
Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data

L. Del Core^{1,*}, D. Pellin², M. A. Grzegorzczak^{1,*} and E. C. Wit^{3,*}

¹University of Groningen - Bernoulli Institute, 9747AG, Groningen, Netherlands, ²Harvard Medical School, MA 02115, Boston, Massachusetts, and


³Università della Svizzera italiana - Institute of Computing, 6962, Lugano, Switzerland

*To whom correspondence should be addressed.

Abstract

Motivation: Clarifying how hematopoietic stem cells differentiate into mature cell types is important for understanding how they attain specific functions and offers the potential for therapeutic manipulation. Over the past decades, clonal tracking has proven to be capable of unveiling population dynamics and hierarchical relationships in vivo. For this reason, clonal tracking studies are required for safety and long-term efficacy assessment in gene therapy. However, many standard clonal tracking studies consider only a subset of cell-types and are subject to noise.

Results: In this work, we propose a stochastic framework that investigates the dynamics of cell differentiation from typical clonal tracking data subject to measurement noise, false-negative errors, and systematically unobserved cell types. Our framework is based on stochastic reaction networks combined with extended Kalman filtering and Rauch-Tung-Striebel smoothing. Our tool can provide statistical support to biologists in gene therapy clonal tracking studies to better understand clonal reconstitution dynamics.

Availability: The stochastic framework is implemented in the  package Karen which is available for download at <https://github.com/delcore-luca/Karen>. The code that supports the findings of this study is openly available at <https://github.com/delcore-luca/CellDifferentiationNetworks>.

Contact: l.del.core@rug.nl

1 Introduction

Hematopoiesis 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 the root node constituted by Hematopoietic Stem Cells (HSC) [1; 2]. Each cell division gives rise to progeny cells that can retain the properties of their parent cell (self-renewal) or differentiate, “moving down” the hematopoietic tree [3–7]. As the progeny move further away from 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 [8]. Several mathematical models have been proposed to describe hematopoiesis in-vivo. One of the first stochastic models of hematopoiesis was introduced in the early '60s [9] suggesting that it is the population as a whole that is regulated rather than individual cells that behave stochastically, and control mechanisms act by varying the cell division and death rates.

More recently, [10–16] analyze data generated by using the most advanced lineage tracing protocols using novel statistical models. Still, to the best of our knowledge, none of the already existing tools considers the presence of false-negative clonal tracking errors. In addition to completely missing cell types, clonal tracking data are characterized by scattered detection (recapture) of clones due to either threshold detection failure or false-negative errors [17]. Usually, threshold detection failure is addressed by assuming that all the missing clone observations correspond to minimal clones and, therefore, set to zero. 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 [18]. It has also been shown that false-negative errors strongly depend on calling pipeline parameters, as well as read coverage [19]. The false-negative diagnosis rate is poorly understood for many NGS applications and is challenging to measure without the use of well-characterized reference standards [20]. The standardization of sequencing coverage depth has also been used to minimize the probability of false-positive and false-negative results. However, there is no consensus on the minimum coverage depth that clonal tracking data have to comply with, creating heterogeneity in the quality of data generated by the different laboratories [21].

We propose a stochastic framework to investigate haematopoiesis while cautiously treating all the undetected values as latent states. More precisely, we describe cell differentiation using stochastic quasi-reaction networks (SqRNs), a framework that allows to (i) model a network of stochastically interacting nodes using an Ito-type SDE formulation, and (ii) describe the dynamics of transition between different states (cell types) in terms of a set of reactions whose rates are unknown. Then we combine SqRNs with extended Kalman filtering (EKF) and Rauch-Tung-Striebel (RTS) smoothing. In particular, we (i) provide an expectation-maximization algorithm to infer the unknown parameters; (ii) extensively test the method on several simulation studies (iii) applied our framework to four in-vivo high dimensional clonal tracking data sets, to compare different biologically plausible models of cell differentiation. A flowchart of the analysis performed in this work is shown in Figure 1.

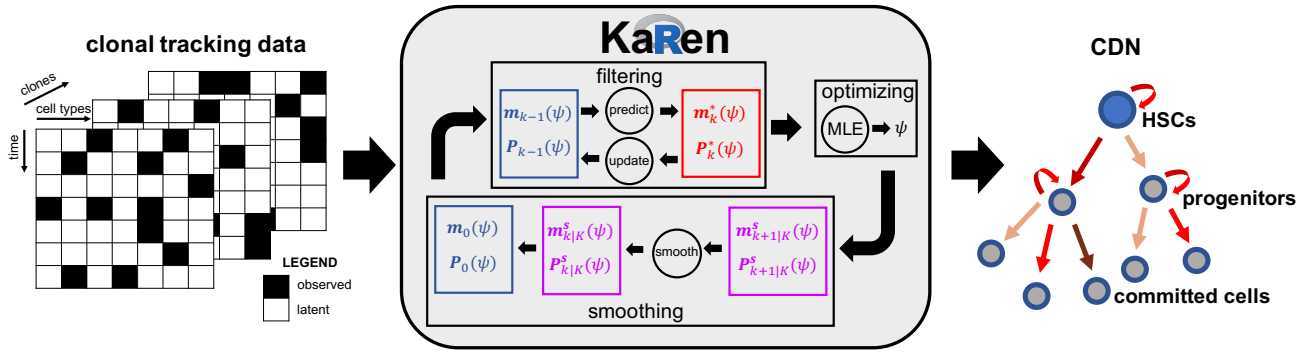


Fig. 1: Schematic representation of the analysis flow: A three-dimensional clonal tracking dataset with partially-observed cells (left panel) is received as input from our proposed stochastic framework KaRen (middle panel). It mainly consists in three parts, such as a filtering step, an optimization (maximum likelihood) step, and a smoothing step which are executed iteratively until a convergence is reached on the unknown vector parameter ψ . Finally, a cell differentiation network (CDN) is returned from KaRen, where each arrow is directed and weighted according to the estimated parameters (right panel).

2 Methods

2.1 KaRen: Kalman Reaction Networks

We consider a non-linear continuous-discrete state space model (CD-SSM) whose dynamic component is represented by the local reaction approximation (LLA) of a stochastic quasi-reaction network [22] defined by a $n \times J$ net effect matrix V , a $p \times 1$ vector parameter θ and a $J \times 1$ hazard vector $h(x; \theta)$ for a n -dimensional counting process $\{x(t) | x(t) \in \mathbb{N}^n\}_t$ (see Section S.1 from Supplementary Information for details). For the measurement function $g_k(x(t_k), r_k)$ we use a time-dependent selection matrix $G_k \in \mathbf{01}^{d \times n}$ (the set of all $d \times n$ binary matrices) which selects only the measurable particles of $x(t_k)$ with an additive noise r_k whose covariance matrix is time-dependent and defined as

$$R_k = \rho_0 I_d + \rho_1 \text{diag}(G_k x(t_k)) \quad \forall k = 1, \dots, K$$

where ρ_0 and ρ_1 are free parameters which we infer from the data, and $\text{diag}(\cdot)$ is a diagonal matrix with diagonal equal to its argument. Therefore

$$\begin{aligned} g_k(x(t_k), r_k) &= G_k x(t_k) + r_k; \quad r_k \sim \mathcal{N}_d(0, R_k); \\ R_k &= \rho_0 I_d + \rho_1 \text{diag}(G_k x(t_k)) \quad \forall k = 1, \dots, K \end{aligned} \quad (1)$$

In the following x_t is a shorthand notation for $x(t)$. Under these assumptions the CD-SSM of Eq. (1) from the **Supplementary Information** reduces to

$$\begin{aligned} \Delta x &= V h(x_t; \theta) \Delta t + \left(V \underbrace{\begin{bmatrix} h_1(x_t; \theta) \\ \vdots \\ h_J(x_t; \theta) \end{bmatrix}}_{\beta(x_t; \theta)} V' \right)^{1/2} dW_t \\ \Delta x &= x_{t+1} - x_t; \quad y_k = G_k x(t_k) + r_k \end{aligned} \quad (2)$$

where

$$\begin{aligned} dW_t &\sim \mathcal{N}_n(0, \Delta t I_n) \\ r_k &\sim \mathcal{N}_d(0, R_k); \quad R_k = \rho_0 I_d + \rho_1 \text{diag}(G_k x(t_k)) \end{aligned} \quad (3)$$

Assuming $x(t_0) \sim \mathcal{N}_n(x(t_0) | m_0, P_0)$ as prior distribution for $x(t)$ at $t = t_0$, the prediction step of Eq. (6) from the **Supplementary Information** reduces to

1. Prediction step:

$$\begin{cases} \frac{dm_k^*(t)}{dt} = V_\theta m_k^*(t) \\ m_k^*(t_{k-1}) = m_{k-1} \end{cases} \quad (4a)$$

$$\begin{cases} \frac{dP_k^*(t)}{dt} = V_\theta P_k^*(t) + P_k^*(t) V_\theta' + \Delta t \beta(m_k^*(t), \theta) \\ P_k^*(t_{k-1}) = P_{k-1} \end{cases} \quad (4b)$$

where we used the fact that for a set of reactions involving only one particle of x as reagent, which is the case for our cell differentiation networks, the mean drift $V h(x_t; \theta)$ reduces to $V_\theta x_t$, where the definition of V_θ depends on V and $h(x_t; \theta)$. The solutions of (4) are given by

$$m_k^*(t) = e^{V_\theta(t-t_{k-1})} m_{k-1} \quad (5a)$$

$$\begin{aligned} P_k^*(t) &= e^{V_\theta(t-t_{k-1})} P_{k-1} e^{V_\theta'(t-t_{k-1})} \\ &+ \int_{t_{k-1}}^t e^{V_\theta(t-s)} \Delta t \beta(m_k^*(s); \theta) e^{V_\theta'(t-s)} ds \end{aligned} \quad (5b)$$

The solution for $m_k^*(t)$ is obtained by applying the integrating factor method [23] to the initial value problem (4a) using an integrating factor $I = e^{-\int_{t_{k-1}}^t V_\theta ds} = e^{-V_\theta(t-t_{k-1})}$. The solution for $P_k^*(t)$ is obtained by applying the well-known solution formula for a differential Sylvester equation [24] to the system (4b). The corresponding update step is defined as follows

2. Update step:

$$\begin{aligned} \mu_k &= G_k m_k^* \\ S_k &= G_k P_k^* G_k' + R_k \\ K_k &= P_k^* G_k' S_k^{-1} \\ m_k &= m_k^* + K_k (y_k - \mu_k) \\ P_k &= P_k^* - K_k S_k K_k' \end{aligned} \quad (6)$$

where m_k , P_k , m_k^* , P_k^* , μ_k and S_k depend on both θ , ρ_0 and ρ_1 . Finally, following Eq. (8) - (10) of the **Supplementary Information**, the optimization and smoothing steps are defined as

3. Optimization step:

$$\begin{aligned} \psi &\leftarrow \underset{\psi \geq 0}{\operatorname{argmin}} -\ell(\psi|y_1, \dots, y_K) \\ y_k &\sim \mathcal{N}(\mu_k(\psi), S_k(\psi)) \quad \forall k = 1, \dots, K \end{aligned} \quad (7)$$

4. Smoothing step: We use the Rauch-Tung-Striebel Smoothing algorithm (RTS) [25] and we estimate the first two-order moments of $p(x_k|y_{1:K}, \theta, \rho_0, \rho_1)$ as

$$\begin{cases} B_{k+1} = P_k(\psi) e^{V'_\psi (P_{k+1}^*(\psi))^{-1}} \\ m_{k|K}^s = m_k(\psi) + B_{k+1}(m_{k+1|K}^s - m_{k+1}^*(\psi)) \\ P_{k|K}^s = P_k(\psi) + B_{k+1}(P_{k+1|K}^s - P_{k+1}^*(\psi)) B_{k+1}' \end{cases} \quad (8)$$

where $e^{(\cdot)}$ is the matrix exponential operator, $\psi = (\theta, \rho_0, \rho_1)$, and the values of m_k , P_k , m_k^* , P_k^* are the ones obtained from the filtering (prediction and update) steps. In order to run the optimization step using a gradient-based method (e.g. Newton-Raphson) we need to compute the gradient $\nabla_{\theta, \rho_0, \rho_1} \varphi(\theta, \rho_0, \rho_1)$ of the energy function $\varphi(\theta, \rho_0, \rho_1)$ which is defined by

$$\begin{aligned} \frac{\partial \varphi(\psi)}{\partial \psi_j} &= \operatorname{tr} \left(S^{-1} \frac{\partial S}{\partial \psi_j} \right) - \left(\frac{\partial \mu}{\partial \psi_j} \right)' S^{-1} (y - \mu) \\ &- (y - \mu)' S^{-1} \frac{\partial S}{\partial \psi_j} S^{-1} (y - \mu) - (y - \mu)' S^{-1} \frac{\partial \mu}{\partial \psi_j} \end{aligned} \quad (9)$$

where

$$S = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_K \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix} \quad (10)$$

This requires, at every time point k , $p + 2$ more prediction and update steps in order to compute the terms $\frac{\partial S_k}{\partial \theta_j}$'s, $\frac{\partial \mu_k}{\partial \theta_j}$'s, $\frac{\partial S_k}{\partial \rho_0}$, $\frac{\partial \mu_k}{\partial \rho_0}$, $\frac{\partial S_k}{\partial \rho_1}$ and $\frac{\partial \mu_k}{\partial \rho_1}$, where p is the dimension of θ . These are obtained by deriving the equations in (4) and (6) w.r.t. θ , ρ_0 and ρ_1 , as shown in Section S.2.1 of **Supplementary Information**. All the results obtained from every prediction/update step at each time point t_k , along with the corresponding derivatives, are then used to compute the energy function $\varphi(\theta, \rho_0, \rho_1)$ and its gradient which, in turn, are used for the optimization step. The proposed extended Kalman filter is summarised in Algorithm 1 from Section S.4 of **Supplementary Information**. The whole procedure returns the estimated parameters $\hat{\theta}_{ekf}$, $\hat{\rho}_{0ekf}^2$, $\hat{\rho}_{1ekf}^2$ and the first two-order moments $m_{k|K}^s$ and $P_{k|K}^s$ of the smoothing distribution $p(x_k|y_{1:K}, \theta, \rho_0, \rho_1)$ at every time point t_k , $k = 1, \dots, K$. All the integrals involved for the computation of P_k^* , $\frac{\partial}{\partial \theta_j} m_k^*$, $\frac{\partial}{\partial \theta_j} P_k^*$, $\frac{\partial}{\partial \rho_0} P_k^*$ and $\frac{\partial}{\partial \rho_1} P_k^*$ are estimated using a 3rd-order Gauss-Legendre method [26].

2.2 Stochastic formulation of clonal dynamics

We assume that the time counting process

$$X_t = (X_{1t}, \dots, X_{Nt}) \quad (11)$$

of N distinct cell types for a single clone evolves in a time interval $(t, t + \Delta t)$ according to a set of reactions and hazard functions defined as

$$v_k = \begin{cases} (\dots 1_i \dots)' \\ (\dots - 1_i \dots)' \\ (\dots - 1_i \dots 2_j \dots)' \end{cases} \quad h_k(X_t, \theta_i) = \begin{cases} X_{it} \alpha_i \\ X_{it} \delta_i \\ X_{it} \lambda_{ij} \end{cases} \quad (12)$$

The hazard functions contain linear terms for duplication and death of cell i with positive rates α_i and δ_i , and a linear term to describe cell

differentiation from lineage i to lineage j with positive rate λ_{ij} for each $i \neq j = 1, \dots, N$. Finally, we use the compact matrix formulations

$$\begin{aligned} V &= [v_1 \dots v_K] \\ h(X_t; \theta) &= [h_1(X_t; \theta) \dots h_K(X_t; \theta)]' \end{aligned} \quad (13)$$

where θ is the vector of all the unknown parameters describing the dynamics. Since for our applications both the HSCs and the progenitors P_i s are missing states, to help parameter inference of the state space model (2)-(3) combined with net-effect matrix and hazard vector (12) we assume the following conservation laws

$$\lambda_{HSC \rightarrow P_i} = \sum_j \lambda_{P_i \rightarrow X_j} \quad (14)$$

where X_j s are all the offspring cell types generated by P_i .


2.3 Transition probabilities

Once the vector parameter θ is estimated for a particular model \mathcal{M} , we define the transition probability p_{ij} from cell type i to cell type j as

$$p_{ij} = \frac{\lambda_{ij} + \alpha_i}{\sum_{k \in S_i(\mathcal{M})} \lambda_{ik}} \quad (15)$$

where $S_i(\mathcal{M})$ is the set of all the possible target cell types associated to cell type i in the model \mathcal{M} .

2.4 Computational implementation

The stochastic framework is implemented in the  package Karen available at <https://github.com/delcore-luca/Karen>. Working examples showing the usage of the package are provided in Section S.7 of **Supplementary Information**.

3 Results

We use our stochastic framework to compare four different biologically-sustained models of hematopoiesis. The graphical representation of the candidate models is shown in Figure 2. For each candidate model, the stochastic differential equation formulation can be obtained from equations (12) - (13). Biological interpretation of the proposed models can be found in Section S.6 of the **Supplementary Information**.

Application to simulated data

We performed several simulations designed to test the proposed inferential procedure under different scenarios. The performance have been investigated by: (i) reducing the number of time points, (ii) reducing the fraction of clones recaptured across lineages and time, which is equivalent to increasing the rate of false-negative errors, (iii) increasing measurement noise, and (iv) selecting a cell differentiation structure among a set of candidates. Additional details on the simulation studies can be found in Section S.3 of the **Supplementary Information**. The results show the accurate recovery by the method of the true parameters, the first two-order smoothing moments, and the true generative model.

Application to Rhesus Macaques study

We analyzed an in-vivo clonal tracking dataset previously used to investigate the hematopoietic reconstitution in Rhesus Macaques [27]. A pool of autologous CD34+ HSPCs barcoded by using lentiviral vectors have been transplanted in three myeloablated animals [28; 29]. Following engraftment, Granulocytes (G), Monocytes (M), T, B and NK cells were

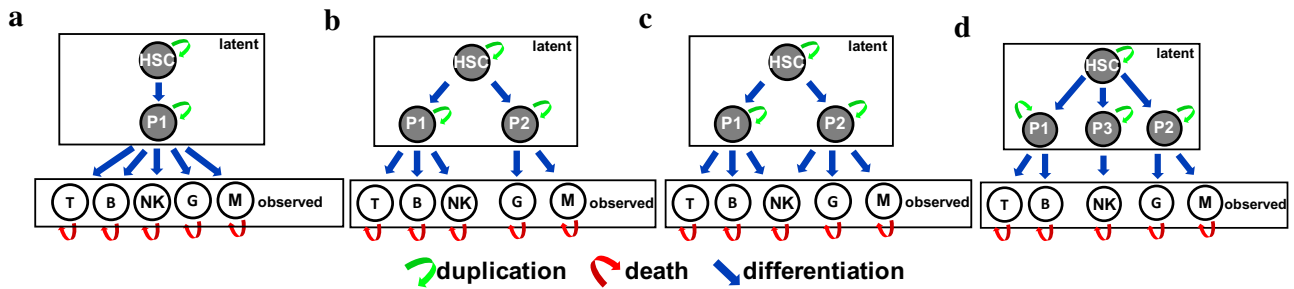


Fig. 2: Graphical representation of the candidate models (a-d): Latent and observed cell types are indicated with grey and white nodes respectively. Red arrows denote a death move, green arrows indicate a duplication move, and blue arrows a differentiation move.

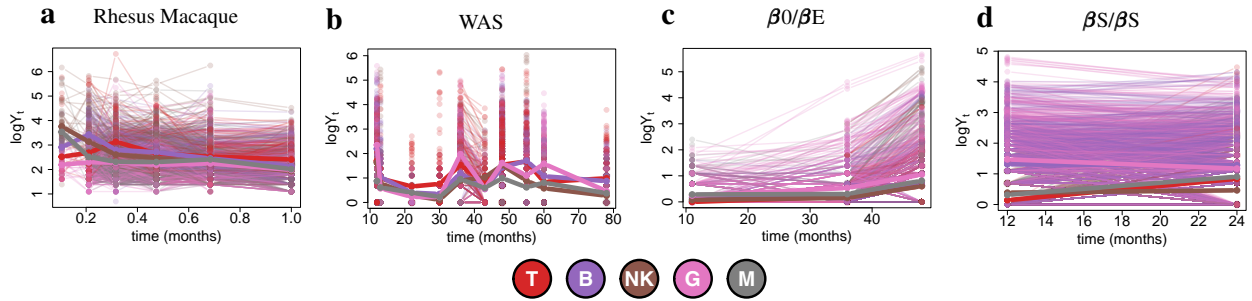


Fig. 3: Clonal tracking data: Logarithmic clonal abundance (y -axis) over time (x -axis) in each cell lineage (colors) for the rhesus macaque study (a) and the clinical trials (b-d).

flow-sorted from peripheral blood (purity median 98.8%), and the majority of transduced cells contained only one barcode. Barcode retrieval by PCR was performed on purified hematopoietic lineage samples monthly for 4.5 months (ZG66), 6.5 months (ZH17), and 9.5 months (ZH33) [30]. Further details on transductions protocol and culture conditions can be found in the original paper study [27]. Although the sample DNA amount was maintained constant during the whole experiment (200 ng for ZH33 and ZG66 or 500 ng for ZH17), the sample collected resulted in different magnitudes of the total number of reads (see Table S.1 in **Supplementary Information**). This discrepancy makes samples not directly comparable across time and cell types. Therefore we rescaled the barcode counts as described in Section S.5 of the **Supplementary Information**. We report the rescaled cell counts, at the clonal level, in Figure 3. The total numbers of clones collected are 1165 (ZH33), 1280 (ZH17), and 1291 (ZG66). We only focused on the top 1000 most recaptured clones across lineages and time to further remove bias.

We fit the four candidate models on the clonal tracking data using Algorithm 1 from Section S.4 of **Supplementary Information**. We report the results in Figure 4 which shows, for each candidate model, the estimated cell differentiation network and the corresponding Akaike Information Criterion (AIC) [31] which we use as a measure of model selection. According to the AIC, model (c) is the one that best fits the clonal tracking data collected from the rhesus macaque study. This result suggests that the classical/dichotomic model (a) fails to describe adequately clonal dynamics in rhesus macaque, whereas the myeloid-based developmental model (c) better explains hematopoietic reconstitution. Therefore our proposed framework Karen clearly indicates that in primate hematopoiesis myeloid progenitors represent a prototype of hematopoietic cells capable to produce both myeloid G/M cells and NK cells.

Application to gene therapy clinical trials

Clonal tracking data derived from the analysis of samples isolated from six patients treated using HSPC-based gene therapy have been used to

investigate human hematopoiesis dynamics. 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) [32], 2 patients with β hemoglobinopathy (1 with $\beta S/\beta S$ sickle cell disease [33] and 1 with $\beta 0/\beta E$ β thalassemia [34]).

Details on procedures, gene therapy protocols, and normalization methods can be found in [32–35]. We report the clonal level logarithmic cell counts in Figure 3. Since data was already normalized to compensate for unbalanced sampling in VCN and DNA [32–35], we did not apply any further transformation. The total clones collected are 156654, 17273, and 230408, respectively, for WAS, $\beta S/\beta S$ and $\beta 0/\beta E$ clinical trials. The following results derive from the analysis of the 1000 most recaptured clones in each clinical trial (top 250 clones per WAS patient).

The same four biologically motivated hematopoietic models (Figure 2) have been scored separately in each clinical trial using our stochastic framework Karen. We report the results in Figure 4 showing the estimated cell differentiation networks for each clinical trial. As a result, according to the AIC, model (d) is the one that always best fits clonal tracking data collected from each clinical trial, thus suggesting that a three-branches developmental model better explains hematopoietic reconstitution in humans after a gene therapy treatment. In particular, while lymphoid T/B and myeloid G/M develop in parallel through separate branches from different progenitors, there is a third developmental branch for the lymphoid NK cells which is independent from the first two branches.

4 Discussion

We have proposed a novel stochastic framework for modeling 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 framework extends stochastic quasi-reaction networks by introducing

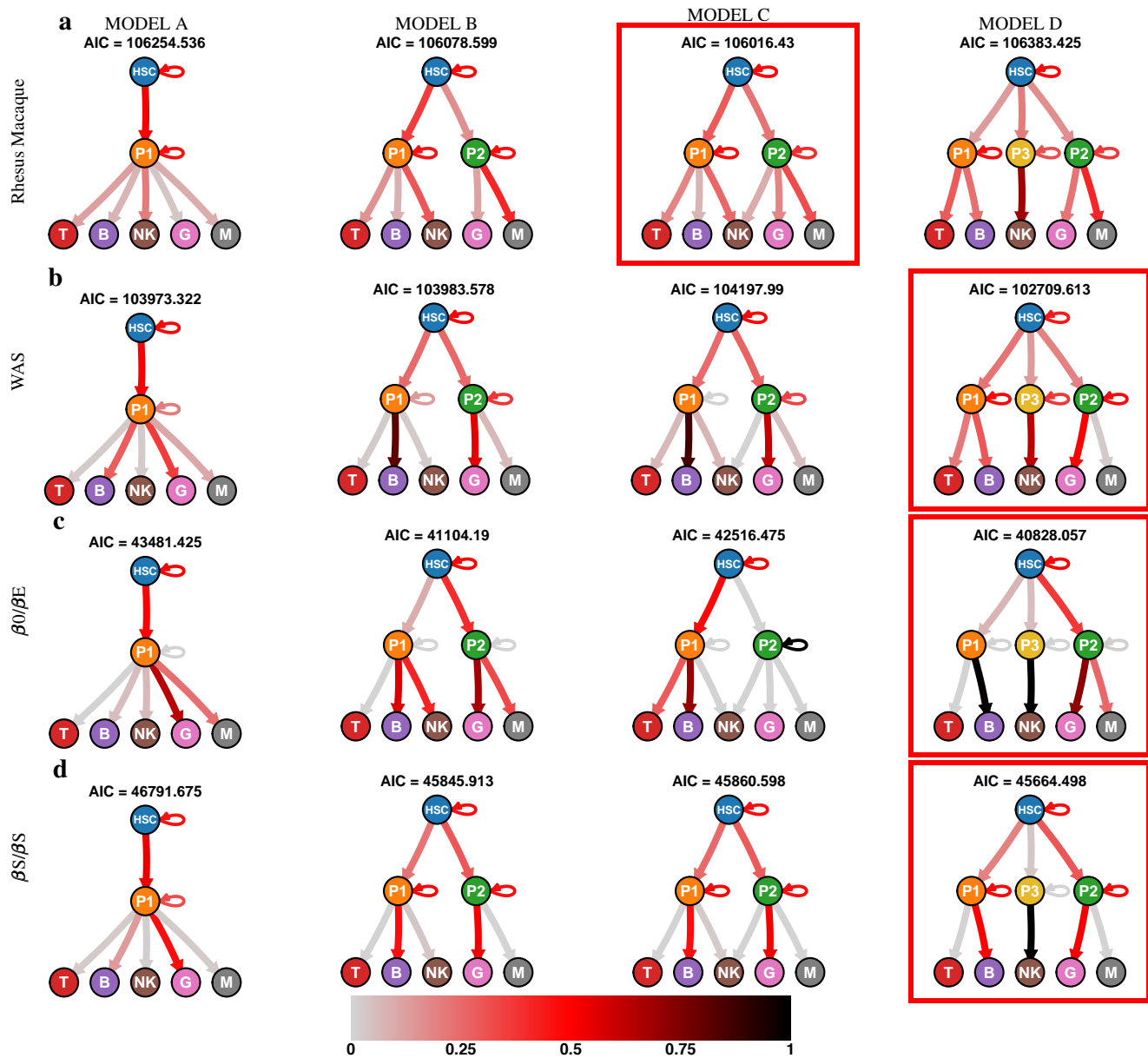


Fig. 4: Results: Inferred cell differentiation networks for the candidate models (columns) for the rhesus macaque study (a) and the clinical trials (b-d). Each arrow is weighted and coloured according to the corresponding transition probability estimated with (15). For each model the AIC is reported and the best model is squared with a red box.

EKF and RTS components. We developed a tailor-made Expectation-Maximization (EM) 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-order smoothing moments of all the process states, and model selection. Simulation results indicated the method's robustness in situations characterized by: the availability of a limited number of time points, limited clonal recapture, and high levels of measurement noise.

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 behaviors. This limitation can be overcome by using a distribution-free approach, such as the Kernel Kalman Rule [36; 37]. Another limitation is that our framework considers reaction rates constant for the whole study period. Extensions that allow for modeling reaction rates as spline functions of

time or depending on clinically relevant variables are within reach and will be the goal of future research.

The application of Karen on 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 myeloid and lymphoid progenitors P1 and P2. Results are consistent with the ones previously reported in [27] where the authors demonstrated the presence of distinct subpopulations within the NK lineage, potentially deriving from alternative maturation processes. Subsequently, we analyzed in-vivo clonal tracking data from three different clinical trials, showing consistency in the selected hematopoietic model structure across the clinical trials. Our stochastic framework can support biologists in understanding hematopoietic 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 cells methods. Applications in alternative contexts, such as the modeling of population dynamics, where similar issues about partial sampling and varying levels of measurement noise are present, could also be explored.

Acknowledgements and Funding

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

Author Contributions

All authors contributed to analysing the data and writing the manuscript. L.D.C. designed and implemented the stochastic framework.

References

- [1] E. McCulloch and J. Till, "Proliferation of hemopoietic colony-forming cells transplanted into irradiated mice," *Radiation Research*, vol. 22, no. 2, pp. 383–397, 1964.
- [2] E. A. McCulloch and J. E. Till, "Perspectives on the properties of stem cells," *Nature Medicine*, vol. 11, no. 10, pp. 1026–1028, 2005.
- [3] M. C. Mackey, "Cell kinetic status of haematopoietic stem cells," *Cell Proliferation*, vol. 34, no. 2, pp. 71–83, 2001.
- [4] C. Haurie, D. C. Dale, and M. C. Mackey, "Cyclical neutropenia and other periodic hematological disorders: a review of mechanisms and mathematical models," *Blood, The Journal of the American Society of Hematology*, vol. 92, no. 8, pp. 2629–2640, 1998.
- [5] C. Haurie, D. C. Dale, and M. C. Mackey, "Occurrence of periodic oscillations in the differential blood counts of congenital, idiopathic, and cyclical neutropenic patients before and during treatment with g-csf," *Experimental Hematology*, vol. 27, no. 3, pp. 401–409, 1999.
- [6] C. Haurie, D. C. Dale, R. Rudnicki, and M. C. Mackey, "Modeling complex neutrophil dynamics in the grey collie," *Journal of Theoretical Biology*, vol. 204, no. 4, pp. 505–519, 2000.
- [7] C. Haurie, R. Person, D. C. Dale, and M. C. Mackey, "Hematopoietic dynamics in grey collies," *Experimental Hematology*, vol. 27, no. 7, pp. 1139–1148, 1999.
- [8] H. Kawamoto, H. Wada, and Y. Katsura, "A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model," *International Immunology*, vol. 22, no. 2, pp. 65–70, 2010.
- [9] J. E. Till, E. A. McCulloch, and L. Siminovitich, "A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 51, no. 1, p. 29, 1964.
- [10] D. Pellin, L. Biasco, A. Aiuti, M. C. Di Serio, and E. C. Wit, "Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking," *Applied Network Science*, vol. 4, no. 1, pp. 1–26, 2019.
- [11] J. Xu, S. Koelle, P. Gutterop, C. Wu, C. Dunbar, J. L. Abkowitz, and V. N. Minin, "Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis," *The Annals of Applied Statistics*, vol. 13, no. 4, pp. 2091–2119, 2019.
- [12] M. A. Newton, P. Gutterop, S. Catlin, R. Assunção, and J. L. Abkowitz, "Stochastic modeling of early hematopoiesis," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1146–1155, 1995.
- [13] I. Roeder and M. Loeffler, "A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity," *Experimental Hematology*, vol. 30, no. 8, pp. 853–861, 2002.
- [14] I. Roeder, L. M. Kamminga, K. Braesel, B. Dontje, G. de Haan, and M. Loeffler, "Competitive clonal hematopoiesis in mouse chimeras explained by a stochastic model of stem cell organization," *Blood*, vol. 105, pp. 609–616, 01 2005.
- [15] D. Dingli and J. M. Pacheco, "Modeling the architecture and dynamics of hematopoiesis," *WIREs Systems Biology and Medicine*, vol. 2, no. 2, pp. 235–244, 2010.
- [16] S. N. Catlin, J. L. Abkowitz, and P. Gutterop, "Statistical inference in a two-compartment model for hematopoiesis," *Biometrics*, vol. 57, no. 2, pp. 546–553, 2001.
- [17] Y.-H. Kim, Y. Song, J.-K. Kim, T.-M. Kim, H. W. Sim, H.-L. Kim, H. Jang, Y.-W. Kim, and K.-M. Hong, "False-negative errors in next-generation sequencing contribute substantially to inconsistency of mutation databases," *PLOS ONE*, vol. 14, no. 9, p. e0222535, 2019.
- [18] K. Robasky, N. E. Lewis, and G. M. Church, "The role of replicates for error mitigation in next-generation sequencing," *Nature Reviews Genetics*, vol. 15, no. 1, pp. 56–62, 2014.
- [19] D. Bobo, M. Lipatov, J. Rodriguez-Flores, A. Auton, and B. Henn, "False negatives are a significant feature of next generation sequencing callsets," 2016.
- [20] S. A. Hardwick, I. W. Deveson, and T. R. Mercer, "Reference standards for next-generation sequencing," *Nature Reviews Genetics*, vol. 18, no. 8, pp. 473–484, 2017.
- [21] A. Petrackova, M. Vasinek, L. Sedlarikova, T. Dyskova, P. Schneiderova, T. Novosad, T. Papajik, and E. Kriegova, "Standardization of sequencing coverage depth in ngs: recommendation for detection of clonal and subclonal mutations in cancer diagnostics," *Frontiers in Oncology*, p. 851, 2019.
- [22] E. Allen, *Modeling with Itô stochastic differential equations*, vol. 22. Springer Science & Business Media, 2007.
- [23] A. D. Polyanin and V. F. Zaitsev, *Handbook of ordinary differential equations: exact solutions, methods, and problems*. CRC Press, 2017.
- [24] M. Behr, P. Benner, and J. Heiland, "Solution formulas for differential sylvester and lyapunov equations," *Calcolo*, vol. 56, no. 4, pp. 1–33, 2019.
- [25] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.
- [26] P. J. Davis and P. Rabinowitz, *Methods of numerical integration*. Courier Corporation, 2007.
- [27] C. Wu, B. Li, R. Lu, S. J. Koelle, Y. Yang, A. Jares, A. E. Krouse, M. Metzger, F. Liang, K. Loré, et al., "Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells," *Cell Stem Cell*, vol. 14, no. 4, pp. 486–499, 2014.
- [28] H. J. Kim, J. F. Tisdale, T. Wu, M. Takatoku, S. E. Sellers, P. Zickler, M. E. Metzger, B. A. Agricola, J. D. Malley, I. Kato, et al., "Many multipotential gene-marked progenitor or stem cell clones contribute to hematopoiesis in nonhuman primates," *Blood, The Journal of the American Society of Hematology*, vol. 96, no. 1, pp. 1–8, 2000.
- [29] B. E. Shepherd, H.-P. Kiem, P. M. Lansdorp, C. E. Dunbar, G. Aubert, A. LaRochelle, R. Seggewiss, P. Gutterop, and J. L. Abkowitz, "Hematopoietic stem-cell behavior in nonhuman primates," *Blood, The Journal of the American Society of Hematology*, vol. 110, no. 6, pp. 1806–1813, 2007.
- [30] R. Lu, N. F. Neff, S. R. Quake, and I. L. Weissman, "Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding," *Nature biotechnology*, vol. 29, no. 10, pp. 928–933, 2011.
- [31] K. P. Burnham, D. R. Anderson, and K. P. Huyvaert, "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 23–35, 2011.
- [32] S. H.-B. Abina, H. B. Gaspar, J. Blondeau, L. Caccavelli, S. Charrier, K. Buckland, C. Picard, E. Six, N. Himoudi, K. Gilmour, et al., "Outcomes following gene therapy in patients with severe wiskott-aldrich syndrome," *Jama*, vol. 313, no. 15, pp. 1550–1563, 2015.
- [33] J.-A. Ribeil, S. Haccin-Bey-Abina, E. Payen, A. Magnani, M. Semeraro, E. Magrin, L. Caccavelli, B. Neven, P. Bourget, W. El Nemer, et al., "Gene therapy in a patient with sickle cell disease," *New England Journal of Medicine*, vol. 376, no. 9, pp. 848–855, 2017.
- [34] A. A. Thompson, M. C. Walters, J. Kwiatkowski, J. E. Rasko, J.-A. Ribeil, S. Hongeng, E. Magrin, G. J. Schiller, E. Payen, M. Semeraro, et al., "Gene therapy in patients with transfusion-dependent β -thalassemia," *New England Journal of Medicine*, vol. 378, no. 16, pp. 1479–1493, 2018.
- [35] E. Sherman, C. Nobles, C. C. Berry, E. Six, Y. Wu, A. Dryga, N. Malani, F. Male, S. Reddy, A. Bailey, et al., "Inspired: a pipeline for quantitative analysis of sites of new dna integration in cellular genomes," *Molecular Therapy-Methods & Clinical Development*, vol. 4, pp. 39–49, 2017.
- [36] G. H. Gebhardt, A. Kupcsik, and G. Neumann, "The kernel kalman rule—efficient nonparametric inference with recursive least squares," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [37] G. H. W. Gebhardt, A. Kupcsik, and G. Neumann, "The kernel kalman rule," *Machine Learning*, vol. 108, pp. 2113–2157, Dec 2019.