fig. 3b and c were used as true generative models that we cross-compared for simulating and fitting. The corresponding biochemical reactions and the system of SDEs were defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). We performed 100 independent simulations of the stochastic trajectories for three clones using the Euler–Maruyama Algorithm S1 of the Supplementary Information. Details on parameters and initial condition $\boldsymbol{x}_0$ used for the simulations are reported in Supplementary Tables S3 and S4. Then we fitted both candidate models using the inference Supplementary Algorithm S2. Forward simulation and fitting has been carried out by using the conservation laws of Equation (16), so as to ensure identifiability of the parameters involving only unobserved cell types. As a result, Fig. 5a indicates that our method performed well in model selection, since the true models had the lowest median AIC, as defined by Equation (18), over 100 independent simulations. Moreover, Fig. 5a suggests that the true models yielded better fits in terms of the smoothing moments $\boldsymbol{m}^s_{k|\tau}$ and $\boldsymbol{P}^s_{k|\tau}$ ($k = 1, \ldots, \tau$) compared to the wrong models.

3.3 Scalability to complex networks

We performed a synthetic study to explore the scalability of our proposed method to more complex network structures. We used the cell differentiation network of Fig. 3d as the

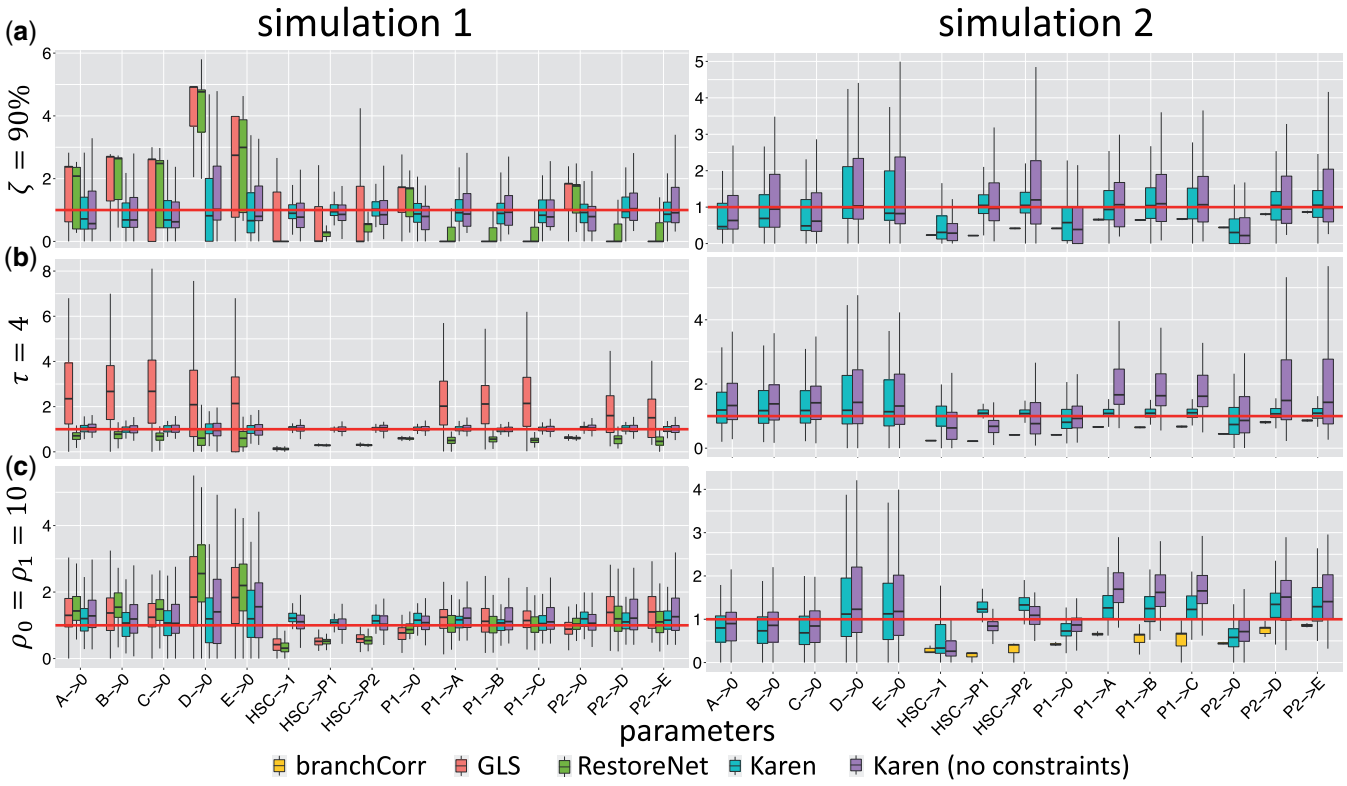

Figure 4. For each comparative synthetic study with observed (left) and systematically missing (right) progenitors HSC, P1, and P2: boxplots (*y*-axis) of the estimated parameters divided by the true ones for each reaction rate (*x*-axis) obtained from each method (colours), across all simulations, under a fraction $\zeta = 90\%$ of false negative errors (top), a sampling frequency $\tau = 4$ (middle), and a measurement noise generated by $\rho_0 = \rho_1 = 10$ (bottom).

true generative model. The corresponding biochemical reactions and the system of SDEs were defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). As displayed in Fig. 3d, we assumed that the clonal cell counts were not collected from the HSCs and all the progenitors (MPP, CMP, CLP, MEP, GMP), and all the cell lineages that are missing for a particular clone at a given time point were also considered as unobserved. Therefore, for the measurement model of Equation (8) the selection matrix $\boldsymbol{G}_t$ was defined accordingly. We performed 50 independent simulations of stochastic trajectories for 100 clones from the generative data process of Fig. 3d using the Euler–Maruyama Algorithm S1 of the Supplementary Information. Details on parameters and initial condition $\boldsymbol{x}_0$ used for the simulations are reported in Supplementary Tables S5 and S6. Then we applied our inference method Karen on the simulated data using Supplementary Algorithm S2. Each simulation and fitting has been carried out by assuming the conservation laws of Equation (16), so as to ensure identifiability of the parameters that involve only unobserved cell types. Results from Fig. 5b and c show low uncertainty of the estimated parameters and a good recovery of the Markov states in terms of the first two smoothing moments $\boldsymbol{m}^s_{k|\tau}$ and $\boldsymbol{P}^s_{k|\tau}$, $k = 1, \ldots, \tau$.

3.4 Genotoxicity study

We analysed an *in vivo* clonal tracking dataset previously used to investigate the impact of vector design on clonal diversity in tumour-prone mice (Del Core *et al.* 2022). Cdkn2a$^{-/-}$ tumour prone Lin$^-$ cells were first *ex vivo* transduced with a lentiviral vector expressing GFP under either spleen focus-forming virus (SFV) or PGK promoter/enhancer sequence. Cells are then transplanted into lethally irradiated wild-type mice. To recover enough DNA material, equal amounts of blood from two or three mice belonging to the same experimental group were pooled before cell sorting. Integration sites were then retrieved by polymerase chain reaction (PCR) at different time points from sorted T (CD3+) and B (CD19+) lymphocytes and myeloid cells (CD11b+). Clonal tracking samples were collected under heterogeneous technical conditions (see Supplementary Table S10), making them not directly comparable. Therefore, we rescaled the data following the description in Supplementary Section S11. The total number of distinct clones that were collected are 45 186 and 20 471 for the PGK and SFV treatments, respectively. To further remove bias, we focused our analyses on the top 1000 most recaptured clones across lineages and time. We used the inference Supplementary Algorithm S2 to fit the four biologically sustained models of cell differentiation from Fig. 2 under the two vector conditions PGK and SFV. For each model the system of SDEs was defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). Inference has been carried out by assuming the conservation laws of Equations (16) and (17), so as to ensure identifiability of the parameters involving only systematically unobserved cell types. For each candidate model we computed the Akaike information criterion (AIC), as defined by Equation (18), and we report the results on model selection in Table 1. According to the AIC, model (b) is the one that best fitted clonal tracking data under each vector design. The corresponding cell differentiation networks are displayed in Fig. 6. This result suggests that the classical/dichotomic model structure (b) adequately described clonal dynamics in tumour-prone mice under both treatments. Also, the arrow weights from Fig. 6 clearly

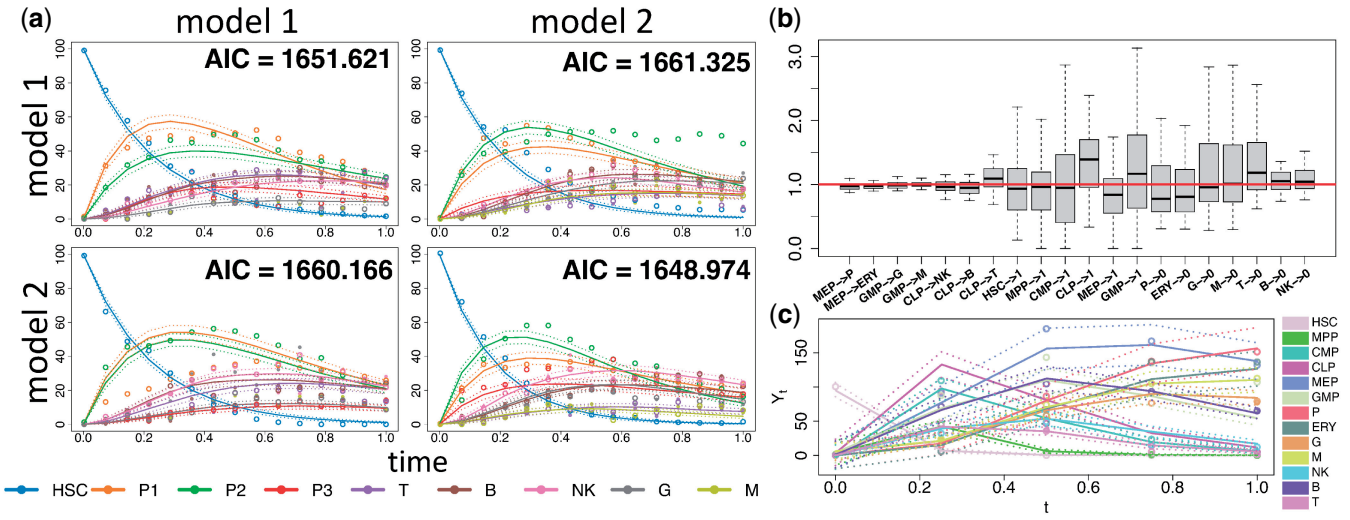

Figure 5. (a) For each true data generating process (row) and for each candidate model (column), the simulated process (empty dots), the synthetic data (full dots), the estimated smoothing moments (lines) of a single clone for each cell type (colours), and the median AIC across all simulations used to evaluate model misspecification. (b) Boxplots (*y*-axis) of the ratio between the estimated and true parameters for each reaction (*x*-axis) of the cell differentiation network of Fig. 3d used in the scalability study. (c) Estimated smoothing moments (lines), the true Markov states (empty dots), and the synthetic data (full dots) of one clone for each cell type (colours) for a single simulation of the cell differentiation network of Fig. 3d used in the scalability study.

indicate that in SFV-treated tumour-prone mice there was a more pronounced unbalance in cell differentiation from the multipotent progenitors (MPP) towards common myeloid progenitors (CMP) compared to the PGK treatment. Therefore our proposed framework Karen suggests that, in this particular study, the design of viral vector did not significantly affect the structure of cell differentiation in tumour-prone mice, but had an impact on the transition probabilities $p(\text{MPP} \to \text{CMP})$ and $p(\text{MPP} \to \text{CLP})$, representing differentiation from the multipotent progenitors compartment (MPP) in either common myeloid (CMP) or lymphoid (CLP) progenitors.

3.5 Rhesus Macaques study

We analysed an *in vivo* clonal tracking dataset collected from Rhesus Macaques (Wu *et al.* 2014). HSCs were first barcoded by using lentiviral vectors and then transplanted in three animals. Barcode retrieval was performed monthly via PCR on Granulocytes (G), Monocytes (M), T, B, and NK cells up to 9.5 months. Further details on transductions protocol and culture conditions can be found in Wu *et al.* (2014). Although the sample DNA amount was maintained constant during the whole experiment, the samples resulted in different magnitudes of reads (see Supplementary Table S11), making the data not directly comparable. Therefore, we rescaled the barcode counts as described in Supplementary Section S12 before analysis. The total numbers of clones that were collected range in 1165–1291, but we focused on the top 1000 most recaptured ones, across lineages and time, so as to further remove bias. We fitted the same four candidate models from the previous section on the clonal tracking data using Supplementary Algorithm S2. For each model the system of SDEs was defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). In analogy with the previous section, inference has been performed by using the conservation laws of Equations (16) and (17) that ensure identifiability of the parameters involving only unobserved cell lineages. Each candidate model has been scored according to the Akaike information criterion (AIC) of Equation (18) and we report the results on model selection in Table 1. According to the AIC, model (c) is the one that best fitted the clonal tracking data collected from the rhesus macaque study. The estimated cell differentiation network is reported in Fig. 6. This result suggests that the classical/dichotomic model (b) failed to describe adequately clonal dynamics in rhesus macaques, whereas the myeloid-based developmental model (c) better explained haematopoietic reconstitution. Our proposed framework Karen clearly indicates that myeloid progenitors represent a prototype of haematopoietic cells capable to produce both myeloid (G, M) cells, erythroid (P, ERY) cells and lymphoid NK cells in primates.

Table 1. For each *in vivo* clonal tracking dataset analysed (rows) and the candidate models a–d (columns) the AIC computed according to Equation (18).

	a	b	c	d
PGK	128 230.94	120 429.29	120 454.55	323 765.06
LTR	76 220.33	75 102.02	75 206.44	75 161.76
RM	138 921.13	802 760.11	108 993.15	109 578.07
WAS	103 931.52	102 959.02	508 888.96	102 444.74
$\beta 0\beta$E	50 900.45	50 142.89	50 936.42	49 960.66
βSβS	46 665.12	45 876.46	45 898.89	45 872.07

3.6 Gene therapy clinical trials

We considered clonal tracking data collected from six patients affected by three different genetic disorders. The six patients underwent a haematopoietic stem and progenitor cell (HSPC) gene therapy treatment. Vector integration sites in five cell lineages (G, M, T, B, and NK) were collected longitudinally from the peripheral blood of four patients affected by Wiskott–Aldrich syndrome (WAS) (Hacein-Bey Abina *et al.* 2015), two patients with β haemoglobinopathy, one with βS/βS sickle cell disease (Ribeil *et al.* 2017), and one with $\beta 0/\beta$E β thalassaemia (Thompson *et al.* 2018). Details on procedures, gene therapy protocols, and normalization methods

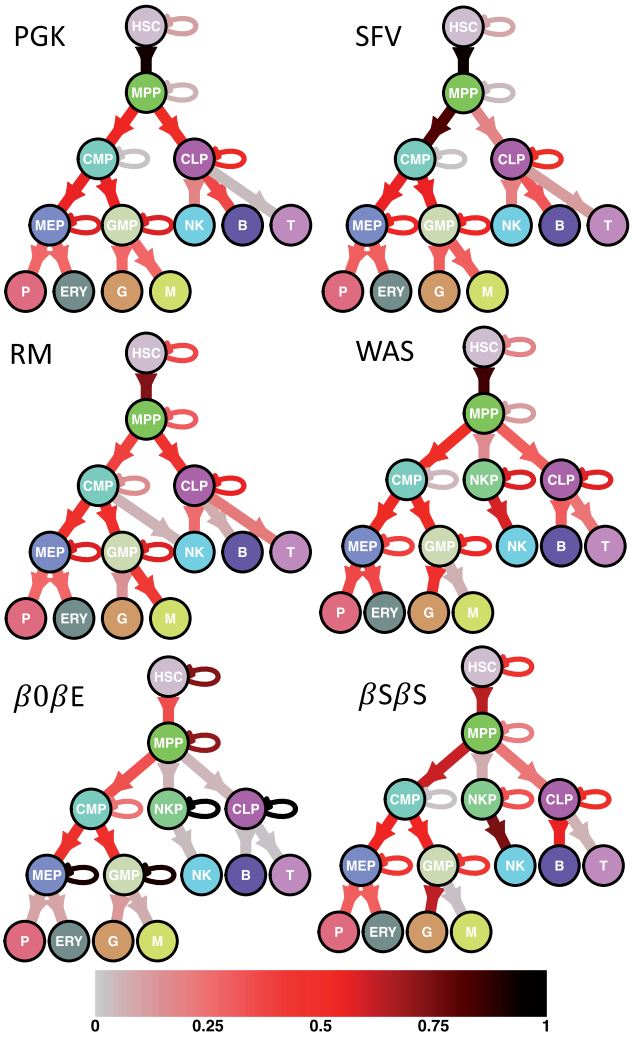

Figure 6. Inferred cell differentiation networks having the lowest AIC, as defined by Equation (18), for the genotoxicity study for the comparison of the two viral vectors PGK and SFV, the rhesus macaque (RM) study, and the clinical trials WAS, $\beta 0\beta E$ and $\beta S\beta S$. Each arrow is weighted and coloured according to the transition probabilities estimated with Equation (15).

can be found in Hacein-Bey Abina *et al.* (2015), Ribeil *et al.* (2017), and Thompson *et al.* (2018). Since data were already normalized to compensate for unbalanced sampling in VCN and DNA (Sherman *et al.* 2017), we did not apply any further transformation. The total clones that were collected are 156 654, 17 273, and 230 408, respectively, for WAS, $\beta S/\beta S$, and $\beta 0/\beta E$ clinical trials. The following results stem from the analysis of the 1000 most recaptured clones, across lineages and time, in each clinical trial. The four haematopoietic models of Fig. 2 have been scored separately in each clinical trial using Supplementary Algorithm S2. For each model the system of SDEs was defined following Equations (1–6), using the state-space formulation of Equations (7) and (8). In analogy with previous sections, inference has been carried out by assuming the linear constraints of Equations (16) and (17) ensuring identifiability of the parameters that involve only the unobserved cell types. For each candidate model we computed the Akaike information criterion (AIC), as defined by Equation (18), and we report the results on model selection in Table 1. According to the AIC, model (d) is the one that best fitted clonal tracking data collected from each clinical trial, and the corresponding cell differentiation networks are reported in Fig. 6. Results suggest that a three-branches developmental model best explained haematopoietic reconstitution in these gene therapy clinical trials. While lymphoid (T, B) and myeloid/erythroid cells (G, M, P, ERY) developed in parallel through separate branches from different progenitors, NK cells appear to be sustained by a dedicated progenitors' cell population.

4 Discussion

We have proposed a novel stochastic framework for calibrating cell differentiation networks from partially observed high-dimensional clonal tracking data. Our model is able to deal with experimental clonal tracking data that suffers from measurement noise and low levels of clonal recapture due to either threshold detection failures or false-negative errors. Our proposed framework Karen extends stochastic quasi-reaction networks by introducing an EKF and an RTS smoother. We have developed an E–M algorithm to infer the corresponding parameters. Simulation studies have shown the method's accuracy regarding inference of the true parameters, estimation of the first two smoothing moments of all the process states, and model selection using AIC. Simulation results indicated higher robustness of our proposed method compared to the state-of-the-art ones in terms of (i) low sampling frequency, (ii) limited clonal recapture, and (iii) high levels of measurement noise. Results from simulations suggest that our proposed framework scales to complex structures of cell differentiations in terms of nodes size and network depth. Although the Gaussian assumption makes the analytical formulations of the likelihoods explicitly available, this approximation may become poor when the data contains outliers or shows non-Gaussian behaviours. This limitation can be overcome by using a distribution-free approach, such as the Kernel Kalman Rule, a recent nonparametric inference technique for high dimensional and possibly non-Gaussian nonlinear state-space models (Gebhardt *et al.* 2019). Besides, our framework considers reaction rates constant for the whole study period. Extensions that allow for modelling reaction rates as smooth functions of time or clinically relevant variables will be the goal of our future research.

Our proposed method allowed to unveil the genotoxic impact on cell differentiation in tumour-prone mice. While the differentiation structure does not seem to be affected by the viral vector design, the transition probabilities from the multipotent progenitors to the intermediate progenitors do, showing a more pronounced unbalance towards the common myeloid progenitors under the SFV treatment compared to PGK. This can be biologically interpreted as a faster immune response to the higher inflammation caused by the toxic SFV treatment, compared to the nontoxic PGK one. Subsequently, the application of Karen to a rhesus macaque clonal tracking study unveiled for the lymphoid NK cells a different developmental pathway from the one detected for lymphoid T and B cells. That is, NK cells are produced by both common myeloid CMP and lymphoid CLP progenitors, whereas T and B cells are sustained only by the common lymphoid progenitors CLP. Results are consistent with those previously reported in Wu *et al.* (2014), where the authors demonstrated the presence of distinct subpopulations within the NK cells lineage, potentially deriving from alternative maturation processes. Finally, it is worth noting the agreement in the inferred

network structure between the different clinical trials. Our modelling approach is able to capture the heterogeneity in cell repopulation dynamics and selective advantage of different contexts, as suggested by the parameter estimates. Our stochastic framework can support biologists to shed light on haematopoietic reconstitution and in designing tailor-made therapies to treat genetic disorders. Our model can be applied to different types of clonal tracking data, such as vector integration sites, clonal barcodes, and single cell methods. Applications in alternative contexts of population dynamics, showing similar issues of partial sampling and measurement noise, could also be explored.

Author contributions

All authors contributed to analysing the data and writing the manuscript. L.D.C. designed and implemented the stochastic framework Karen. E.C.W. and M.A.G. jointly supervised this work.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This publication is based on work from COST Action CA15109, supported by the European Cooperation in Science and Technology. E.C.W. acknowledges support from the Fondazione Leonardo [514.7.010.098-4] and the Swiss National Science Foundation [SNSF 188534].

References

Bobo D, Lipatov M, Rodriguez-Flores JL *et al.* False negatives are a significant feature of next generation sequencing callsets. bioRxiv, https://doi.org/10.1101/066043, 2016, preprint: not peer reviewed.

Burnham KP, Anderson DR, Huyvaert KP *et al.* AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol* 2011;**65**:23–35. https://doi.org/10.1007/s00265-010-1029-6.

Catlin SN, Abkowitz JL, Guttorp P *et al.* Statistical inference in a two-compartment model for hematopoiesis. *Biometrics* 2001;**57**:546–53. https://doi.org/10.1111/j.0006-341X.2001.00546.x.

Cooper GM, Adams K. *The Cell: A Molecular Approach*. New York, NY: Oxford University Press, 2023.

Del Core L. The stochastic route of haematopoiesis: modelling and inference methods in clonal tracking studies. PhD Thesis, University of Groningen, 2023. https://doi.org/10.33612/diss.

Del Core L, Cesana D, Gallina P *et al.* Normalization of clonal diversity in gene therapy studies using shape constrained splines. *Sci Rep* 2022;**12**:3836. https://doi.org/10.1038/s41598-022-05837-0.

Del Core L, Pellin D, Wit EC *et al.* A mixed-effects stochastic model reveals clonal dominance in gene therapy safety studies. *BMC Bioinformatics* 2023;**24**:228. https://doi.org/10.1186/ s12859-023-05269-1.

Di Serio C, Scala S, Vicard P *et al.* Bayesian networks for cell differentiation process assessment. *Stat* 2020;**9**:e287. https://doi.org/10.1002/sta4.287.

Dingli D, Pacheco JM. Modeling the architecture and dynamics of hematopoiesis. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**:235–44. https://doi.org/10.1002/wsbm.56.

Érdi P, Tóth J. *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Manchester University Press, 1989.

Gebhardt GHW, Kupcsik A, Neumann G *et al.* The kernel kalman rule. *Mach Learn* 2019;**108**:2113–57. https://doi.org/10.1007/s10994-019-05816-z.

Hacein-Bey Abina S, Gaspar HB, Blondeau J *et al.* Outcomes following gene therapy in patients with severe Wiskott–Aldrich syndrome. *JAMA* 2015;**313**:1550–63. https://doi.org/10.1001/jama.2015.3253.

Kawamoto H, Katsura Y. A new paradigm for hematopoietic cell lineages: revision of the classical concept of the myeloid–lymphoid dichotomy. *Trends Immunol* 2009;**30**:193–200. https://doi.org/10.1016/j.it.2009.03.001.

Kawamoto H, Wada H, Katsura Y *et al.* A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. *Int Immunol* 2010;**22**:65–70. https://doi.org/10.1093/intimm/dxp125.

Kim Y-H, Song Y, Kim J-K *et al.* False-negative errors in next-generation sequencing contribute substantially to inconsistency of mutation databases. *PLoS One* 2019;**14**:e0222535. https://doi.org/10.1371/journal.pone.0222535.

Pellin D, Biasco L, Aiuti A *et al.* Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking. *Appl Netw Sci* 2019;**4**:115. https://doi.org/10.1007/s41109-019-0225-1.

Ribeil J-A, Hacein-Bey-Abina S, Payen E *et al.* Gene therapy in a patient with sickle cell disease. *N Engl J Med* 2017;**376**:848–55. https://doi.org/10.1056/NEJMoa1609677.

Roeder I, Loeffler M. A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. *Exp Hematol* 2002;**30**:853–61. https://doi.org/10.1016/S0301-472X(02)00832-9.

Roeder I, Kamminga LM, Braesel K *et al.* Competitive clonal hematopoiesis in mouse chimeras explained by a stochastic model of stem cell organization. *Blood* 2005;**105**:609–16. https://doi.org/10.1182/blood-2004-01-0282.

Sherman E, Nobles C, Berry CC *et al.* INSPIIRED: a pipeline for quantitative analysis of sites of new dna integration in cellular genomes. *Mol Ther Methods Clin Dev* 2017;**4**:39–49. https://doi.org/10.1016/j.omtm.2016.11.002.

Thompson AA, Walters MC, Kwiatkowski J *et al.* Gene therapy in patients with transfusion-dependent β-thalassemia. *N Engl J Med* 2018;**378**:1479–93. https://doi.org/10.1056/NEJMoa1705342.

Till JE, Mcculloch EA, Siminovitch L *et al.* A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc Natl Acad Sci USA* 1964;**51**:29–36. https://doi.org/10.1073/pnas.51.1.29.

Wu C, Li B, Lu R *et al.* Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* 2014;**14**:486–99. https://doi.org/10.1016/j.stem.

Xu J, Koelle S, Guttorp P *et al.* Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *Ann Appl Stat* 2019;**13**:2091–119. https://doi.org/10.1214/19-AOAS1272.