

University of Groningen

## The stochastic route of haematopoiesis

Del Core, Luca

DOI:  
[10.33612/diss.603416389](https://doi.org/10.33612/diss.603416389)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2023

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*  
Del Core, L. (2023). *The stochastic route of haematopoiesis: modelling and inference methods in clonal tracking studies*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.603416389>

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

# **The stochastic route of haematopoiesis**

Modelling and inference methods in clonal tracking studies

**Luca Del Core**



**university of  
 groningen**

faculty of science  
 and engineering

bernoulli institute

ISBN: 9789083310954

The research output presented in this thesis was carried out at the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen (The Netherlands).

Cover design: made by Luca Del Core

Printed by: Print Service Ede



university of  
 groningen

# The stochastic route of haematopoiesis

Modelling and inference methods in clonal tracking studies

**PhD thesis**

to obtain the degree of PhD at the  
 University of Groningen  
 on the authority of the  
 Rector Magnificus Prof. C. Wijmenga  
 and in accordance with  
 the decision by the College of Deans.

This thesis will be defended in public on

Monday 3 April 2023 at 11:00 hours

by

**Luca Del Core**

born on 8 September 1989  
 in Taranto, Italy

## **Supervisors**

Prof. M.A. Grzegorzcyk  
Prof. E.C. Wit

## **Assessment Committee**

Prof. D. Husmeier  
Prof. C.J. Albers  
Prof. C. Di Serio

*Dedicated to my family*



# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Aim. . . . .	2
1.2	Thesis outline . . . . .	3
1.3	Clonal tracking of haematopoiesis . . . . .	5
1.4	Current strategies and our proposal . . . . .	6
1.4.1	Cell differentiation . . . . .	6
1.4.2	Clonal dominance . . . . .	8
1.4.3	Clonal diversity . . . . .	8
	References . . . . .	13
<b>2</b>	<b>A stochastic state space model of cell differentiation</b>	<b>17</b>
2.1	Introduction . . . . .	19
2.2	Methods. . . . .	20
2.2.1	A stochastic model for cell differentiation . . . . .	20
2.2.2	State space formulation . . . . .	22
2.2.3	Optimal filtering and smoothing. . . . .	22
2.2.4	Transition probabilities . . . . .	24
2.2.5	Reaction constraints . . . . .	24
2.2.6	Model selection . . . . .	24
2.2.7	Computational implementation . . . . .	25
2.3	Results. . . . .	25
2.3.1	In-silico validation studies . . . . .	25
2.3.2	Genotoxicity study . . . . .	26
2.3.3	Rhesus Macaques study . . . . .	27
2.3.4	Gene therapy clinical trials . . . . .	28
2.4	Discussion . . . . .	29
2.5	Availability of data and materials . . . . .	30



<b>Appendices</b>	<b>39</b>
2.A Stochastic quasi-reaction networks . . . . .	39
2.B The Master Equation . . . . .	40
2.C Monte Carlo simulation of Ito-SDEs . . . . .	42
2.D Differential Sylvester Equation . . . . .	43
2.E Integrating factor method . . . . .	43
2.F Kalman Reaction Networks (Karen) . . . . .	44
2.G Simulation studies . . . . .	50
2.H Comparison with the state-of-the-art . . . . .	52
2.I Genotoxicity data rescaling . . . . .	53
2.J Rhesus macaque data rescaling. . . . .	54
2.K Hematopoietic models . . . . .	54
References . . . . .	66
<b>3 Mixed-effects reaction networks of clonal dominance</b>	<b>69</b>
3.1 Background. . . . .	71
3.2 Methods. . . . .	72
3.2.1 A stochastic model for cell differentiation . . . . .	73
3.2.2 LLA formulation of clonal dominance . . . . .	74
3.2.3 Inference procedure . . . . .	76
3.2.4 Model selection . . . . .	77
3.3 Results. . . . .	78
3.3.1 In silico validation study. . . . .	78
3.3.2 Comparison with GLS method . . . . .	79
3.3.3 Clonal dynamics in rhesus macaques . . . . .	81
3.3.4 Genotoxic effects on clonal dynamics. . . . .	83
3.4 Discussion and conclusion . . . . .	86
3.5 Availability of data and materials . . . . .	88
<b>Appendices</b>	<b>101</b>
3.A $\tau$ -leaping algorithm . . . . .	101
3.B Euler-Maruyama approximation . . . . .	102
3.C Inference of the base GLM model . . . . .	103
3.D Random-effects reaction networks. . . . .	104
References . . . . .	109

---

<b>4</b>	<b>A normalized measure of clonal diversity</b>	<b>115</b>
4.1	Introduction . . . . .	116
4.2	Potential artefacts in clonal tracking. . . . .	119
4.3	Methods: Shape Constrained Splines . . . . .	121
4.3.1	Definition of the model . . . . .	121
4.3.2	Shape-constrained splines normalization . . . . .	122
4.3.3	Inference procedure . . . . .	125
4.3.4	Pseudocode . . . . .	128
4.4	Applications of SCS in NGS data . . . . .	128
4.4.1	In-vitro assay . . . . .	128
4.4.2	Viral vector safety in a genotoxicity study. . . . .	131
4.5	Discussion . . . . .	135
4.6	Availability of data and materials . . . . .	136
	<b>Appendices</b>	<b>145</b>
4.A	In-vitro Assay. . . . .	145
4.B	Supplementary figures . . . . .	145
4.B.1	Fitting quadratic and cubic splines . . . . .	145
4.B.2	Comparisons of rescaling methods . . . . .	146
	References . . . . .	156
	<b>Summary</b>	<b>163</b>
	References . . . . .	166
	<b>Samenvatting</b>	<b>167</b>
	Bibliografie . . . . .	171
	<b>Acknowledgements</b>	<b>173</b>
	<b>Biography</b>	<b>175</b>



# 1

## INTRODUCTION

## ABSTRACT

*In this chapter we first state the goals of this work and we list the thesis content. The rest of the chapter is then aimed at providing relevant biology background, such as the concept of haematopoiesis, and its connection to gene therapy clonal tracking studies. Finally we review the state of the art methodologies for investigating haematopoietic reconstitution in terms of cell differentiation, clonal dominance, and clonal diversity. We first briefly discuss on their limitations, and then we introduce our proposed methods.*

### 1.1. AIM

Mathematical models of haematopoiesis are increasingly in demand since they can provide relevant insights on how blood cells develop, thus supporting the design of novel clinical strategies. Clonal tracking is a recent high-throughput technology that allows to quantitatively calibrate such mathematical models. In gene therapy clonal tracking studies, cell differentiation networks describe the cellular hierarchical relationships involved in haematopoiesis, whereas clonality is aimed at quantifying the clonal population complexity (clonal diversity) and to early detect therapy side effects, such as events of clonal dominance. The work of this thesis is aimed at providing stochastic frameworks to shed more light on those mechanisms given the information provided by clonal tracking data. We investigate cell differentiation using stochastic quasi-reaction networks, a framework that allows to model stochastic biochemical reactions. Since clonal tracking data suffers from missing cell types and false negative errors, we combined stochastic quasi-reaction networks with extended Kalman filtering, leading to our Kalman Reaction Networks framework (Karen). We first test our state space model in several simulation studies, showing an accurate recovery of the true parameters and the generative differentiation structure. Then we use our proposed method to compare different biologically

plausible models of cell differentiation in five in-vivo clonal tracking studies. Subsequently, we combine stochastic quasi-reaction networks with random-effects (RestoreNet) to provide clone-specific expansion parameters, thus allowing to detect possible events of clonal dominance. We have shown in synthetic studies and in two in-vivo models that our framework RestoreNet is able to detect events of clonal dominance, and thus it can provide statistical support in gene therapy safety studies. Finally, to objectively assess clonal diversity, we have developed a shape-constrained regression approach (SCS) that removes the effect of several technical artefacts from the Shannon entropy index, thus providing an unbiased diversity measure. Our SCS-rescaling method was first validated in a specifically designed in-vitro assay, and then used to objectively evaluate the impact of vector genotoxicity on the entropy decays of tumor prone mice.

## 1.2. THESIS OUTLINE

The thesis content is structured as follows:

- **Chapter 1** provides relevant biology background, such as cell differentiation and clonal complexity, their connection with gene therapy, and how these can be computationally investigated by means of next generation sequencing (NGS) clonal tracking. Subsequently, (i) we report the state-of-the-art methodologies for investigating haematopoiesis in clonal tracking studies, (ii) we discuss on their limitations, and (iii) we briefly introduce our proposed methods.
- **Chapter 2** focuses on our proposed Kalman reaction network framework Karen and its application to infer cell differentiation networks from clonal tracking data. Our framework is based on stochastic reaction networks combined with extended Kalman filtering and Rauch-Tung-Striebel smoothing. We calibrate the parameters of our framework on typical clonal tracking data subject to measurement noise, false-negative errors, and systematically unobserved cell types. Given a clonal tracking dataset and a set of candidate network structures, Karen infers the unknown dynamics parameters, the generative differentiation structure, and the first two smoothing moments. After introducing the method, we first validate it with

several in-silico studies, then (i) we compare the dynamics of cell differentiation in tumour-prone mice that were treated with two different viral vector designs, (ii) we analyse cell differentiation in an in-vivo study of Rhesus Macaques, and (iii) we infer cell differentiation in gene-therapy treated patients from three distinct clinical trials. In each in-vivo study we compared different plausible models of cell differentiation. Our tool can provide statistical support in gene therapy clonal tracking studies to better understand clonal reconstitution dynamics.

- In **Chapter 3** we introduce our proposed random-effects stochastic framework RestoreNet to detect possible events of clonal dominance in gene therapy safety studies. In particular, starting from an Ito-type formulation of a stochastic reaction network, the dynamics of cells duplication, death and differentiation at clonal level without clonal dominance is described by a local linear approximation. The parameters of this model are assumed to be shared across the clones. In order to incorporate the possibility of clonal dominance, we extend the base model by introducing random effects for the clonal parameters. This extended formulation is estimated using a tailor-made expectation maximization algorithm. We first validate our framework with several in-silico studies, then (i) we analyse a clonal tracking dataset from a rhesus macaque study, and (ii) we compare the dynamics of clonal expansion in a genotoxicity mice study. Our proposed framework can guide gene therapy surveillance studies to detect possible adverse events of clonal dominance.
- In **Chapter 4** we propose a shape-constrained method to quantify and remove the effect of technical artefacts from the Shannon entropy index, so as to get an unbiased measure of clonal diversity in gene therapy longitudinal studies. In particular, we first show that the Shannon diversity index, a well-established measure of heterogeneity of the clonal population, is affected by several technical confounders such as the DNA amount of the collected samples and the sequencing depth of the NGS library. In particular, with an exploratory analysis, we first provide evidence that a direct comparison across clonal tracking samples, collected under different

technical conditions, may provide misleading conclusions. Subsequently, (i) we define our shape-constrained rescaling method SCS, (ii) we validate it on a tailor-made in-vitro study, and (iii) we apply it to an in-vivo mice study to objectively evaluate the impact of vector genotoxicity on the entropy decays. Our method does not only allow to compare the complexity of different clonal populations, but does also provide insights on the artefacts mainly affecting diversity measurements.

### 1.3. CLONAL TRACKING OF HAEMATOPOIESIS

Haematopoiesis (from Greek *αἷμα*, 'blood' and *ποιεῖν* 'to make') is the process by which blood cells are produced by haematopoietic stem cells (HSCs) [1]. HSCs reside in the bone marrow niche and have the unique ability to give rise to all of the different mature blood cell types and tissues. HSCs are self-renewing cells, meaning that as soon as they differentiate, they can still produce HSCs, so as to ensure that the pool of stem cells is not exhausted. In an asymmetric HSC cell division a single HSC first duplicates into two HSCs, and then one of the two differentiates into a more specialised cell, such as a myeloid or lymphoid progenitor cell [2]. Otherwise, if both HSC copies keep the HSC status, this leads to a symmetric cell division [3]. The other daughter (non-HSC) cells can continue to follow their differentiation path towards more specific blood cell types, but cannot renew themselves, except for the multipotent intermediate progenitors that can be considered as only-transiently (short-term) self-renewing [4]. As a stem cell matures, several changes in gene expression occur that move the cell closer to the final cell type, and further limit its potential to differentiate. The dynamics of haematopoiesis may differ under normal (healthy) and malignant circumstances [5], and therefore understanding this process is important for therapeutic applications [6]. This is the case of gene therapy, a medical treatment aimed at curing a genetic disease at its source by delivering a functioning copy of the missing/corrupted gene which is causing the disease [7].

Several high-throughput technologies based on next generation sequencing (NGS) and single-cell systems, allow to investigate haematopoiesis both in in-vitro preclinical studies and in-vivo gene therapy studies [8].



NGS is a recent approach for DNA and RNA sequencing, which consists of a complex interplay of chemistry, hardware, optical sensors and software [9–12]. One of the most used NGS-based approaches in gene therapy is clonal tracking which consists in labelling the haematopoietic stem cells by the random insertion in its genome of several copies of a genetically-modified virus (viral vector). More precisely the labels, called the clones, are the genomic coordinates where the viral vector integrates on the HSC genome. As the stem cells grow and differentiate, all the offspring cells inherit the clones from their ancestor. During follow-up, the labels are collected from tissues and blood samples using bioinformatics pipelines. As a result, clonal tracking allows to calibrate mathematical models of clonal dynamics and hierarchical relationships of haematopoiesis. In particular, mathematical models of cell differentiation, clonal dominance and clonal complexity (divertisty) can provide useful insights for safety and efficacy of gene therapy strategies.

## **1.4. CURRENT STRATEGIES AND OUR PROPOSAL**

### **1.4.1. CELL DIFFERENTIATION**

Cell differentiation describes the cellular hierarchical relationships underlying haematopoiesis. Clarifying how HSCs differentiate into mature cell types is important for understanding how they attain specific functions and offers the potential for therapeutic manipulation [13]. All the models of cell differentiation that have been postulated can be divided into two major categories, such as deterministic and stochastic models [6]. While deterministic models assume that all the possible scenarios of cell differentiation are mainly characterized by deterministic (non-random) environmental stimulating factors, stochastic models assume that undifferentiated blood cells differentiate to specific cell types due to the randomness of the stimulating factors. Recent studies support the stochastic theory of cell differentiation by showing that the variability of certain factors characterizes the cellular population into several groups exhibiting different dynamics of differentiation [14]. There is also some evidence that stochasticity plays an important role in regulating apoptosis (cell death) and self-renewal, and that the process residing in the bone marrow aimed

at balancing cell production has a stochastic nature [15]. Despite the multiple studies in this regard, it has only been agreed that intermediate progenitor cells lose the ability to proliferate, and each progenitor type can produce one or several types of mature blood cells before exhausting its own lifespan. Besides, it has not been reached a consensus about how many types of progenitors exist in this intermediate stage and how they differentiate [16]. Therefore mathematical modelling and inference of cell differentiation from clonal tracking data can play a key role in shedding more light on those mechanisms.

Several mathematical models have been proposed to describe cell differentiation in-vivo. For example, Pellin et al. [17] proposed a continuous-time Markov model able to describe the process of cell differentiation from clonal tracking data. The inferential procedure provides parameter estimates and structure selection of the differentiation network. More recently Xu et al. [16] proposed a novel quantitative framework based on a continuous-time, multi-type branching process aimed at describing the mechanistic models of cell division and differentiation. In addition to the dynamic parameters, the framework allows to compare structurally distinct models of hematopoiesis using cross validation. The model takes also into account that data is partially observed, that is some of the cell types are not collected. Many other mathematical models aimed at describing cell differentiation have been proposed in research literature [18–23]. Still, none of the already existent tools takes into account that clonal tracking data contains many missing values due to either threshold detection failure or to false negative errors [24]. In this work we present a novel stochastic framework to investigate cell differentiation while prudentially treating all the missing values as latent states [25]. Our proposed framework Karen is a continuous-discrete state space model having a stochastic reaction network as dynamic model combined with a linear Gaussian measurement model that links the noisy observations to the underlying stochastic states. Parameters inference consists in three steps, such as Kalman filtering, Rauch-Tung-Striebel smoothing, and a non-linear constrained optimization problem that are iterated until convergence on the unknown parameters. A graphical representation of our proposed method is reported in Figure 2.1.

### 1.4.2. CLONAL DOMINANCE

Gene therapy treatments commonly use several multiple copies of a modified viral vector containing a functioning copy of the gene which is needed to cure the disease (high gene transfer rate) [26, 27]. But genetic modification of large numbers of cells is associated with the higher probability of unintentional vector insertions near proto oncogenes that may lead to insertional mutagenesis [28–30]. Insertional mutagenesis causes a significant change in clone fitness that can lead to the clones' abnormal expansion and to an unbalanced contribution of different clones to blood cells production. Thus, identifying such clones can guide the design of safer and more effective viral vector designs for gene therapies.

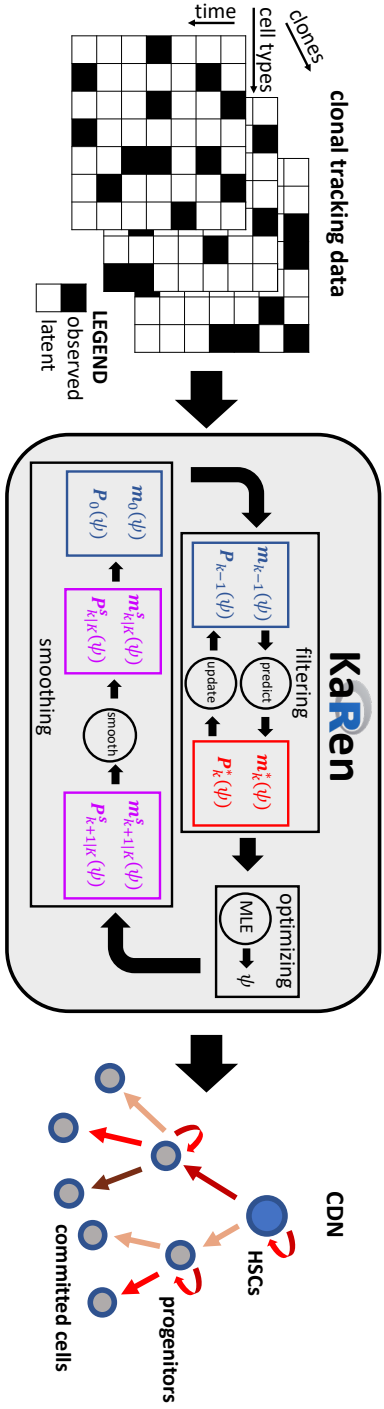
To the best of our knowledge, none of the existing mathematical models of clonal dynamics can identify clones that exhibit deviant behaviours during haematopoietic reconstitution. For example in the continuous-time Markov model proposed by Pellin et al. [17] all the dynamic parameters are shared across the clones, so that the model does not allow to identify specific clonal patterns of duplication and differentiation. Also, in the multi-type branching process of Xu et al. [16] the kinetics of cell division and differentiation are assumed to be the same for each clone, therefore not even this model is able to describe clone-specific growing dynamics. Here we propose a stochastic framework aimed at detecting possible events of clonal dominance based on clone-specific birth, death and differentiation dynamics [31]. Our proposed model combines Ito-type stochastic reaction networks with clone-specific random-effects for the dynamic parameters. In this formulation all the unknown parameters are estimated with a tailor-made expectation maximization algorithm. Our random-effects stochastic reaction networks technique (RestoreNet) provides a quantification of clonal dominance in terms of random expansion rates on a network of cell lineages. Furthermore, our proposed method can also be used to infer clone-specific dynamics of cell differentiation. Figure 3.2.1 shows a graphical representation of our proposed framework.

### 1.4.3. CLONAL DIVERSITY

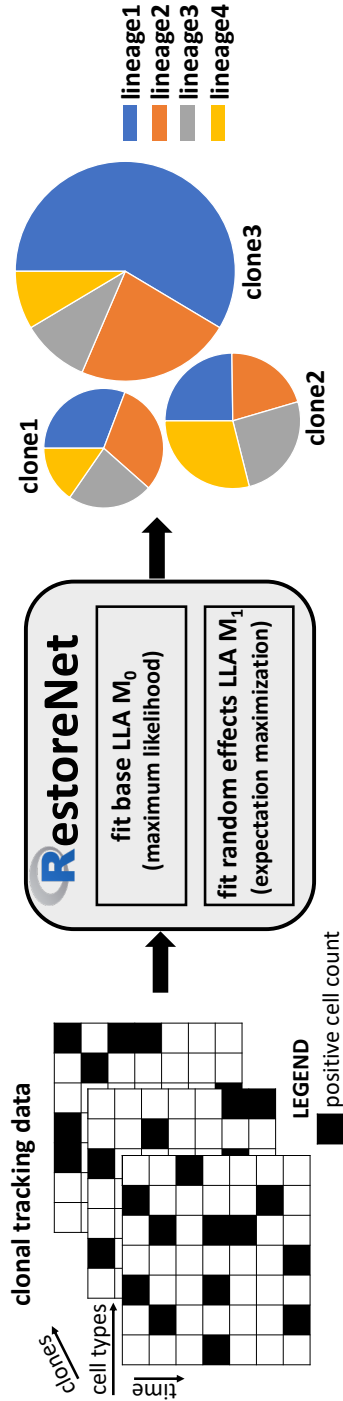
Clonal population complexity (clonality or clonal diversity) provides insights on the safety and efficacy of a gene therapy treatment. A high level of

clonality, called polyclonality, is characterized by a high number of distinct clones whose abundances are evenly distributed. Whereas a low level of clonality is characterized by either few distinct clones or by a degenerate distribution of their abundances to few clones, which we refer to as oligoclonality. Usually, a polyclonal population is associated with a normal (healthy) haematopoiesis, whereas an oligoclonal distribution suggests the occurrence of a malignant event, such as a disease progression. Therefore quantifying and monitoring clonality after a gene therapy treatment is important in safety studies to prevent or moderate potential side effects [32–34].

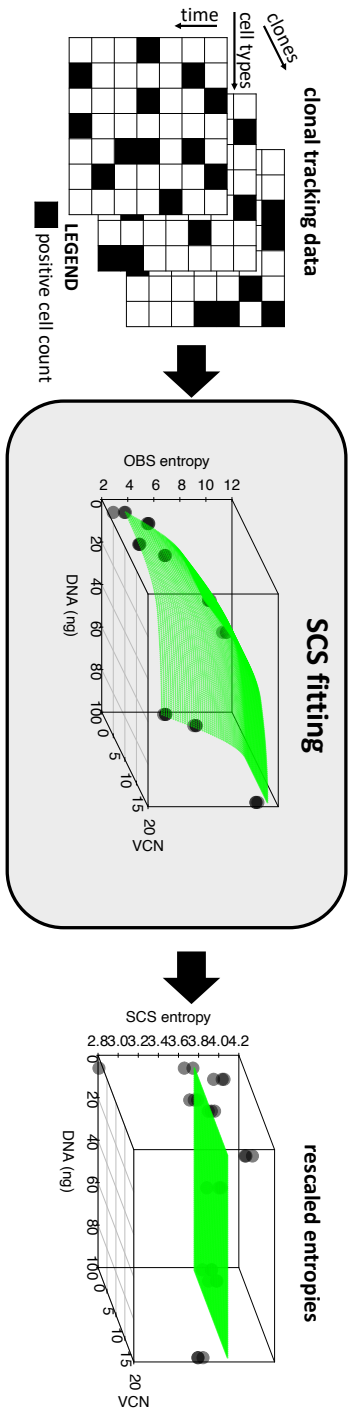
The Shannon entropy index is one of the most used measure of clonal diversity in medical applications [35, 36]. However, this index does not take into account the measurement noise due to technical variations and artefacts, such as the amount of the sequenced DNA, the sequencing depth and many other tuning parameters of the NGS platforms. Several rescaling methods have been proposed in research literature to remove the confounding effects, such as the rarefaction method [37] and its scaled version [38]. The application of such methods revealed that rarefied microbiome count data may be strongly biased [39–41]. Besides, none of the already existent methods can quantify the effect of each confounder on clonal diversity. In this work we propose a method based on the combination of the Shannon entropy index [35, 36] with shape-constrained splines (SCS) for objectively measure clonal complexity while taking into account measurement noise due to technical confounding factors [42]. Our proposed SCS method first quantifies the effect of each confounder on the observed Shannon entropies. The quantification is based on a B-spline basis whose shape is restricted according to a biological-sustained hypothesis, such as the positive correlation between the number of clones observed and the DNA amount of the sample being collected. Subsequently, those confounding effects are removed from the observed entropies by means of the estimated residuals. Effectively, our proposed SCS rescaling method allows to distinguish between biological and artefacts related changes in clonal complexity, thus providing an unbiased measure of clonal diversity. A graphical representation of our proposed method is reported in Figure 1.3.



**Figure 1.1 |** Schematic representation of Karen: A 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 step, and a smoothing step which are iterated until convergence on the unknown vector parameter  $\psi$ . Finally, a cell differentiation network (CDN) is returned from Karen, where each arrow is directed and weighted according to the estimated parameters (right panel).



**Figure 1.2** | Schematic representation of RestoreNet: A clonal tracking dataset (left) is received as input from our proposed stochastic framework RestoreNet (middle). It mainly consists in two parts, such as a maximum likelihood step to infer the fixed-effects model, and an expectation-maximization step to infer the mixed-effects formulation. Finally, a clonal piechart is returned, where each clone is identified by a pie whose slices are lineage specific and proportional to their expansion rates (right).



**Figure 1.3 |** Schematic representation of SCS: A clonal tracking dataset (left) is received as input from our proposed framework SCS (middle). It mainly consists in two parts, such as a the computation of the Shannon entropy index ( $\gamma$ -axis dots) and a shape-constrained fitting step (green surface). Finally, the rescaled Shannon entropies are returned (right).

## REFERENCES

- [1] I. Monga, K. Kaur, and S. K. Dhanda, *Revisiting hematopoiesis: applications of the bulk and single-cell transcriptomics dissecting transcriptional heterogeneity in hematopoietic stem cells*, Briefings in Functional Genomics **21**, 159 (2022).
- [2] J.-P. Tassan and J. Z. Kubiak, *Asymmetric Cell Division in Development, Differentiation and Cancer* (Springer, 2017).
- [3] S. J. Morrison and J. Kimble, *Asymmetric and symmetric stem-cell divisions in development and cancer*, Nature **441**, 1068 (2006).
- [4] J. F. Zhong, Y. Zhao, S. Sutton, A. Su, Y. Zhan, L. Zhu, C. Yan, T. Gallaher, P. B. Johnston, W. F. Anderson, *et al.*, *Gene expression profile of murine long-term reconstituting vs. short-term reconstituting hematopoietic stem cells*, Proceedings of the National Academy of Sciences **102**, 2448 (2005).
- [5] D. Peixoto, D. Dingli, and J. M. Pacheco, *Modelling hematopoiesis in health and disease*, Mathematical and Computer Modelling **53**, 1546 (2011).
- [6] M. Kimmel, *Stochasticity and determinism in models of hematopoiesis*, A Systems Biology Approach to Blood , 119 (2014).
- [7] T. Wirth, N. Parker, and S. Ylä-Herttua, *History of gene therapy*, Gene **525**, 162 (2013).
- [8] J. D. Robin, A. T. Ludlow, R. LaRanger, W. E. Wright, and J. W. Shay, *Comparison of DNA quantification methods for next generation sequencing*, Scientific Reports **6**, 24067 (2016).
- [9] C. Ledergerber and C. Dessimoz, *Base-calling for next-generation sequencing platforms*, Briefings in Bioinformatics **12**, 489 (2011).
- [10] F. Chang and M. M. Li, *Clinical application of amplicon-based next-generation sequencing in cancer*, Cancer Genetics, Cancer Genetics **206**, 413 (2013).
- [11] A. Kohlmann, V. Grossmann, and T. Haferlach, *Integration of Next-Generation Sequencing into clinical practice: Are we there yet?* Seminars in Oncology **39**, 26 (2012), molecular Pathogenesis of Hematologic Malignancies.
- [12] A. S. Gargis, L. Kalman, D. P. Bick, C. da Silva, D. P. Dimmock, B. H. Funke, S. Gowrisankar, M. R. Hegde, S. Kulkarni, C. E. Mason, R. Nagarajan, K. V. Voelkerding, E. A. Worthey, N. Aziz, J. Barnes, S. F. Bennett, H. Bisht, D. M. Church, Z. Dimitrova, S. R. Gargis, N. Hafez, T. Hambuch, F. C. L. Hyland, R. A. Luna, D. MacCannell, T. Mann, M. R. McCluskey, T. K. McDaniel, L. M. Ganova-Raeva, H. L. Rehm,



- J. Reid, D. S. Campo, R. B. Resnick, P. G. Ridge, M. L. Salit, P. Skums, L.-J. C. Wong, B. A. Zehnauer, J. M. Zook, and I. M. Lubin, *Good laboratory practice for clinical next-generation sequencing informatics pipelines*, *Nature Biotechnology* **33**, 689 (2015).
- [13] H. Kawamoto, H. Wada, and Y. Katsura, *A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model*, *International Immunology* **22**, 65 (2010).
- [14] H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang, *Transcriptome-wide noise controls lineage choice in mammalian progenitor cells*, *Nature* **453**, 544 (2008).
- [15] F. Q. Alenzi, B. Q. Alenazi, S. Y. Ahmad, M. L. Salem, A. A. Al-Jabri, and R. K. Wyse, *The haemopoietic stem cell: between apoptosis and self renewal*, *The Yale Journal of Biology and Medicine* **82**, 7 (2009).
- [16] 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* **13**, 2091 (2019).
- [17] 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* **4**, 1 (2019).
- [18] C. Di Serio, S. Scala, and P. Vicard, *Bayesian networks for cell differentiation process assessment*, *Stat* **9**, e287 (2020), e287 STAT-20-0009.R1, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.287>.
- [19] M. A. Newton, P. Gutterop, S. Catlin, R. Assunção, and J. L. Abkowitz, *Stochastic modeling of early hematopoiesis*, *Journal of the American Statistical Association* **90**, 1146 (1995), <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476621>.
- [20] I. Roeder and M. Loeffler, *A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity*, *Experimental Hematology* **30**, 853 (2002).
- [21] 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* **105**, 609 (2005), <https://ashpublications.org/blood/article-pdf/105/2/609/1706042/zh800205000609.pdf>.

- [22] D. Dingli and J. M. Pacheco, *Modeling the architecture and dynamics of hematopoiesis*, WIREs Systems Biology and Medicine **2**, 235 (2010), <https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.56>.
- [23] S. N. Catlin, J. L. Abkowitz, and P. Guttorp, *Statistical inference in a two-compartment model for hematopoiesis*, Biometrics **57**, 546 (2001).
- [24] 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 **14**, e0222535 (2019).
- [25] L. Del Core, D. Pellin, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data*, bioRxiv (2022).
- [26] O. S. Kustikova, A. Wahlers, K. Kühnlcke, B. Stähle, A. R. Zander, C. Baum, and B. Fehse, *Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population*, Blood **102**, 3934 (2003).
- [27] B. Fehse, O. Kustikova, M. Bubenheim, and C. Baum, *Pois (s) on—it's a question of dose...*, Gene Therapy **11**, 879 (2004).
- [28] C. Baum, J. Düllmann, Z. Li, B. Fehse, J. Meyer, D. A. Williams, and C. Von Kalle, *Side effects of retroviral gene transfer into hematopoietic stem cells*, Blood, The Journal of the American Society of Hematology **101**, 2099 (2003).
- [29] U. Modlich, O. S. Kustikova, M. Schmidt, C. Rudolph, J. Meyer, Z. Li, K. Kamino, N. Von Neuhoff, B. Schlegelberger, K. Kuehlcke, *et al.*, *Leukemias following retroviral transfer of multidrug resistance 1 (mdr1) are driven by combinatorial insertional mutagenesis*, Blood **105**, 4235 (2005).
- [30] C. Baum, O. Kustikova, U. Modlich, Z. Li, and B. Fehse, *Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors*, Human Gene Therapy **17**, 253 (2006).
- [31] L. Del Core, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic inference of clonal dominance in gene therapy studies*, bioRxiv (2022).
- [32] D. B. Kohn, C. Booth, E. M. Kang, S.-Y. Pai, K. L. Shaw, G. Santilli, M. Armant, K. F. Buckland, U. Choi, S. S. De Ravin, M. J. Dorsey, C. Y. Kuo, D. Leon-Rico, C. Rivat, N. Izotova, K. Gilmour, K. Snell, J. X.-B. Dip, J. Darwish, E. C. Morris, D. Terrazas, L. D. Wang, C. A. Bauser, T. Paprotka, D. B. Kuhns, J. Gregg, H. E. Raymond, J. K. Everett, G. Honnet, L. Biasco, P. E. Newburger, F. D. Bushman, M. Grez, H. B. Gaspar, D. A. Williams, H. L. Malech, A. Galy, A. J. Thrasher, K. F. Buckland, C. A. Bauser, H. B. Gaspar, A. J. Thrasher, and the Net4CGD consortium, *Lentiviral gene therapy for X-linked chronic granulomatous disease*, Nature Medicine **26**, 200 (2020).

- [33] C. F. Magnani, G. Gaipa, F. Lussana, D. Belotti, G. Gritti, S. Napolitano, G. Matera, B. Cabiati, C. Buracchi, G. Borleri, *et al.*, *Sleeping Beauty–engineered CAR T cells achieve antileukemic activity without severe toxicities*, *The Journal of Clinical Investigation* **130**, 6021 (2020).
- [34] S. Marktelt, S. Scaramuzza, M. P. Cicalese, F. Giglio, S. Galimberti, M. R. Lidonnici, V. Calbi, A. Assanelli, M. E. Bernardo, C. Rossi, *et al.*, *Intrabone hematopoietic stem cell gene therapy for adult and pediatric patients affected by transfusion-dependent  $\beta$ -thalassemia*, *Nature Medicine* **25**, 234 (2019).
- [35] S. Fuhrman, M. J. Cunningham, X. Wen, G. Zweiger, J. J. Seilhamer, and R. Somogyi, *The application of shannon entropy in the identification of putative drug targets*, *Biosystems* **55**, 5 (2000).
- [36] A. Monaco, N. Amoroso, L. Bellantuono, E. Lella, A. Lombardi, A. Monda, A. Tateo, R. Bellotti, and S. Tangaro, *Shannon entropy approach reveals relevant genes in alzheimer's disease*, *PLOS One* **14**, e0226190 (2019).
- [37] H. L. Sanders, *Marine benthic diversity: A comparative study*, *The American Naturalist* **102**, 243 (1968), <https://doi.org/10.1086/282541>.
- [38] K. P. Beule L, *Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities*. PeerJ (2020), 10.7717/peerj.9593.
- [39] P. J. McMurdie and S. Holmes, *Waste not, want not: Why rarefying microbiome data is inadmissible*, *PLOS Computational Biology* **10**, 1 (2014).
- [40] A. D. Willis, *Rarefaction, alpha diversity, and statistics*, *Frontiers in Microbiology* **10**, 2407 (2019).
- [41] R. H. Whittaker, *Evolution and measurement of species diversity*, *TAXON* **21**, 213 (1972), <https://onlinelibrary.wiley.com/doi/pdf/10.2307/1218190>.
- [42] L. Del Core, D. Cesana, P. Gallina, Y. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. Grzegorzczuk, *Normalization of clonal diversity in gene therapy studies using shape constrained splines*, *Scientific Reports* **12**, 1 (2022).


# 2

## **A STOCHASTIC STATE SPACE MODEL OF CELL DIFFERENTIATION**

---

Parts of this chapter have been published in “Stochastic inference of cell differentiation networks in gene therapy clonal tracking studies” [1].

## ABSTRACT

*Investigating cell differentiation under different disease settings offers the potential for improving current gene therapy strategies. Clonal tracking is a recent high throughput technology that allows to quantitatively track the evolution and fate of stem cells at clonal level, thus supporting computational modelling of engraftment dynamics and hierarchical relationships in vivo. However, many clonal tracking protocols relies on a subset of cell-types for the characterisation of HSC output and the data generated are subject to measurement errors and noise. This means that mathematical tools not accounting for these aspects may provide misleading conclusions. We propose a stochastic framework for inferring dynamic models of cell differentiation based on clonal tracking data collected in clinical and preclinical gene therapy studies. Our framework is based on stochastic reaction networks combined with extended Kalman filtering and Rauch-Tung-Striebel smoothing. We tested and validated our framework in in-silico studies, showing superiority over the state-of-the-art methods in terms of parameters inference and robustness against measurement noise, false-negative errors, and systematically unobserved cell types characterizing clonal tracking data. The application of our method on five in-vivo gene therapy studies revealed different dynamics of cell differentiation. Our tool can provide statistical support to biologists and clinicians to better understand clonal reconstitution dynamics. The stochastic framework is implemented in the  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/delcoreluca/CellDifferentiationNetworks>.*

## 2.1. 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 hematopoietic stem cells (HSCs) at the root node [2]. 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 hematopoietic tree. As the progeny 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 [3]. Several mathematical models have been proposed to describe hematopoiesis in-vivo. One of the first stochastic models of hematopoiesis 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, and control mechanisms act by varying the cell division and death rates [4].

More recently, various studies analyzed data generated by advanced lineage tracing protocols using novel statistical models [5–10]. Some of these methods are able to take into account missing cell types, such as those that are difficult to collect, e.g. in the bone marrow [11]. Still, to the best of our knowledge, none of the already existing tools considers the bias provided by false-negative clonal tracking errors. State of the art methods usually 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 [12]. Besides, it has also been shown that false-negative errors strongly depend on calling pipeline parameters, as well as read coverage [13].

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 latent states. More precisely, we model cell differentiation using a continuous-discrete state-space formulation including an Ito-type stochastic differential equation (SDE) describing the

clonal dynamics coupled with a measurement model that links the noisy corrupted measurements to the underlying process states. In Section 2.2, a formal definition of our modelling approach is provided along with an expectation-maximization algorithm based on extended Kalman filtering (EKF) and Rauch-Tung-Striebel (RTS) smoothing to infer the unknown parameters. In Section 2.3 we extensively test the method on several simulation studies including a direct comparison with the already existing state-of-the-art approaches and we apply our framework to five in-vivo high dimensional clonal tracking datasets, comparing different biologically plausible models of cell differentiation. In Section 4.5 we discuss our results from both a methodological and biological perspective.

## 2.2. METHODS

A concise graphical representation of our proposed framework is provided in Figure 2.1. It consists in an expectation-maximization (EM) algorithm. The E-step is based on a Kalman filter/smoothen aimed at estimating the state variables given the parameters inferred from the M-step. While in the M-step, a non-linear optimization method updates the unknown parameters given the states estimated by the E-step. Thus, given the cell measurements, both steps are iterated until convergence of the unknown parameters. The following subsections provide details on the state-space formulation of cell differentiation and the expectation-maximization algorithm.

### 2.2.1. A STOCHASTIC MODEL FOR CELL DIFFERENTIATION

Consistently with the definition of a stochastic quasi-reaction network of Section 2.A, we consider a Markov process

$$\mathbf{x}_t = (x_{1t}, \dots, x_{nt}), \quad (2.2.1)$$

for a single clone and  $n$  cell types ( $i = 1, \dots, n$ ) that evolve, in a time interval  $(t, t + \Delta t)$ , according to a set of net-effect vectors  $\{\mathbf{v}_{i_k}\}_{k=1}^{K_i}$  and hazard

functions  $\{h_{i_k}(\mathbf{x}_t, \boldsymbol{\theta})\}_{k=1}^{K_i}$  defined as

$$\mathbf{v}_{i_k} = \begin{cases} (\cdots \underset{i}{1} \cdots)' \\ (\cdots - \underset{i}{1} \cdots)' \\ (\cdots - \underset{i}{1} \cdots \underset{\mathcal{O}(i)}{2} \cdots)' \end{cases} \quad h_{i_k}(\mathbf{x}_t, \boldsymbol{\theta}) = \begin{cases} x_{it} \alpha_i \\ x_{it} \delta_i \\ x_{it} \lambda_{i\mathcal{O}(i)} \end{cases} \quad (2.2.2)$$

where

$$\mathcal{O}(i) = \{j | \lambda_{ij} > 0\} \quad (2.2.3)$$

is the offspring set of cell type  $i$ , and  $K_i$  is the total number of reactions that involve cell type  $i$  and its offspring set  $\mathcal{O}(i)$ . The definitions of the hazard functions and the net-effects follow from the law of mass action, consistently with Eq. (2.A.6) of Section 2.A. The hazard functions include a linear growth term  $x_{it} \alpha_i$  for cell lineage  $i$  with a duplication rate parameter  $\alpha_i > 0$ , a linear term  $x_{it} \delta_i$  for cell death of lineage  $i$  with a death rate parameter  $\delta_i > 0$ , and a linear term  $x_{it} \lambda_{ij}$  describing cell differentiation from lineage  $i$  to any lineage  $j \in \mathcal{O}(i)$  with a differentiation rate  $\lambda_{ij} > 0$ . The vector parameter

$$\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_n, \delta_1, \dots, \delta_n, \boldsymbol{\lambda}'_{1\mathcal{O}(1)}, \dots, \boldsymbol{\lambda}'_{n\mathcal{O}(n)})', \quad (2.2.4)$$

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

$$\mathbf{V} = [\mathbf{v}_{1_1} \cdots \mathbf{v}_{1_{K_1}} \cdots \mathbf{v}_{n_1} \cdots \mathbf{v}_{n_{K_n}}] \in \mathbb{Z}^{n \times K}, \quad (2.2.5)$$

$$\mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}) = (h_{1_1}(\mathbf{x}_t, \boldsymbol{\theta}), \dots, h_{1_{K_1}}(\mathbf{x}_t, \boldsymbol{\theta}) \cdots h_{n_1}(\mathbf{x}_t, \boldsymbol{\theta}), \dots, h_{n_{K_n}}(\mathbf{x}_t, \boldsymbol{\theta}))'$$

where  $K = \sum_{i=1}^n K_i$  is the total number of reactions involved in the network. Finally, as probabilistic assumption, we use the Kolmogorov-forward ODE

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = -\nabla_{\mathbf{x}} \{\boldsymbol{\mu}(\mathbf{x}; \boldsymbol{\theta}) P(\mathbf{x}; t)\} + \frac{1}{2} \nabla_{\mathbf{x}}^2 \{\boldsymbol{\beta}(\mathbf{x}; \boldsymbol{\theta}) P(\mathbf{x}; t)\}, \quad (2.2.6)$$

$$\boldsymbol{\mu}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{V} \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}), \quad \boldsymbol{\beta}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{V} \begin{bmatrix} h_1(\mathbf{x}; \boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_K(\mathbf{x}; \boldsymbol{\theta}) \end{bmatrix} \mathbf{V}',$$

obtained from a continue approximation of the Master equation (see details in Section 2.B).



### 2.2.2. STATE SPACE FORMULATION

We consider a continuous-discrete state space model (CD-SSM) whose dynamic component is the Ito type SDE formulation defined by Eqs. (2.2.1)-(2.2.6), that is

$$\begin{aligned} d\mathbf{x} &= \boldsymbol{\mu}(\mathbf{x}_t; \boldsymbol{\theta}) dt + \boldsymbol{\beta}(\mathbf{x}_t; \boldsymbol{\theta})^{1/2} d\mathbf{W}_t \\ d\mathbf{W}_t &\sim \mathcal{N}_n(\mathbf{0}, dt\mathbf{I}_n), \end{aligned} \quad (2.2.7)$$

combined with the measurement model

$$\begin{aligned} \mathbf{y}_k &= \mathbf{g}_k(\mathbf{x}_{t_k}, \mathbf{R}_k) = \mathbf{G}_k \mathbf{x}_{t_k} + \mathbf{r}_k, \quad \mathbf{r}_k \sim \mathcal{N}_d(\mathbf{0}, \mathbf{R}_k), \\ \mathbf{R}_k &= \rho_0 \mathbf{I}_d + \rho_1 \begin{bmatrix} (\mathbf{G}_k \mathbf{x}_{t_k})_1 & & \\ & \ddots & \\ & & (\mathbf{G}_k \mathbf{x}_{t_k})_d \end{bmatrix}, \end{aligned} \quad (2.2.8)$$

where  $\mathbf{G}_k$  is a  $d \times n$  time-dependent matrix selecting only the measurable states of  $\mathbf{x}_{t_k}$  subject to an additive noise  $\mathbf{r}_k$ , and  $\mathbf{x}_t$  is a shorthand notation for  $\mathbf{x}(t)$ . The covariance matrix  $\mathbf{R}_k$  models the measurement noise as a linear function of the process states  $\mathbf{G}_k \mathbf{x}_{t_k}$ , thus allowing to increase noise intensity with the magnitude of cell counts.

### 2.2.3. OPTIMAL FILTERING AND SMOOTHING

Assuming the Markov properties

$$\begin{aligned} p(\mathbf{x}_k | \mathbf{x}_{1:k-1}, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k | \mathbf{x}_{k-1}) \\ p(\mathbf{x}_{k-1} | \mathbf{x}_{k:T}, \mathbf{y}_{k:T}) &= p(\mathbf{x}_{k-1} | \mathbf{x}_k) \\ p(\mathbf{y}_k | \mathbf{x}_{1:k}, \mathbf{y}_{1:k-1}) &= p(\mathbf{y}_k | \mathbf{x}_k), \end{aligned} \quad (2.2.9)$$

and given the measurements

$$\mathbf{y}_{1:\tau} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_\tau\}, \quad (2.2.10)$$

the aim of optimal filtering and smoothing is to estimate the distributions

$$p(\mathbf{x}_k | \mathbf{y}_{1:\tau}; \boldsymbol{\theta}, \boldsymbol{\rho}) \quad \begin{cases} k > \tau & \text{predictive} \\ k = \tau & \text{filtering} \\ k < \tau & \text{smoothing} \end{cases} \quad (2.2.11)$$

in place of  $p(\mathbf{x}_{0:\tau}|\mathbf{y}_{1:\tau})$ , while inferring  $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\rho})$ . Assuming a prior distribution

$$\mathbf{x}_{t_0} \sim \mathcal{N}_n(\mathbf{x}_{t_0}|\mathbf{m}_0, \mathbf{P}_0), \quad (2.2.12)$$

for  $\mathbf{x}_t$  at  $t = t_0$ , our Kalman filtering and smoothing algorithm can be summarised in the following steps (for details see Section 2.F):

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

$$\begin{cases} \frac{d\mathbf{m}^*(t)}{dt} = \mathbf{V}_\theta \mathbf{m}^*(t) \\ \mathbf{m}^*(t_{k-1}) = \mathbf{m}_{k-1}, \end{cases} \quad (2.2.13a)$$

$$\begin{cases} \frac{d\mathbf{P}^*(t)}{dt} = \mathbf{V}_\theta \mathbf{P}^*(t) + \mathbf{P}^*(t) \mathbf{V}'_\theta + \Delta t \boldsymbol{\beta}(\mathbf{m}^*(t), \boldsymbol{\theta}) \\ \mathbf{P}^*(t_{k-1}) = \mathbf{P}_{k-1}, \end{cases} \quad (2.2.13b)$$

where  $\mathbf{V}_\theta \mathbf{x}$  is a re-formulation of  $\mathbf{V} \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$  as a linear function of  $\mathbf{x}$ .

**2. Update:** Update the initial conditions of the DMEs via

$$\begin{aligned} \boldsymbol{\mu}_k &= \mathbf{G}_k \mathbf{m}_k^* \\ \mathbf{S}_k &= \mathbf{G}_k \mathbf{P}_k^* \mathbf{G}'_k + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^* \mathbf{G}'_k \mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^* + \mathbf{K}_k (\mathbf{y}_k - \boldsymbol{\mu}_k) \\ \mathbf{P}_k &= \mathbf{P}_k^* - \mathbf{K}_k \mathbf{S}_k \mathbf{K}'_k, \end{aligned} \quad (2.2.14)$$

where  $\mathbf{m}_k = \mathbf{m}(t_k)$ ,  $\mathbf{P}_k = \mathbf{P}(t_k)$ ,  $\mathbf{m}_k^* = \mathbf{m}^*(t_k)$ ,  $\mathbf{P}_k^* = \mathbf{P}^*(t_k)$ ,  $\boldsymbol{\mu}_k$  and  $\mathbf{S}_k$  depend on  $\boldsymbol{\psi}$ .

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

$$\boldsymbol{\psi} \leftarrow \underset{\boldsymbol{\psi} \geq \mathbf{0}}{\operatorname{argmin}} - \ell(\boldsymbol{\psi}|\mathbf{y}_1, \dots, \mathbf{y}_k), \quad (2.2.15)$$

$$\mathbf{y}_k \sim \mathcal{N}(\boldsymbol{\mu}_k(\boldsymbol{\psi}), \mathbf{S}_k(\boldsymbol{\psi})), \quad \forall k = 1, \dots, K.$$

**4. Smoothing:** Estimate  $\mathbf{x}_k|\mathbf{y}_{1:K} \sim \mathcal{N}(\mathbf{m}_{k|K}^s, \mathbf{P}_{k|K}^s)$  through the following backward step

$$\begin{cases} \mathbf{B}_{k+1} = \mathbf{P}_k e^{\mathbf{V}'_\theta (\mathbf{P}_{k+1}^*)^{-1}} \\ \mathbf{m}_{k|K}^s = \mathbf{m}_k + \mathbf{B}_{k+1} (\mathbf{m}_{k+1|K}^s - \mathbf{m}_{k+1}^*) \\ \mathbf{P}_{k|K}^s = \mathbf{P}_k + \mathbf{B}_{k+1} (\mathbf{P}_{k+1|K}^s - \mathbf{P}_{k+1}^*) \mathbf{B}'_{k+1}, \end{cases} \quad (2.2.16)$$

where  $e^{(\cdot)}$  is the matrix exponential operator. We used a gradient-based method for the optimization step of Eq. (2.2.15). The gradient

$$\nabla_{\boldsymbol{\psi}} \ell(\boldsymbol{\psi} | \mathbf{y}_1, \dots, \mathbf{y}_k) \quad (2.2.17)$$

of the marginal log-likelihood can be computed with  $p+2$  more prediction and update steps, at every time point  $t_k$ , aimed at computing all the partial derivatives of  $\boldsymbol{\mu}_k(\boldsymbol{\psi})$  and  $\mathbf{S}_k(\boldsymbol{\psi})$  w.r.t.  $\boldsymbol{\psi}$ , where  $p$  is the dimension of  $\boldsymbol{\theta}$  (see Section 2.F for details). The proposed inference procedure is summarised in Algorithm 2 of Section 2.F. The whole procedure returns the estimated parameters  $\hat{\boldsymbol{\psi}}$ , and the first two-order moments  $\mathbf{m}_{k|K}^s$  and  $\mathbf{P}_{k|K}^s$  of the smoothing distribution  $p(\mathbf{x}_k | \mathbf{y}_{1:K}, \hat{\boldsymbol{\psi}})$  at every time point  $t_k$ .

#### 2.2.4. TRANSITION PROBABILITIES

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

$$p_{ij} = \frac{\lambda_{ij} + \alpha_i}{\sum_{k \in \mathcal{O}_{\mathcal{M}}(i)} \lambda_{ik}}, \quad (2.2.18)$$

where  $\mathcal{O}_{\mathcal{M}}(i)$  is the set of all the offspring cells of cell  $i$  in a model  $\mathcal{M}$ , consistently with Eq. (2.2.3).

#### 2.2.5. REACTION CONSTRAINTS

Since in many clonal tracking studies both the HSCs and the progenitors  $P_i$ s are missing states, we assume the following conservation laws

$$\lambda_{HSC \rightarrow P_i} = \sum_j \lambda_{P_i \rightarrow x_j}, \quad (2.2.19)$$

to help parameter inference in Eq. (2.2.15), where  $x_j$ s are all the offspring cell types of  $P_i$ .


#### 2.2.6. MODEL SELECTION

Each candidate model  $\mathcal{M}$  of cell differentiation is scored according to the Akaike Information Criterion (AIC) [14], that is

$$AIC(\mathcal{M}) = 2p_{\mathcal{M}} - 2\ell_{\mathcal{M}}(\boldsymbol{\psi} | \mathbf{y}_1, \dots, \mathbf{y}_k), \quad (2.2.20)$$

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

### 2.2.7. COMPUTATIONAL IMPLEMENTATION

The stochastic framework is implemented in the  package Karen available for download at <https://cran.r-project.org/package=Karen>.

2

## 2.3. RESULTS

### 2.3.1. IN-SILICO VALIDATION STUDIES

We first compared our proposed method Karen with the state-of-the-art approaches, such as the generalised least squares (GLS) method [6], the maximum likelihood method (RestoreNet) [15] and the branchCorr method [11]. The comparisons have been made in terms of robustness against (i) the sampling frequency  $T$ , (ii) the fraction  $f$  of false-negatives, and (iii) the magnitude of the measurement noise parameters  $\rho_0$  and  $\rho_1$ . To this end we used the Euler-Maruyama Algorithm 1 from Section 2.C to simulate the stochastic trajectories of 3 clones obeying to the cell differentiation network of Figure 2.2. To allow the comparison of our method with the other candidate ones we used the definition of net-effect matrix and hazard functions from [11], and the corresponding system of SDEs was defined accordingly. 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 latent states, and therefore GLS and RestoreNet were excluded from this comparison since both methods do not allow for latent states.

Results from Figure 2.3 clearly indicate the superiority of our proposed method over the competitor ones. In particular, Figure 2.3a provides evidence that our proposed method was the most robust against false negative errors compared to the other methods, which provided more biased estimates for the parameters under an high value  $f = 90\%$  of missing data. Subsequently, plot panels 2.3b show that a low sampling frequency ( $T = 4$ ) of the simulated trajectories did not affect the estimates provided by our

proposed method, whereas the ones obtained with any of the competitor approaches were biased. Finally, after increasing the magnitude of the measurement noise parameters  $\rho_0$  and  $\rho_1$  up to 10, our proposed method still provided better estimates compared to the other candidate methods. In conclusion, the results of our synthetic studies show that our method outperformed the competitor ones overall against false negative errors, sample size and measurement noise. This gives confidence in using our method on real in-vivo datasets for providing better parameter estimates and clonal dynamics predictions. Further results under different values of  $f$ ,  $T$ ,  $\rho_0$  and  $\rho_1$  can be found in Sections 2.G-2.H.

We tested our method against model misspecification with an additional synthetic study. We considered two distinct candidate models (Figure 2.4b) that we cross-compared for simulating and fitting. The corresponding system of SDEs was defined according to Eqs. (2.2.1)-(2.2.6). We performed 100 independent simulations for the clonal trajectories using the Euler-Maruyama Algorithm 1 from Section 2.C, and we fitted both candidate models using our proposed framework Karen. As a result, Figure 2.4a indicates that our method was able to identify the true generative model structure, having the lowest median AIC over 100 independent simulations.

### 2.3.2. GENOTOXICITY STUDY

We analyzed an in-vivo clonal tracking dataset previously used to investigate the impact of vector design on clonal diversity in tumor-prone mice [16]. *Cdkn2a*<sup>-/-</sup> tumor prone *Lin*<sup>-</sup> 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, from myeloid cells (CD11b+) and unsorted blood cells (total MNC). Clonal tracking samples were collected under heterogeneous technical conditions (see Table 2.I.1 from Section 2.I), making them not directly comparable. Therefore

we rescaled the data following the description in Section 2.I.

The total number of distinct clones that were collected are 45186 and 20471 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 our stochastic framework Karen to compare four biologically-sustained models of cell differentiation under the two vector conditions PGK and SFV. We reported the results in Figure 2.5 which shows, for each candidate model, the estimated cell differentiation network and the corresponding Akaike Information Criterion (AIC) as defined in Eq. (3.2.12). According to the AIC, model (b) is the one that best fitted clonal tracking data under each vector design. This result suggests that the classical/dichotomic model structure (b) adequately described clonal dynamics in tumor-prone mice under both treatments. Also, the arrow weights from Figure 2.5 clearly indicate that in SFV-treated tumor-prone mice there was a more pronounced unbalance in cell differentiation towards lymphoid progenitors 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 tumor-prone mice, but had an impact on the transition probabilities  $p(HSC \rightarrow P1)$  and  $p(HSC \rightarrow P2)$ , representing HSC differentiation in lymphoid (P1) and myeloid (P2) progenitors.

### 2.3.3. RHESUS MACAQUES STUDY

We analyzed an in-vivo clonal tracking dataset collected from Rhesus Macaques [17]. 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 the original paper study [17]. Although the sample DNA amount was maintained constant during the whole experiment the samples resulted in different magnitudes of reads (see Table 2.J.1 from Section 2.J), making the data not directly comparable. Therefore we rescaled the barcode counts as described in Section 2.J before analysis. The total numbers of clones that were collected range in 1165 - 1291, but we focused on the top 1000 most recaptured ones, so as to further remove

bias.

We fitted the same four candidate models from previous section on the clonal tracking data using Algorithm 2 of Section 2.F. We reported the results in Figures 2.5-2.6 which shows, for each candidate model, the estimated cell differentiation network. According to the AIC from Eq. (3.2.12), model (c) is the one that best fitted the clonal tracking data collected from the rhesus macaque study. 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 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 lymphoid NK cells.

2

#### 2.3.4. GENE THERAPY CLINICAL TRIALS

We considered clonal tracking data collected from six patients affected by three different genetic disorders and that undergo a 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) [18], 2 patients with  $\beta$  hemoglobinopathy, 1 with  $\beta^S/\beta^S$  sickle cell disease [19] and 1 with  $\beta^0/\beta^E$   $\beta$  thalassemia [20]. Details on procedures, gene therapy protocols, and normalization methods can be found in [18–20]. Since data were already normalized to compensate for unbalanced sampling in VCN and DNA [21], we did not apply any further transformation. The total clones that were 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 from previous section have been scored separately in each clinical trial using our stochastic framework Karen. We reported the results in Figures 2.7-2.8 showing the estimated cell differentiation networks for each clinical trial. As a result, according to the AIC, model (d) is the one that best fitted clonal tracking data collected from each clinical trial, thus suggesting that

a three-branches developmental model better explained hematopoietic reconstitution in these clinical trials. In particular, while lymphoid T/B and myeloid G/M developed in parallel through separate branches from different progenitors, NK cells appear to be sustained by a dedicated progenitors cell population.

## 2.4. 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 framework extends stochastic quasi-reaction networks by introducing EKF and RTS components. We have 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 smoothing moments of all the process states, and model selection. Simulation results indicated higher robustness of our proposed method compared to the state-of-the-art ones against (i) a limited number of time points, (ii) limited clonal recapture, and (iii) 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 [22]. Another limitation is that our framework considers reaction rates constant for the whole study period. Extensions that allow for modeling reaction rates as smooth functions of time or depending on clinically relevant variables are within reach and will be the goal of future research.


Our proposed method allowed to unveil the genotoxic impact on cell differentiation in tumor-prone mice. While the differentiation structure it does not seem to be affected by the viral vector design, the transition probabilities from the HSCs to the intermediate progenitors do, showing a more pronounced unbalance towards lymphoid progenitors under the

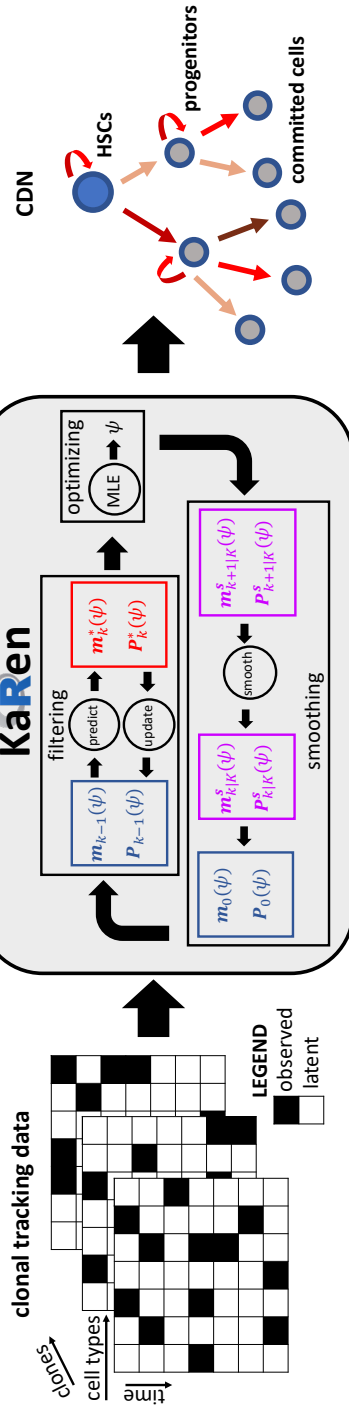


SFV treatment compared to PGK. This can be biologically interpreted as a faster immune response to the higher inflammation caused by the toxic SFV treatment compared to the non-toxic PGK one. Subsequently, the application of Karen to a rhesus macaque clonal tracking study unveiled for the lymphoid NK cells a different developmental pathway from the one detected for lymphoid T and B cells. That is, NK cells are produced by both myeloid and lymphoid progenitors P1 and P2, whereas T and B cells are sustained only by the lymphoid progenitor P1. Results are consistent with the ones previously reported in [17] where the authors demonstrated the presence of distinct subpopulations within the NK lineage, potentially deriving from alternative maturation processes. Finally, we analyzed in-vivo clonal tracking data from three different clinical trials. It is worth noting the degree of agreement between the network structure inferred using the different clinical datasets. Nonetheless, our modelling approach is able to capture the heterogeneity of engraftment dynamics and selective advantage characteristic of the different context as demonstrated by the different parameter estimates.

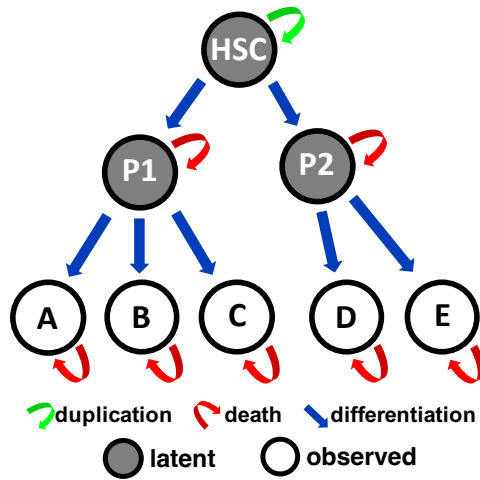
Our stochastic framework can support biologists in understanding hematopoietic reconstitution and in designing tailor-made therapies to treat genetic disorders. Our model can be applied to different types of clonal tracking data, such as vector integration sites, clonal barcodes, and single cell methods. 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.

## 2.5. AVAILABILITY OF DATA AND MATERIALS

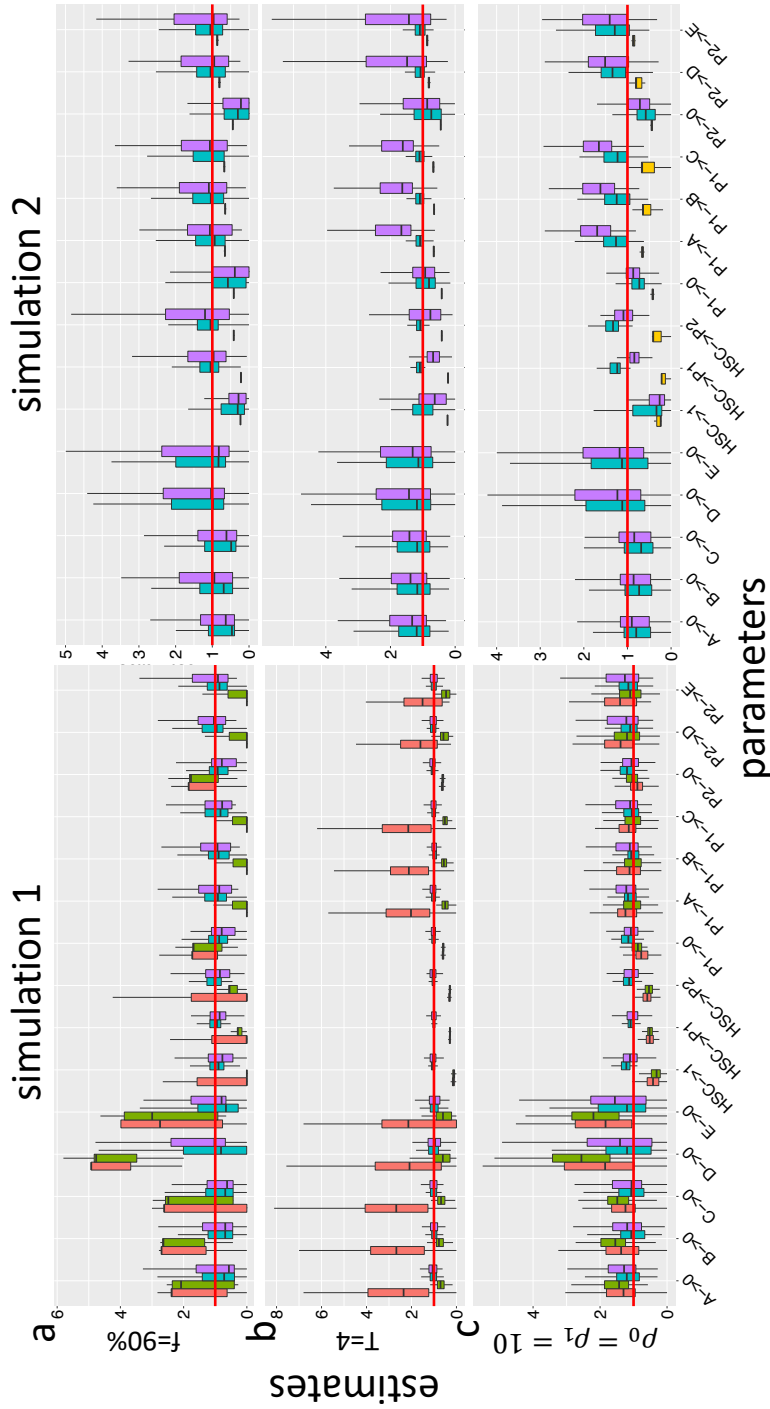
The stochastic framework is implemented in the  package Karen which is available for download at <https://cran.r-project.org/package=Karen>. The code that supports the findings of this study is openly available at <https://github.com/delcore-luca/CellDifferentiationNetworks>.



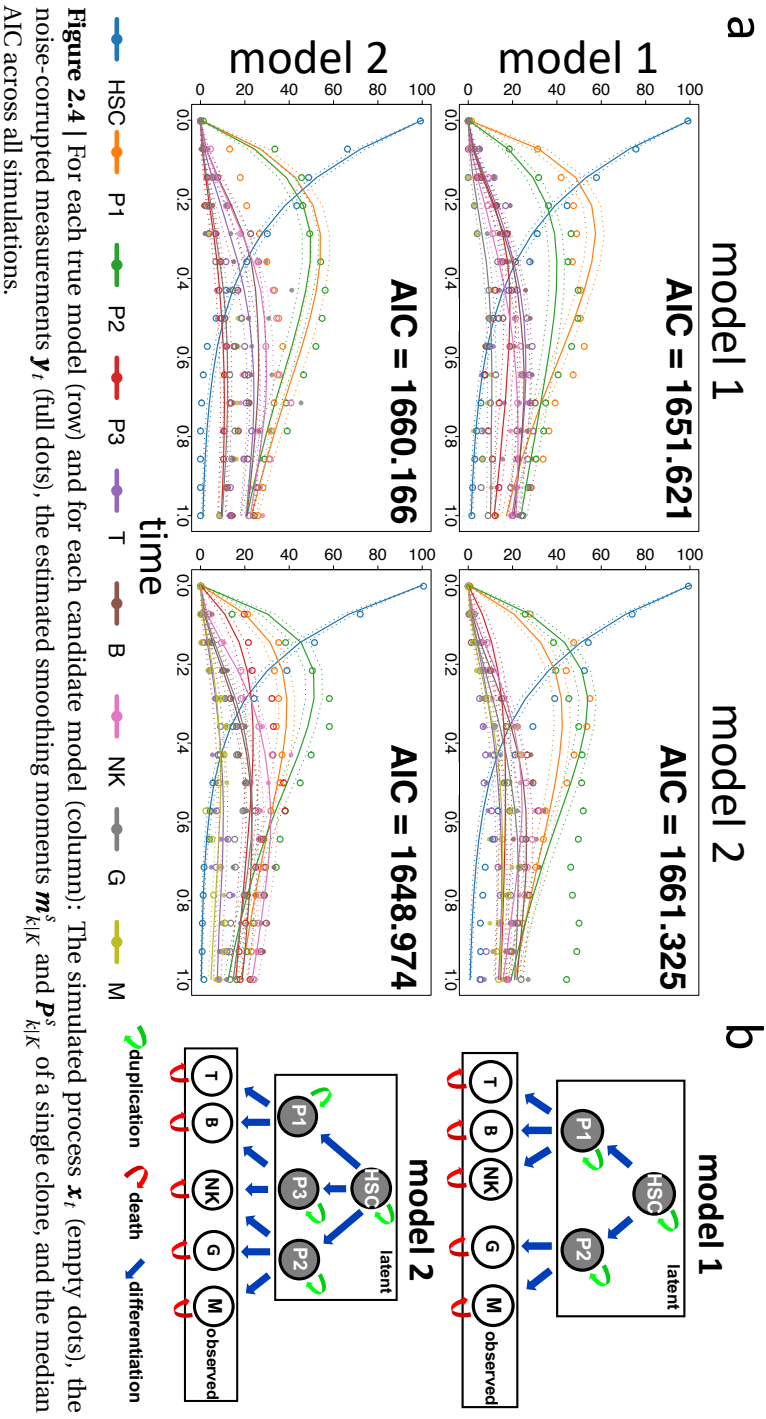
**Figure 2.1** | Schematic representation of the analysis flow: A three-dimensional clonal tracking dataset with partially-observed cells (left panel) is received as input from our proposed stochastic framework Karen (middle panel). It mainly consists in three parts, such as a filtering step, an optimization (maximum likelihood) step, and a smoothing step which are executed iteratively until a convergence is reached on the unknown vector parameter  $\psi$ . Finally, a cell differentiation network (CDN) is returned from Karen, where each arrow is directed and weighted according to the estimated parameters (right panel).

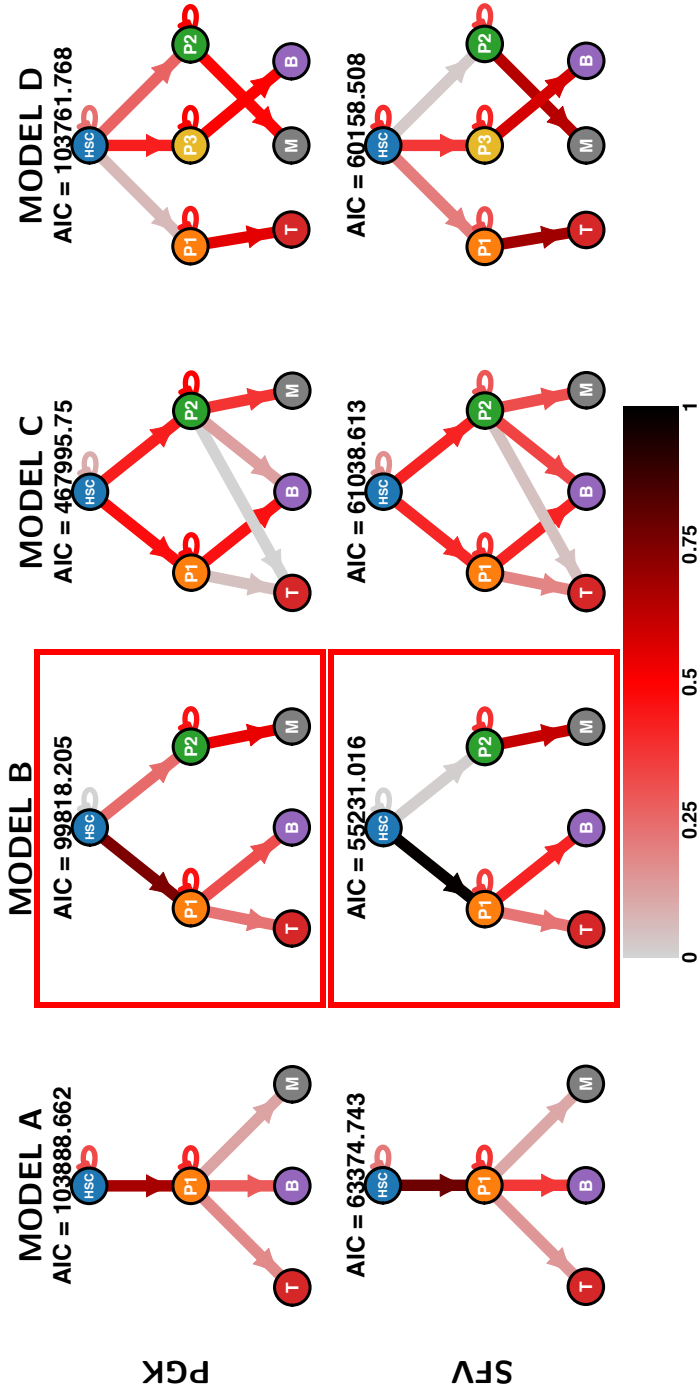


**Figure 2.2** | Graphical representation of the cell differentiation network used in the comparisons in-silico studies. Grey and white nodes represent latent and observed cell types. Arrows represent cell duplication (green), death (red) and differentiation (blue).

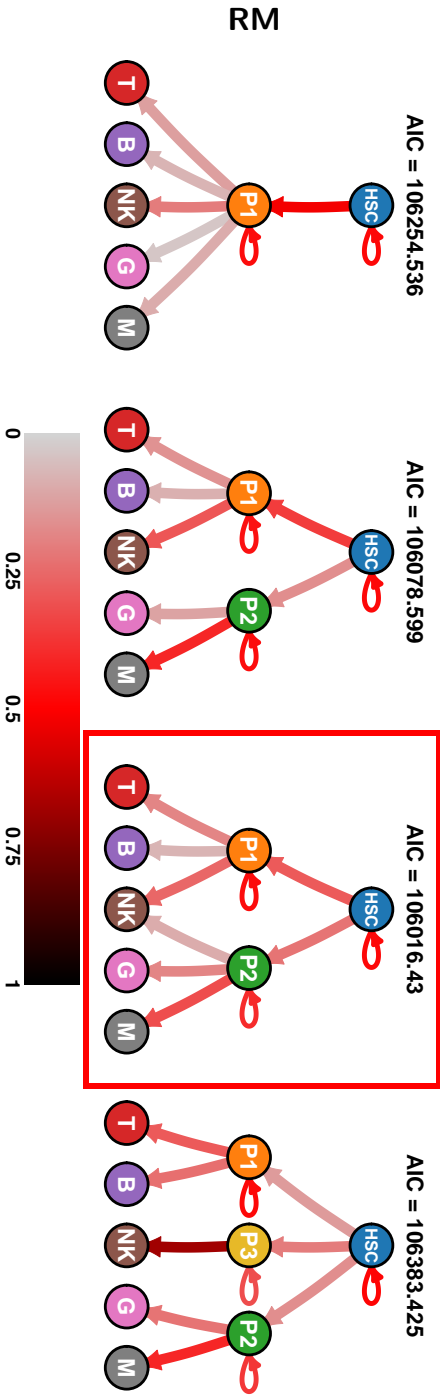


**Figure 2.3** | For each simulation with observed (left) and systematically missing (right) progenitors HSC, P1 and P2: boxplots (y-axis) of the estimated parameters divided by the true ones for each reaction rate (x-axis) obtained from each method (colors) under a fraction of 90% false negative errors (top), a sample size of  $T = 4$  time points (middle), and a measurement noise generated by  $\rho_0 = \rho_1 = 10$  (bottom).

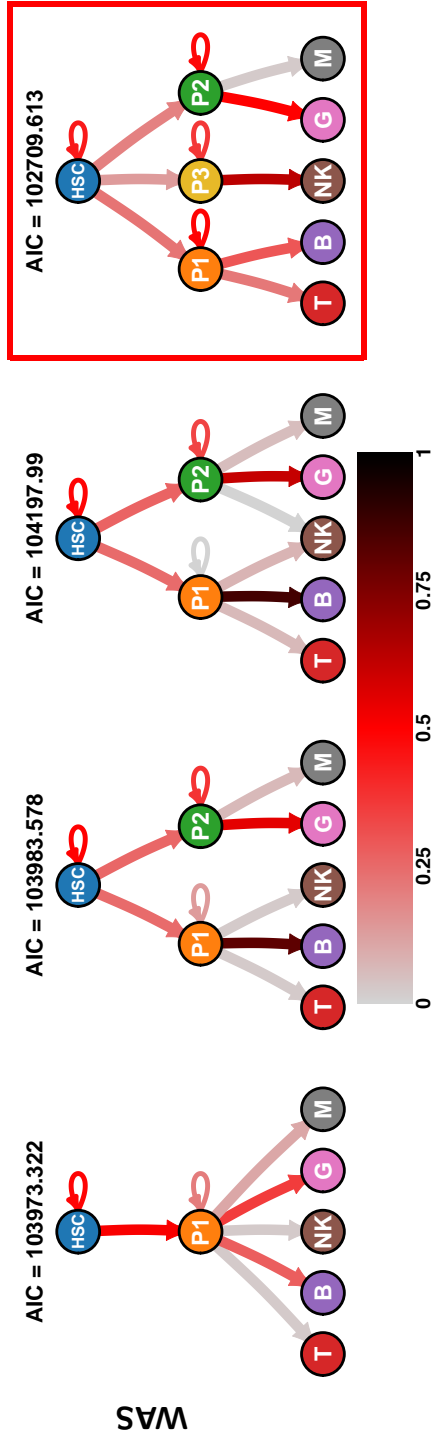




**Figure 2.5** | Results for the mice study: Inferred cell differentiation networks for the candidate models (columns) under the two treatments (rows). Each arrow is weighted and coloured according to the corresponding transition probability estimated with Eq. (2.2.18). For each model the AIC is reported and the best model is squared with a red box.

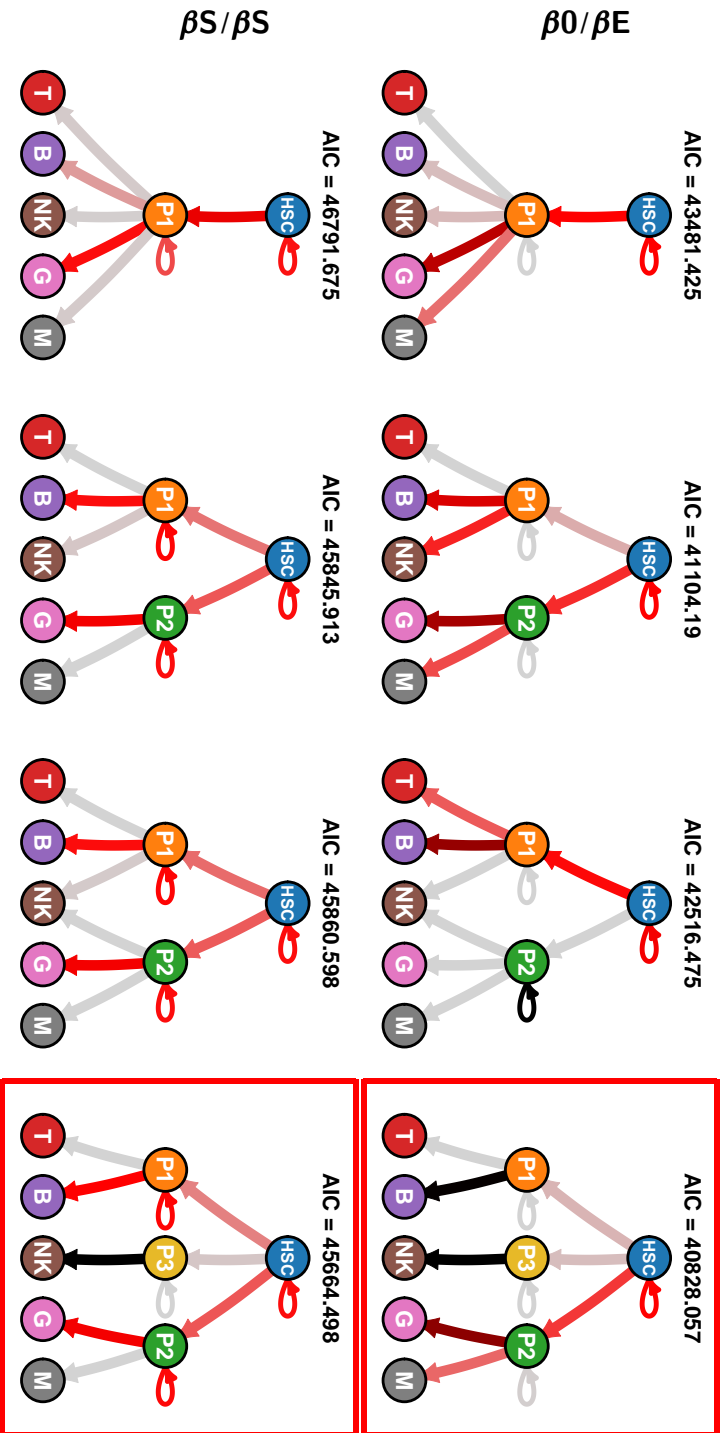


**Figure 2.6** | Results for the rhesus macaque study: Inferred cell differentiation networks for the candidate models (columns). Each arrow is weighted and coloured according to the corresponding transition probability estimated with Eq. (2.2.18). For each model the AIC is reported and the best model is squared with a red box.



**Figure 2.7** | Results for WAS: Inferred cell differentiation networks for the candidate models (columns) in the clinical trial WAS. Each arrow is weighted and coloured according to the corresponding transition probability estimated with Eq. (2.2.18). For each model the AIC is reported and the best model is squared with a red box.





**Figure 2.8** | Results for the clinical trials  $\beta_0/\beta_E$  and  $\beta_S/\beta_S$ : Inferred cell differentiation networks for the candidate models (columns) in each clinical trial (rows). Each arrow is weighted and coloured according to the corresponding transition probability estimated with Eq. (2.2.18). For each model the AIC is reported and the best model is squared with a red box.

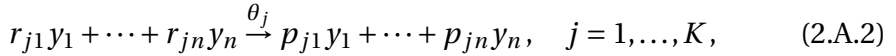
# APPENDIX

## 2.A. STOCHASTIC QUASI-REACTION NETWORKS

Stochastic quasi-reaction networks (S-QRNs) allow to implement a particular class of stochastic differential equations that can be used to model biochemical reactions. More formally, let

$$\mathbf{y}_t = (y_{1t}, \dots, y_{nt})' \in \mathbb{N}_0^n \quad (2.A.1)$$

be a collection of molecules of  $n$  different types observed at time  $t$ , and consider  $K$  distinct (and competing) reactions



each occurring with its own rate  $\theta_j$ . The coefficients  $r_{ji}$ 's defining the left-side of the reaction are called reagents and represent the minimum amount of molecules of type  $i$  needed for the  $j$ -th reaction to occur. Similarly, the coefficients  $p_{ji}$  defining the right-side of the reaction are called products and represent the amount of produced molecules of type  $i$  after the  $j$ -th reaction is triggered. We assume that, if we observe  $\mathbf{y}_0 = (r_{j1}, \dots, r_{jn})'$  molecules at time  $t = 0$ , the  $j$ -th reaction will occur after

$$T_j \sim \text{Exp}(\theta_j), \quad j = 1, \dots, K, \quad (2.A.3)$$

Namely, if exactly  $r_{ij}$  molecules of each type  $i$  would be present, then the  $j$ -th reaction can only take place in one way, with the exponential hazard rate  $\theta_j$ . The interpretation is that, after a waiting time  $T_j$ ,  $r_{ji}$  molecules of type  $i$  collide with each other and produce  $p_{ji}$  molecules of type  $i$  ( $\forall i = 1, \dots, n$ ), while the molecules move randomly in a hosting "cellular" environment. However, in general at time  $t = 0$  we might observe  $Y_{i0} \geq r_{ji}$  molecules of each type  $i$  and, therefore, the  $j$ -th reaction can take place

in a combinatorial number of ways leading to the following waiting time formulation

$$T_j \sim \text{Exp} \left( \theta_j \prod_{i=1}^n \binom{y_{i0}}{r_{ji}} \right), \quad \text{where} \binom{x}{y} = 0, \quad \text{for } x < y, \quad (2.A.4)$$

where

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)' \quad (2.A.5)$$

is the vector parameter for the reaction rates, and

$$h_j(\mathbf{y}_0, \boldsymbol{\theta}) = \theta_j \prod_{i=1}^n \binom{y_{i0}}{r_{ji}} \quad (2.A.6)$$

is the  $j$ -th hazard rate. In this case, the effect will be that at time  $t + T_j$  we have the following expression for the number of molecules of substrate  $i$ ,

$$y_{i,t+T_j} = y_{it} + p_{ji} - r_{ji} = y_{it} + v_{ji}, \quad (2.A.7)$$

where  $v_{ji} = p_{ji} - r_{ji}$  is the  $j$ -th net effect. More compactly, for a set of  $K$  reactions and  $n$  species, the molecular transfer from reagent to product species is a net change of

$$\mathbf{V} = \mathbf{P} - \mathbf{R}, \quad (2.A.8)$$

where  $\mathbf{P} = [p_{ji}]'$  denotes the  $n \times r$  dimensional matrix of products,  $\mathbf{R} = [r_{ji}]'$  is the  $n \times r$  dimensional matrix of reactants, and  $\mathbf{V} = [v_{ji}]'$  is an  $n \times r$  dimensional matrix called net-effect matrix. Therefore, a S-QRN of  $K$ -distinct reactions is fully identified by a net-effect matrix  $\mathbf{V}$  and by the hazard vector

$$\mathbf{h}(\mathbf{y}, \boldsymbol{\theta}) = (h_1(\mathbf{y}, \boldsymbol{\theta}), \dots, h_K(\mathbf{y}, \boldsymbol{\theta}))'. \quad (2.A.9)$$

## 2.B. THE MASTER EQUATION

In practice it is common that the reaction rates of a stochastic reaction network are unknown, and the goal is to estimate them given a collected

dataset. In order to estimate the rates  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)'$  using a likelihood-based approach, we need to define an underlying probabilistic model. One of the most natural choices for describing stochastic chemical kinetics of Eqs. (2.A.2)-(2.A.9) is the chemical master equation

$$\frac{dP(\mathbf{y}; t)}{dt} = \sum_{j=1}^K \{h_j(\mathbf{y} - \mathbf{V}_{\cdot j}; \boldsymbol{\theta})P(\mathbf{y} - \mathbf{V}_{\cdot j}; t) - h_j(\mathbf{y}; \boldsymbol{\theta})P(\mathbf{y}; t)\}, \quad (2.B.1)$$

with transition rates

$$h_j(\mathbf{y}; \boldsymbol{\theta}) = \theta_j \prod_{i=1}^n \binom{y_i}{r_{ji}}, \quad (2.B.2)$$

consistently with Eq. (2.A.4). It describes the temporal evolution of the probability density function  $P(\mathbf{y}; t)$  of the state vector  $\mathbf{y}$  of the chemical system of Eq. (2.A.2). Roughly speaking, the first part of the right-hand side of Eq. (2.B.1) models all the reactions letting the state out of  $k (\neq j)$ , whereas the second part models all the reactions which brings the state back to  $k$ . It is often the case that the Master equation is computationally intractable, especially when the state vector  $\mathbf{y}$  is high-dimensional, so that the number of possible states the system may occupy is too large. Several approximations of the Master equation exist [23, 24], and here we describe a procedure for “continuizing” the discrete-state chemical Markov process defined by Eqs. (2.A.2)-(2.B.1). The procedure is summarized in the following theorem.

**Theorem 2.B.1.** *Assume that  $h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)$  are analytical functions in  $\mathbf{x}$ . Then, a second order Taylor expansion of the products  $h_j(\mathbf{x} - \mathbf{V}_{\cdot j}; \boldsymbol{\theta})P(\mathbf{x} - \mathbf{V}_{\cdot j}; t)$  around  $\mathbf{x}$  leads to the Ito-type stochastic differential equation*

$$d\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_t; \boldsymbol{\theta})dt + \boldsymbol{\beta}^{1/2}(\mathbf{x}_t; \boldsymbol{\theta})d\mathbf{W}(t), \quad d\mathbf{W}(t) \sim N(\mathbf{0}, dt\mathbf{I}), \quad (2.B.3)$$

called the Kramers-Moyal approximation where the drift function and the dispersion matrix are given by

$$\boldsymbol{\mu}(\mathbf{x}_t; \boldsymbol{\theta}) = \mathbf{V}\mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}) \quad (2.B.4)$$

$$\boldsymbol{\beta}(\mathbf{x}_t; \boldsymbol{\theta}) = \mathbf{V} \underbrace{\begin{bmatrix} h_1(\mathbf{x}_t; \boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_K(\mathbf{x}_t; \boldsymbol{\theta}) \end{bmatrix}}_{d(\mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}))} \mathbf{V}'. \quad (2.B.5)$$

*Proof.* The analytical assumption of  $h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{y}; t)$  in  $\mathbf{x}$  allows us to consider a second-order Taylor expansion of  $h_j(\mathbf{x} - V_{.j}; \boldsymbol{\theta})P(\mathbf{x} - V_{.j}; t)$  around  $\mathbf{x}$ , that is

$$\begin{aligned} & h_j(\mathbf{x} - V_{.j}; \boldsymbol{\theta})P(\mathbf{x} - V_{.j}; t) \\ &= h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t) + \nabla_{\mathbf{x}} h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t) \left( (\mathbf{x} - V_{.j}) - \mathbf{x} \right) \\ & \quad + \frac{1}{2} \left( (\mathbf{x} - V_{.j}) - \mathbf{x} \right)' H_{\mathbf{x}} h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t) \left( (\mathbf{x} - V_{.j}) - \mathbf{x} \right) \\ &= h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t) - \nabla_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j} + \frac{1}{2} V'_{.j} H_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j}, \end{aligned}$$

and therefore

$$\begin{aligned} & h_j(\mathbf{x} - V_{.j}; \boldsymbol{\theta})P(\mathbf{x} - V_{.j}; t) - h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t) \\ &= -\nabla_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j} + \frac{1}{2} V'_{.j} H_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j}, \end{aligned}$$

and by plugging it in the Master equation (2.B.1) we have

$$\begin{aligned} \frac{\partial P(\mathbf{x}, t)}{\partial t} &= \sum_{j=1}^K \left\{ -\nabla_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j} + \frac{1}{2} V'_{.j} H_{\mathbf{x}} \{h_j(\mathbf{x}; \boldsymbol{\theta})P(\mathbf{x}; t)\} V_{.j} \right\} \\ &= -\nabla_{\mathbf{x}} \{ \mathbf{V} \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}) P(\mathbf{x}; t) \} + \frac{1}{2} \nabla_{\mathbf{x}}^2 \left\{ \mathbf{V} \begin{bmatrix} h_1(\mathbf{x}; \boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_K(\mathbf{x}; \boldsymbol{\theta}) \end{bmatrix} \mathbf{V}' P(\mathbf{x}; t) \right\}, \end{aligned}$$

which we recognize as a Kolmogorov forward (Fokker-Plank) equation with drift function  $\mathbf{V} \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$  and dispersion matrix  $\mathbf{V} d(\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})) \mathbf{V}'$ , which completes the proof.  $\square$

## 2.C. MONTE CARLO SIMULATION OF ITO-SDES

Stochastic simulation can be stated as forming a Monte Carlo approximation to the probability density  $p(\mathbf{x})$  of the state  $\mathbf{x}$  generated by the Ito-type stochastic differential equation

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t) dt + \mathbf{L}(\mathbf{x}, t) d\boldsymbol{\beta}(t), \quad (2.C.1)$$

where  $\boldsymbol{\beta}$  is a Brownian motion with diffusion matrix  $\mathbf{Q}_c(t)$ . Simulation methods are usually based on discrete approximations to the continuous solution of Eq. (2.C.1).

**EULER-MARUYAMA SIMULATION**

One of the simplest algorithms for simulating trajectories of stochastic differential equations is the Euler-Maruyama method [25] which can be formulated as follows.

**Input:**  $\mathbf{x}_0, \Pi_K = \{0 = t_0 < t_1 < \dots < t_{K+1} = T\}, t_{k+1} - t_k = \Delta t$

**Output:**  $\{\mathbf{x}_k\}_k$

**for**  $k = 1 : K$  **do**

1. Draw  $\Delta\boldsymbol{\beta}_k$  from

$$\Delta\boldsymbol{\beta}_k \sim N(\mathbf{0}, \mathbf{Q}_c(t_k)\Delta t); \quad t_k = k\Delta t \quad (2.C.2)$$

2. Compute

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, t_k)\Delta t + \mathbf{L}(\mathbf{x}_k, t_k)\Delta\boldsymbol{\beta}_k \quad (2.C.3)$$

**end**

**Algorithm 1:** Pseudocode of Euler-Maruyama method.

**2.D. DIFFERENTIAL SYLVESTER EQUATION**

**Theorem 2.D.1.** *Let  $I \subseteq \mathbb{R}$  be an open interval with  $t_0 \in I$  and  $\mathbf{X}(t) \in \mathbb{R}^{n \times m}$ ,  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{m \times m}$ ,  $\mathbf{C}(t) \in \mathbb{R}^{n \times m}$ , and  $\mathbf{D} \in \mathbb{R}^{n \times m}$ . The differential Sylvester equation*

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{A}\mathbf{X}(t) + \mathbf{X}(t)\mathbf{B} + \mathbf{C}(t) \\ \mathbf{X}(t_0) &= \mathbf{D} \end{aligned} \quad (2.D.1)$$

*has the unique solution [26]*

$$\mathbf{X}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{D}e^{\mathbf{B}(t-t_0)} + \int_{t_0}^t e^{\mathbf{A}(t-s)}\mathbf{C}(s)e^{\mathbf{B}(t-s)}ds. \quad (2.D.2)$$

**2.E. INTEGRATING FACTOR METHOD**

A system of first order differential equations in standard form

$$\dot{\mathbf{y}} + \mathbf{A}(t)\mathbf{y} = \mathbf{b}, \quad (2.E.1)$$

where  $\mathbf{y} \in \mathbb{R}^n$ , and  $\mathbf{A}(t) \in \mathbb{R}^{n \times n}$  and  $\mathbf{b} \in \mathbb{R}^n$  has an explicit solution given by

$$\mathbf{y} = e^{-\int \mathbf{A}(t) dt} \left\{ \int e^{\int \mathbf{A}(t) dt} \mathbf{b} dt + \mathbf{C} \right\}, \quad (2.E.2)$$

where

$$I = e^{\int \mathbf{A}(t) dt} \quad (2.E.3)$$

is the integrating factor.

## 2.F. KALMAN REACTION NETWORKS (Karen)

In this work we combine stochastic quasi-reaction networks from Section 2.A with Kalman filtering and smoothing. That is, we consider a non-linear continuous-discrete state space model (CD-SSM) [27] whose dynamic component is represented by the Kramers-Moyal approximation

$$d\mathbf{x} = \mathbf{V} \mathbf{h}(\mathbf{x}_t; \boldsymbol{\theta}) dt + \underbrace{\left( \mathbf{V} \begin{bmatrix} h_1(\mathbf{x}_t; \boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_J(\mathbf{x}_t; \boldsymbol{\theta}) \end{bmatrix} \mathbf{V}' \right)^{1/2}}_{\boldsymbol{\beta}(\mathbf{x}_t; \boldsymbol{\theta})} d\mathbf{W}_t \quad (2.F.1)$$

$$d\mathbf{W}_t \sim \mathcal{N}_n(\mathbf{0}, dt \mathbf{I}_n)$$

of a stochastic quasi-reaction network defined by a  $n \times J$  net effect matrix  $\mathbf{V}$ , a  $p \times 1$  vector parameter  $\boldsymbol{\theta}$  and a  $J \times 1$  hazard vector  $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$  for a  $n$ -dimensional counting process  $\{\mathbf{x}(t) | \mathbf{x}(t) \in \mathbb{N}^n\}_t$ . As measurement model we use

$$\begin{aligned} \mathbf{g}_k(\mathbf{x}(t_k), \mathbf{r}_k) &= \mathbf{G}_k \mathbf{x}(t_k) + \mathbf{r}_k, \quad \mathbf{r}_k \sim \mathcal{N}_d(\mathbf{0}, \mathbf{R}_k), \\ \mathbf{R}_k &= \rho_0 \mathbf{I}_d + \rho_1 \text{diag}(\mathbf{G}_k \mathbf{x}(t_k)), \quad \forall k = 1, \dots, K, \end{aligned} \quad (2.F.2)$$

where  $\mathbf{G}_k \in 0\mathbf{I}^{d \times n}$  (the set of all  $d \times n$  binary matrices) is a time-dependent selection matrix which selects only the measurable particles of  $\mathbf{x}(t_k)$  with an additive noise  $\mathbf{r}_k$  with covariance matrix  $\mathbf{R}_k$ . Here  $\rho_0$  and  $\rho_1$  are free parameters which we infer from the data, and  $\text{diag}(\cdot)$  is a diagonal matrix with diagonal equal to its argument. In the following  $\mathbf{x}_t$  is a shorthand notation for  $\mathbf{x}(t)$ .

Assuming the Markov properties

$$\begin{aligned} p(\mathbf{x}_k | \mathbf{x}_{1:k-1}, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k | \mathbf{x}_{k-1}) \\ p(\mathbf{x}_{k-1} | \mathbf{x}_{k:T}, \mathbf{y}_{k:T}) &= p(\mathbf{x}_{k-1} | \mathbf{x}_k) \\ p(\mathbf{y}_k | \mathbf{x}_{1:k}, \mathbf{y}_{1:k-1}) &= p(\mathbf{y}_k | \mathbf{x}_k), \end{aligned} \quad (2.F3)$$

and given the measurements

$$\mathbf{y}_{1:\tau} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_\tau\}, \quad (2.F4)$$

the aim of optimal filtering and smoothing [28] is to estimate the following distributions

$$p(\mathbf{x}_k | \mathbf{y}_{1:\tau}; \boldsymbol{\theta}, \boldsymbol{\rho}) \quad \begin{cases} k > \tau & \text{predictive} \\ k = \tau & \text{filtering} \\ k < \tau & \text{smoothing} \end{cases} \quad (2.F5)$$

in place of the posterior distribution  $p(\mathbf{x}_{0:\tau} | \mathbf{y}_{1:\tau})$  of the states given the measurements, while inferring the unknown parameters  $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\rho})$ . The Kalman filter/smoothen provides an estimate of the distributions defined in Eq. (2.F5) under a Gaussian assumption of a continuous-discrete state space model [27]. Our proposed Kalman filtering / optimizing / smoothing algorithm can be summarised as follows. Assuming

$$\mathbf{x}(t_0) \sim \mathcal{N}_n(\mathbf{x}(t_0) | \mathbf{m}_0, \mathbf{P}_0) \quad (2.F6)$$

as prior distribution for  $\mathbf{x}(t)$  at  $t = t_0$ , the solution  $\mathbf{x}(t)$  of Eq. (2.F1) is a Gaussian process whose first two-order moments are the solutions of the differential moment equations [27, 29], obtained through the following

**1. Prediction step:**

$$\begin{cases} \frac{d\mathbf{m}_k^*(t)}{dt} = \mathbf{V}_\theta \mathbf{m}_k^*(t) \\ \mathbf{m}_k^*(t_{k-1}) = \mathbf{m}_{k-1} \end{cases} \quad (2.F7a)$$

$$\begin{cases} \frac{d\mathbf{P}_k^*(t)}{dt} = \mathbf{V}_\theta \mathbf{P}_k^*(t) + \mathbf{P}_k^*(t) \mathbf{V}_\theta' + \Delta t \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta}) \\ \mathbf{P}_k^*(t_{k-1}) = \mathbf{P}_{k-1}, \end{cases} \quad (2.F7b)$$



where  $\mathbf{V}_\theta \mathbf{x}_t$  is a linear formulation of  $\mathbf{V}\mathbf{h}(\mathbf{x}_t; \boldsymbol{\theta})$ , and the definition of  $\mathbf{V}_\theta$  depends on  $\mathbf{V}$  and  $\mathbf{h}(\mathbf{x}_t; \boldsymbol{\theta})$ . The solutions of Eq. (2.F.7) are given by

$$\mathbf{m}_k^*(t) = e^{\mathbf{V}_\theta(t-t_{k-1})} \mathbf{m}_{k-1}, \quad (2.F.8a)$$

$$\begin{aligned} \mathbf{P}_k^*(t) &= e^{\mathbf{V}_\theta(t-t_{k-1})} \mathbf{P}_{k-1} e^{\mathbf{V}'_\theta(t-t_{k-1})} \\ &+ \int_{t_{k-1}}^t e^{\mathbf{V}_\theta(t-s)} \Delta t \boldsymbol{\beta}(\mathbf{m}_k^*(s); \boldsymbol{\theta}) e^{\mathbf{V}'_\theta(t-s)} ds. \end{aligned} \quad (2.F.8b)$$

The solution for  $\mathbf{m}_k^*(t)$  is obtained by applying the integrating factor method of Eq. (2.E.2) from Section 2.E to the initial value problem of Eq. (2.F.7a) using an integrating factor

$$I = e^{-\int_{t_{k-1}}^{t_k} \mathbf{V}_\theta ds} = e^{-\mathbf{V}_\theta(t-t_{k-1})}. \quad (2.F.9)$$

The solution for  $\mathbf{P}_k^*(t)$  is obtained by applying the solution formula of Eq. (2.D.2) for a differential Sylvester equation (2.D.1) to the system (2.F.7b). These time-discretized solutions allow to use the update steps of a discrete-time Kalman filter [27], whose equations are given by

## 2. Update step:

$$\begin{aligned} \boldsymbol{\mu}_k &= \mathbf{G}_k \mathbf{m}_k^* \\ \mathbf{S}_k &= \mathbf{G}_k \mathbf{P}_k^* \mathbf{G}'_k + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^* \mathbf{G}'_k \mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^* + \mathbf{K}_k (\mathbf{y}_k - \boldsymbol{\mu}_k) \\ \mathbf{P}_k &= \mathbf{P}_k^* - \mathbf{K}_k \mathbf{S}_k \mathbf{K}'_k, \end{aligned} \quad (2.F.10)$$

where  $\mathbf{m}_k$ ,  $\mathbf{P}_k$ ,  $\mathbf{m}_k^*$ ,  $\mathbf{P}_k^*$ ,  $\boldsymbol{\mu}_k$  and  $\mathbf{S}_k$  depend on the parameters  $\boldsymbol{\theta}$ ,  $\rho_0$  and  $\rho_1$ .

## 3. Optimization step:

For a linear Gaussian continuous-discrete state space model the marginal likelihood of the measurements  $\mathbf{y}_{1:T}$  [30, 31] is the following Gaussian distribution

$$\mathbf{y}_{1:T} \sim \mathcal{N} \left( \begin{bmatrix} \boldsymbol{\mu}_1(\boldsymbol{\psi}) \\ \vdots \\ \boldsymbol{\mu}_T(\boldsymbol{\psi}) \end{bmatrix}, \begin{bmatrix} \mathbf{S}_1(\boldsymbol{\psi}) & & \\ & \ddots & \\ & & \mathbf{S}_T(\boldsymbol{\psi}) \end{bmatrix} \right), \quad (2.F.11)$$

whose optimal parameters can be found via

$$\begin{aligned} \boldsymbol{\psi} &\leftarrow \underset{\boldsymbol{\psi} \geq \mathbf{0}}{\operatorname{argmin}} -\ell(\boldsymbol{\psi}|\mathbf{y}_1, \dots, \mathbf{y}_K), \\ \mathbf{y}_k &\sim \mathcal{N}(\boldsymbol{\mu}_k(\boldsymbol{\psi}), \mathbf{S}_k(\boldsymbol{\psi})), \quad \forall k = 1, \dots, K, \end{aligned} \quad (2.F.12)$$

where

$$\ell(\boldsymbol{\psi}|\mathbf{y}_1, \dots, \mathbf{y}_K) = -\frac{1}{2} \sum_{k=1}^K \log |2\pi \mathbf{S}_k| - \frac{1}{2} \sum_{k=1}^K (\mathbf{y}_k - \boldsymbol{\mu}_k)' \mathbf{S}_k^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k) \quad (2.F.13)$$

is the marginal loglikelihood of the measurements.

**4. Smoothing step:** Following [27], the backward smoothing recursion formula to estimate the first two-order moments  $\mathbf{m}_{k|K}^s$  and  $\mathbf{P}_{k|K}^s$  of the smoothing distribution  $p(\mathbf{x}_k|\mathbf{y}_{1:K}; \boldsymbol{\psi})$  are defined as

$$\begin{cases} \mathbf{B}_{k+1} = \mathbf{P}_k(\boldsymbol{\psi}) e^{V'_{\boldsymbol{\psi}}} (\mathbf{P}_{k+1}^*(\boldsymbol{\psi}))^{-1} \\ \mathbf{m}_{k|K}^s = \mathbf{m}_k(\boldsymbol{\psi}) + \mathbf{B}_{k+1} \left( \mathbf{m}_{k+1|K}^s - \mathbf{m}_{k+1}^*(\boldsymbol{\psi}) \right) \\ \mathbf{P}_{k|K}^s = \mathbf{P}_k(\boldsymbol{\psi}) + \mathbf{B}_{k+1} \left( \mathbf{P}_{k+1|K}^s - \mathbf{P}_{k+1}^*(\boldsymbol{\psi}) \right) \mathbf{B}'_{k+1}, \end{cases} \quad (2.F.14)$$

where  $\boldsymbol{\psi} = (\boldsymbol{\theta}, \rho_0, \rho_1)$  and the values of  $\mathbf{m}_k$ ,  $\mathbf{P}_k$ ,  $\mathbf{m}_k^*$ ,  $\mathbf{P}_k^*$  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  $\nabla_{\boldsymbol{\psi}, \rho_0, \rho_1} -\ell(\boldsymbol{\psi}|\mathbf{y}_1, \dots, \mathbf{y}_K)$  of the marginal negative log-likelihood  $-\ell(\boldsymbol{\psi}|\mathbf{y}_1, \dots, \mathbf{y}_K)$  which is defined by the following partial derivatives

$$\begin{aligned} -\frac{\partial \ell(\boldsymbol{\psi})}{\partial \psi_j} &= \operatorname{tr} \left( \mathbf{S}^{-1} \frac{\partial \mathbf{S}}{\partial \psi_j} \right) - \left( \frac{\partial \boldsymbol{\mu}}{\partial \psi_j} \right)' \mathbf{S}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \\ &-(\mathbf{y} - \boldsymbol{\mu})' \mathbf{S}^{-1} \frac{\partial \mathbf{S}}{\partial \psi_j} \mathbf{S}^{-1} (\mathbf{y} - \boldsymbol{\mu}) - (\mathbf{y} - \boldsymbol{\mu})' \mathbf{S}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \psi_j}, \end{aligned} \quad (2.F.15)$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 & & \\ & \ddots & \\ & & \mathbf{s}_K \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_K \end{bmatrix}. \quad (2.F.16)$$

This requires, at every time point  $k$ ,  $p+2$  more prediction and update steps in order to compute the terms  $\frac{\partial \mathbf{S}_k}{\partial \theta_j}$ 's,  $\frac{\partial \boldsymbol{\mu}_k}{\partial \theta_j}$ 's,  $\frac{\partial \mathbf{S}_k}{\partial \rho_0}$ ,  $\frac{\partial \boldsymbol{\mu}_k}{\partial \rho_0}$ ,  $\frac{\partial \mathbf{S}_k}{\partial \rho_1}$  and  $\frac{\partial \boldsymbol{\mu}_k}{\partial \rho_1}$ , where  $p$  is the dimension of  $\boldsymbol{\theta}$ . These are obtained by differentiating Eqs. (2.F.7) and

(2.F.10) w.r.t.  $\theta$ ,  $\rho_0$  and  $\rho_1$ , as shown below.

**1'. Prediction derivatives step:**

“ $\frac{\partial \mathbf{m}_k^*}{\partial \psi_j}$ ” : If we differentiate the system (2.F.7a) w.r.t.  $\psi_j$  we get

$$\begin{cases} \frac{d}{dt} \left( \frac{\partial}{\partial \psi_j} \mathbf{m}_k^*(t) \right) = \frac{\partial}{\partial \psi_j} \mathbf{V}_\theta \mathbf{m}_k^*(t) = \mathbf{V}_\theta \frac{\partial}{\partial \psi_j} \mathbf{m}_k^*(t) + \left( \frac{\partial}{\partial \psi_j} \mathbf{V}_\theta \right) \mathbf{m}_k^*(t) \\ \frac{\partial}{\partial \psi_j} \mathbf{m}_k^*(t_{k-1}) = \frac{\partial}{\partial \psi_j} \mathbf{m}_{k-1}. \end{cases} \quad (2.F.17)$$

By using the integrating factor

$$I = e^{-\int_{t_{k-1}}^t \mathbf{V}_\theta ds} = e^{-\mathbf{V}_\theta(t-t_{k-1})}, \quad (2.F.18)$$

we get

$$\begin{aligned} \frac{\partial \mathbf{m}_k^*}{\partial \psi_j} &= e^{\mathbf{V}_\theta(t-t_{k-1})} \left\{ \int_{t_{k-1}}^t e^{-\mathbf{V}_\theta(s-t_{k-1})} \frac{\partial \mathbf{V}_\theta}{\partial \psi_j} \mathbf{m}_k^*(s) ds + \frac{\partial}{\partial \psi_j} \mathbf{m}_{k-1} \right\} \\ &= \int_{t_{k-1}}^t e^{\mathbf{V}_\theta(t-s)} \frac{\partial \mathbf{V}_\theta}{\partial \psi_j} e^{\mathbf{V}_\theta(s-t_{k-1})} \mathbf{m}_{k-1} ds + e^{\mathbf{V}_\theta(t-t_{k-1})} \frac{\partial}{\partial \psi_j} \mathbf{m}_{k-1}. \end{aligned} \quad (2.F.19)$$

“ $\frac{\partial \mathbf{P}_k^*}{\partial \psi_j}$ ” : By differentiating the system (2.F.7b) w.r.t.  $\psi_j$  we get

$$\begin{aligned} \frac{d}{dt} \left( \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) \right) &= \frac{\partial}{\partial \psi_j} \left\{ \mathbf{V}_\theta \mathbf{P}_k^*(t) + \mathbf{P}_k^*(t) \mathbf{V}'_\theta + \Delta t \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta}) \right\} \\ &= \frac{\partial}{\partial \psi_j} \mathbf{V}_\theta \mathbf{P}_k^*(t) + \mathbf{V}_\theta \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) + \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) \mathbf{V}'_\theta + \mathbf{P}_k^*(t) \frac{\partial}{\partial \psi_j} \mathbf{V}'_\theta + \\ &\quad + \Delta t \left\{ \sum_{i=1}^n \frac{\partial \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta})}{\partial x_i} \frac{\partial m_{ki}^*(t)}{\partial \psi_j} + \frac{\partial \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta})}{\partial \psi_j} \right\} \\ &= \mathbf{V}_\theta \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) + \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) \mathbf{V}'_\theta + \mathbf{Q}(t), \end{aligned} \quad (2.F.20)$$

where

$$\begin{aligned} \mathbf{Q}(t) &= \frac{\partial}{\partial \psi_j} \mathbf{V}_\theta \mathbf{P}_k^*(t) + \mathbf{P}_k^*(t) \frac{\partial}{\partial \psi_j} \mathbf{V}'_\theta \\ &+ \Delta t \left\{ \sum_{i=1}^n \frac{\partial \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta})}{\partial x_i} \frac{\partial m_{ki}^*(t)}{\partial \psi_j} + \frac{\partial \boldsymbol{\beta}(\mathbf{m}_k^*(t), \boldsymbol{\theta})}{\partial \psi_j} \right\}, \end{aligned} \quad (2.F.21)$$

which is a differential Sylvester equation. The corresponding initial value problem is

$$\begin{cases} \frac{d}{dt} \left( \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) \right) = \mathbf{V}_\theta \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) + \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) \mathbf{V}'_\theta + \mathbf{Q}(t) \\ \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t_{k-1}) = \frac{\partial}{\partial \psi_j} \mathbf{P}_{k-1}, \end{cases} \quad (2.F.22)$$

whose solution is given, by applying Eq. (2.D.2), as

$$\begin{aligned} \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*(t) &= e^{(t-t_{k-1})\mathbf{V}_\theta} \frac{\partial}{\partial \psi_j} \mathbf{P}_{k-1} e^{(t-t_{k-1})\mathbf{V}'_\theta} \\ &+ \int_{t_{k-1}}^t e^{(t-s)\mathbf{V}_\theta} \mathbf{Q}(s) e^{(t-s)\mathbf{V}'_\theta} ds. \end{aligned} \quad (2.F.23)$$

## 2'. Update derivatives step:

The resulting solutions  $\frac{\partial \mathbf{m}_k^*}{\partial \theta_j}$ ,  $\frac{\partial \mathbf{P}_k^*}{\partial \theta_j}$ ,  $\frac{\partial \mathbf{m}_k^*}{\partial \rho_0}$ ,  $\frac{\partial \mathbf{P}_k^*}{\partial \rho_0}$ ,  $\frac{\partial \mathbf{m}_k^*}{\partial \rho_1}$  and  $\frac{\partial \mathbf{P}_k^*}{\partial \rho_1}$  are then used to update the corresponding initial values via a set of equations obtained by differentiating Eq. (2.F.10) w.r.t. each component of  $\boldsymbol{\psi} = (\boldsymbol{\theta}, \rho_0, \rho_1)$ , that is

$$\begin{aligned} \frac{\partial \boldsymbol{\mu}_k}{\partial \psi_j} &= \mathbf{G}_k \frac{\partial \mathbf{m}_k^*}{\partial \psi_j}, \quad \frac{\partial \mathbf{S}_k}{\partial \psi_j} = \mathbf{G}_k \frac{\partial \mathbf{P}_k^*}{\partial \psi_j} \mathbf{G}'_k + \frac{\partial \mathbf{R}_k}{\partial \psi_j}, \\ \frac{\partial \mathbf{K}_k}{\partial \psi_j} &= \frac{\partial \mathbf{P}_k^*}{\partial \psi_j} \mathbf{G}'_k \mathbf{S}_k^{-1} - \mathbf{P}_k^* \mathbf{G}'_k \mathbf{S}_k^{-1} \frac{\partial \mathbf{S}_k}{\partial \psi_j} \mathbf{S}_k^{-1}, \\ \frac{\partial \mathbf{m}_k}{\partial \psi_j} &= \frac{\partial \mathbf{m}_k^*}{\partial \psi_j} + \frac{\partial \mathbf{K}_k}{\partial \psi_j} (\mathbf{y}_k - \boldsymbol{\mu}_k) - \mathbf{K}_k \frac{\partial \boldsymbol{\mu}_k}{\partial \psi_j}, \\ \frac{\partial \mathbf{P}_k}{\partial \psi_j} &= \frac{\partial \mathbf{P}_k^*}{\partial \psi_j} - \frac{\partial \mathbf{K}_k}{\partial \psi_j} \mathbf{S}_k \mathbf{K}'_k - \mathbf{K}_k \frac{\partial \mathbf{S}_k}{\partial \psi_j} \mathbf{K}'_k - \mathbf{K}_k \mathbf{S}_k \frac{\partial \mathbf{K}_k'}{\partial \psi_j}. \end{aligned} \quad (2.F.24)$$

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 marginal log-likelihood  $\ell(\boldsymbol{\psi} | \mathbf{y}_1, \dots, \mathbf{y}_K)$  and its gradient which, in turn, are used for the optimization step. The proposed extended Kalman filter procedure, is summarised in Algorithm 2. All the integrals involved for the computation of  $\mathbf{P}_k^*$ ,  $\frac{\partial}{\partial \theta_j} \mathbf{m}_k^*$ ,  $\frac{\partial}{\partial \theta_j} \mathbf{P}_k^*$ ,  $\frac{\partial}{\partial \rho_0} \mathbf{P}_k^*$  and  $\frac{\partial}{\partial \rho_1} \mathbf{P}_k^*$  are estimated using a 3rd-order Gauss-Legendre method [32].

**Input:**  $\{x_k\}_k, \mathbf{V}, \mathbf{h}(\mathbf{x}, \boldsymbol{\theta}), \mathbf{x}_0 \sim \mathcal{N}_n(\mathbf{m}_0, \mathbf{P}_0)$   
**Output:**  $\hat{\boldsymbol{\theta}}_{ekf}, \hat{\rho}_{0ekf}, \hat{\rho}_{1ekf}, \mathbf{m}_{k|K}^s$  and  $\mathbf{P}_{k|K}^s$

```

while  $\epsilon > \text{tol}$  do
   $\boldsymbol{\psi}_{old} \leftarrow \boldsymbol{\psi}$ 
  for  $k = 1 : K$  do
    1. Prediction: get  $\mathbf{m}_k^*$  and  $\mathbf{P}_k^*$ 
    1'. Prediction derivatives: get  $\frac{\partial \mathbf{m}_k^*}{\partial \psi_j}, \frac{\partial}{\partial \psi_j} \mathbf{P}_k^*$ 
    2. Update: get  $\mathbf{m}_k, \mathbf{P}_k, \boldsymbol{\mu}_k$  and  $\mathbf{S}_k$ 
    2'. Update derivatives: get  $\frac{\partial \mathbf{m}_k}{\partial \psi_j}, \frac{\partial}{\partial \psi_j} \mathbf{P}_k, \frac{\partial \boldsymbol{\mu}_k}{\partial \psi_j}, \frac{\partial}{\partial \psi_j} \mathbf{S}_k$ 
  end
  3. Optimization:  $\boldsymbol{\psi} \leftarrow \underset{\boldsymbol{\psi} > 0}{\text{argmin}} - \ell(\boldsymbol{\psi} | \mathbf{y}_{1:K})$ 
  4. Smoothing: Get  $\mathbf{m}_{k|K}^s$  and  $\mathbf{P}_{k|K}^s$ 
  5. Update prior for  $\mathbf{x}_0$ :  $\mathbf{m}_0 \leftarrow \mathbf{m}_{1|K}^s$  and  $\mathbf{P}_0 \leftarrow \mathbf{P}_{1|K}^s$ 
   $\epsilon \leftarrow \frac{\|\boldsymbol{\psi} - \boldsymbol{\psi}_{old}\|_2}{\|\boldsymbol{\psi}_{old}\|_2}$ 
end

```

**Algorithm 2:** Pseudocode of Karen.

## 2.G. SIMULATION STUDIES

We performed several simulations designed to test and validate our proposed method. The performance has 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 candidate models. We consider the cell differentiation network of Figure 2.K.1-b written in the state space formulation of Eqs. (2.F.1) - (2.F.2) from Section 2.F. The net-effect matrix  $\mathbf{V}$  and an hazard vector  $\mathbf{h}(\mathbf{x}, \boldsymbol{\theta})$  of the dynamic component are defined according to Eq. (3.2.2) from Section 2.2.1, where  $\boldsymbol{\theta}$  is the vector of the unknown dynamic parameters.

We assume that the information on the HSCs and progenitors P1 and P2 is not available at every time point, and therefore we consider them as a latent states which cannot be measured. Furthermore, all the lineages that have not been recaptured for a particular clone at a given time point are

also considered as latent states. Therefore, for the measurement model of Eq. (2.F.2) the selection matrix  $\mathbf{G}_k$  is defined accordingly. In our simulation studies we also assume the following conservation laws

$$\begin{aligned}\lambda_{HSC \rightarrow P1} &= \lambda_{P1 \rightarrow T} + \lambda_{P1 \rightarrow B} + \lambda_{P1 \rightarrow NK} \\ \lambda_{HSC \rightarrow P2} &= \lambda_{P2 \rightarrow G} + \lambda_{P2 \rightarrow M},\end{aligned}\tag{2.G.1}$$

so as to facilitate the inference of parameters related to the systematically missing cell types. In each simulation study, to generate the clone-specific trajectories, we use the Euler-Maruyama algorithm 1 with an initial condition  $\mathbf{x}_0$  of 100 cells for the HSCs and zero otherwise. Each trajectory, starting from  $t_0 = 0$  and terminating at  $t_1 = 1$ , has a sample size equal to 1000 with  $\Delta t = 1/1000$ . Then, we select a subset of  $T$  equidistant time points, where  $T$  is chosen depending on the particular simulation design. Each simulation study is designed to test parameter uncertainty when reducing  $T$  (see Figures 2.G.1-2.G.2), reducing the fraction  $0 < f < 1$  of clones recaptured across lineages and time (see Figures 2.G.3-2.G.4), increasing measurement noise parameters  $(\rho_0, \rho_1)$  (see Figures 2.G.5-2.G.6), selecting a cell differentiation structure among the candidates (Figure 2.4).

Results from simulations show accurate performance of the method in the identification of the missing states and for the inference of the true parameters. In particular, results from simulation 1 suggest that reducing the number of time points of the complete simulated trajectories does not affect parameter inference, even in the extreme case where we fitted our model to the data with only five time points out of 1000 of the complete trajectories. Second simulation clearly indicates that our method still provides good estimates if we reduce the fraction  $f$  of the observed states, even for an high fraction of missing states ( $f = 0.1$ ). Robustness on measurement noise has been assessed in simulation 3, whose results show that parameters are identifiable, even under extreme noise settings ( $\rho_0 = \rho_1 = 100$ ) where still we get some sensible estimates. Finally, in the fourth simulation study our method combined with Akaike Information Criterion was able to select the true generative structure among the candidates.

	branchCorr	GLS	MLE	Karen	Karen-noConstr
sim 1	no	<b>yes</b>	<b>yes</b>	<b>yes</b>	<b>yes</b>
sim 2	<b>yes</b>	no	no	<b>yes</b>	<b>yes</b>

**Table 2.H.1** | Candidate methods (columns) involved (yes/no) in the two simulations studies (rows) with either the progenitor cells considered as measured states (sim 1) or as latent states (sim 2).

2

## 2.H. COMPARISON WITH THE STATE-OF-THE-ART

We compared our proposed method Karen with the state-of-the-art approaches, such as the generalised least squares (GLS) method [6], the maximum likelihood (MLE) [15] and the branchCorr method [11]. The comparisons have been made in terms of robustness against (i) the sample size  $T$ , (ii) the fraction  $f$  of false-negative errors, and (iii) the magnitude of the measurement noise parameters  $\rho_0$  and  $\rho_1$ . To this end we used the Euler-Maruyama algorithm 1 to simulate the stochastic trajectories of 3 clones obeying to the cell differentiation network of Figure 2.2. To allow the comparison of our method with the other candidate ones we used the definition of net-effect matrix and hazard functions from [11]. Two different simulation studies have been designed for the comparisons, depending on whether the competitor approach allows to consider systematically missing cell types or not, as reported in Table 2.H.1.

Results from Figures 2.H.1-2.H.3 clearly indicate the superiority of our proposed method over the competitor ones under different values of  $f$ ,  $T$ ,  $\rho_0$  and  $\rho_1$ . In particular, Figure 2.H.1 provide evidence that our proposed method is the most robust against false negative errors compared to the other methods, which provide more biased estimates for the parameters as we increase the fraction  $f$  of missing data. Subsequently, Figure 2.H.2 show that decreasing the number of time points  $T$  of the simulated trajectories does not affect the estimates provided by our proposed method, whereas the ones obtained with any of the competitor approaches are increasingly biased. Finally, Figure 2.H.3 shows that after increasing the magnitude of the measurement noise parameters  $\rho_0$  and  $\rho_1$  from 0 up to 10, our proposed method still provides better estimates compared to the other candidate methods. In conclusion, the results of our synthetic studies show that our method outperforms the competitor ones overall against

	PGK				LTR			
	DNA	VCN	PS	SD	DNA	VCN	PS	SD
Min.	8.64	1.31	1	60	8.64	0.240	1	189
1st Qu.	106.56	10.90	2	1969	94.50	5.320	1	1130
Median	200.00	13.59	2	5881	200.00	6.300	2	2973
Mean	181.07	12.80	1.96	9351	222.88	6.219	2.1	4695
3rd Qu.	200.25	13.90	2	14055	222.50	7.800	3	7390
Max.	973.00	27.00	3	49853	973.00	10.500	7	15375

**Table 2.I.1 | Mice study:** Quartiles and range of the DNA amount, VCN, PS and SD for the  $n = 242$  samples and separately for PGK (left) and LTR (right) treatments.

false negative errors, sample size and measurement noise. As a result, we would trust more in using our method on real in-vivo datasets for providing better parameter estimates and clonal dynamics predictions.

## 2.I. GENOTOXICITY DATA RESCALING

Clonal tracking samples were collected under heterogeneous technical conditions as reported in Table 2.I.1. The variability of these confounding factors makes clonal tracking samples not directly comparable across time and cell types. Here we consider the DNA amount (in nanograms), the vector copy number (VCN), the pool size (PS) and the PCR protocol (SLiM or Sonic-LAM) as potential confounders. By analogy to the SCS method [16], we first evaluate and then remove the effect of the confounders from the observed data using a regression approach. More precisely, we first perform a log-link Poisson regression on the collected cell counts  $\mathbf{y}$  against the corresponding confounding factors and the possibly factors of interest, leading to the following model

$$\log(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta}, \quad y_i \sim \text{Poisson}(\lambda_i), \quad (2.I.1)$$

where  $y_i$  is the  $i$ -th component of  $\mathbf{y}$ ,  $\lambda_i$  is the  $i$ -th component of  $\boldsymbol{\lambda}$  for  $i = 1, \dots, n$ ,  $\mathbf{X} = [\mathbf{1} \quad \mathbf{X}_c]$  is the full design matrix including a term  $\mathbf{1} \in \mathbb{R}^{n \times 1}$  for the intercept and a term  $\mathbf{X}_c \in \mathbb{R}^{n \times 4}$  with confounder-specific columns. After having estimated the parameters  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_c)'$  with a Fisher scoring algorithm, the rescaled clonal tracking data has been defined as the partial



residuals corresponding to the confounders, that is

$$\mathbf{y}^{res} = \exp\left(\log(\mathbf{y}) - \mathbf{X}_c \hat{\boldsymbol{\beta}}_c\right), \quad (2.I.2)$$

where  $\hat{\boldsymbol{\beta}}_c$  are the optimal parameters for the confounders.

2

## 2.J. RHESUS MACAQUE DATA RESCALING

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 total number of reads. Table 2.J.1 shows the total number of reads collected in each sample of the rhesus macaque clonal tracking dataset. This discrepancy makes all the samples not comparable across time and cell types. Therefore we define the rescaled barcode counts  $Y_{ijk}^{res}$  as

$$Y_{ijk}^{res} = Y_{ijk} \frac{\min_{lm} \sum_n Y_{lmn}}{\sum_n Y_{lmn}}, \quad (2.J.1)$$

where  $Y_{ijk}$  is the  $ijk$ -entry of the barcode matrix with dimensions  $(i, j, k)$  mapping respectively time, cell type and clone.

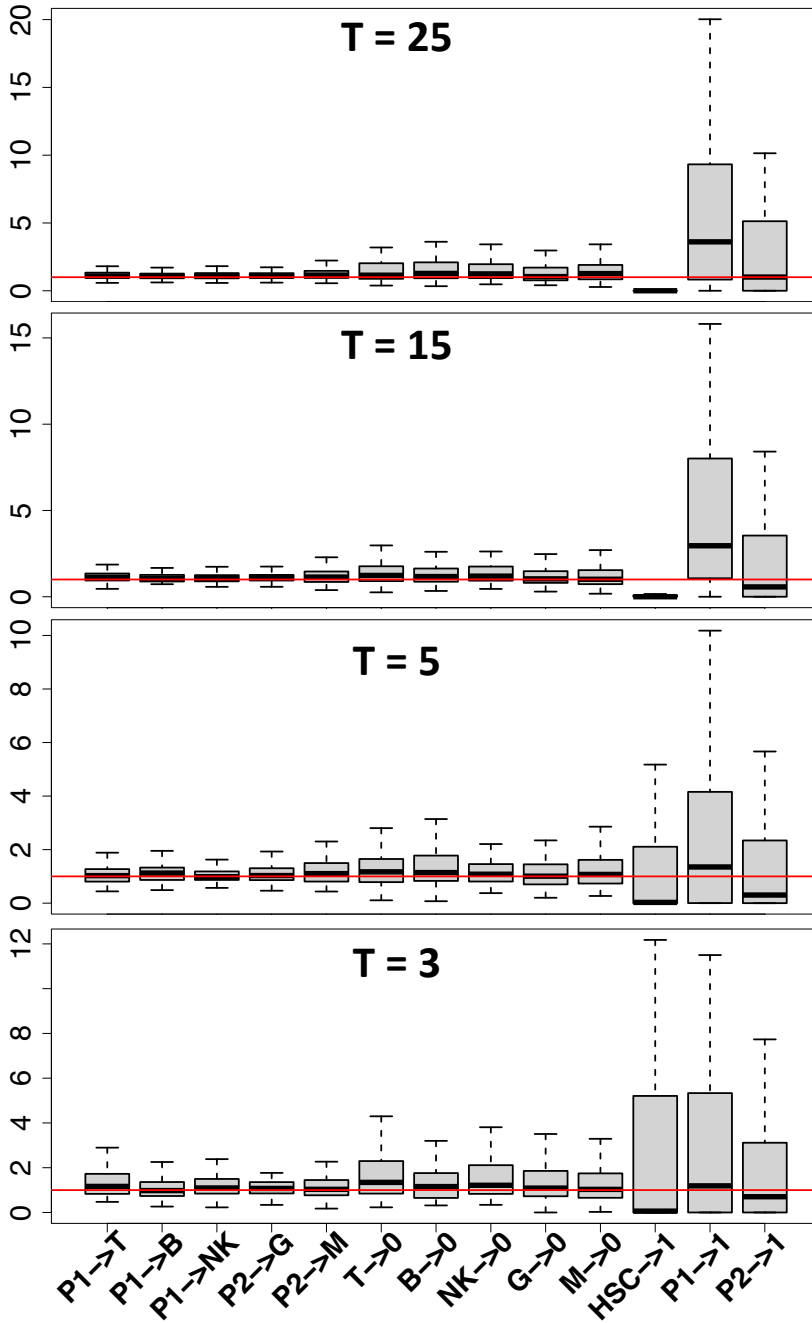
## 2.K. HEMATOPOIETIC MODELS

In this work we consider four different biologically-sustained models of hematopoiesis whose graphical representation is shown in Figure 2.K.1. Model (A) is a single-branch developmental tree where the hematopoietic stem cells produce all the mature cell type through a single multipotent intermediate progenitor  $P_1$ . According to model (B) the lymphoid cells (T, B, NK) and the myeloid cells (G, M) are generated through separate branches of differentiation. Therefore, this is very similar to the well known classical/dichotomic model of hematopoiesis [33]. This model classifies blood cells into two major lineages, but finally differentiated cells are placed in parallel. In contrast, model (C) proposes the idea that myeloid cells represent a prototype of hematopoietic cells capable to produce both myeloid G/M cells and lymphoid NK cells, whereas T, B and NK cells represent specialized types. Therefore model C can be interpreted as the myeloid-based

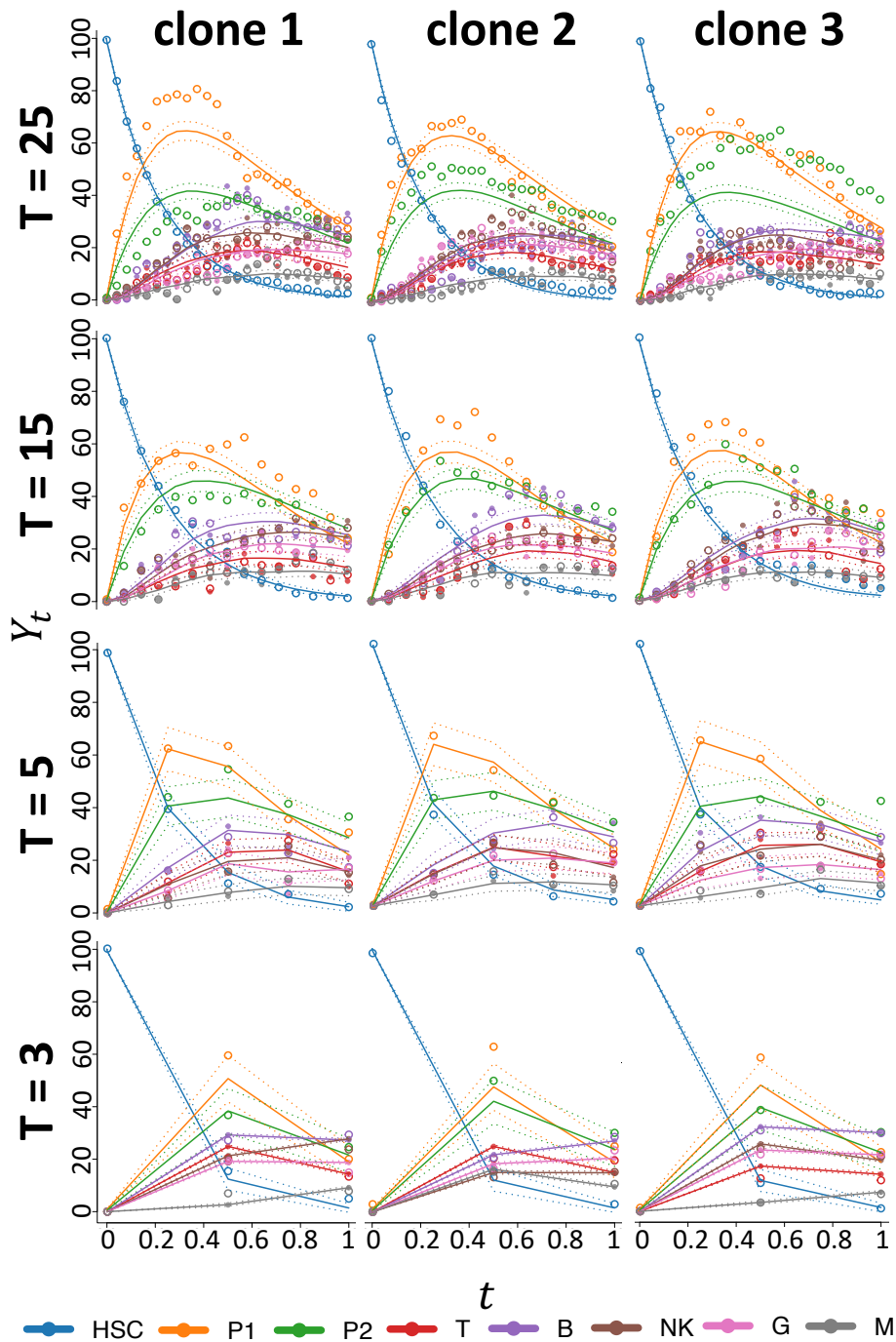
		T	B	NK	M	G
ZH33	1	1465289	74735	135092	119331	2831
	2	225797	216844	335789	1035270	908685
	3	243986	413757	663184	886682	816990
	4.5	485542	479493	834064	985821	987171
	6.5	645005	676413	926089	895309	911637
	9.5	829073	962325	1057398	1229233	1220506
ZH17	1	51802	1347050	1288718	1351450	707382
	2	826190	1342700	1350703	1354355	1213749
	3	1303922	1347692	1338024	1347177	1283250
	4.5	190591	1206361	489098	572877	1195585
	6.5	887851	610999	1344488	381552	1339299
ZG66	1	752127	0	211350	13382	0
	2	692133	58890	308800	363310	145252
	3	339292	209137	424458	808404	704331
	4.5	617281	338977	718472	887183	897672

**Table 2.J.1** | Total number of reads (sum across the clones) collected from each animal (outer rows) at each time point (inner rows) and for all the cell types (columns).

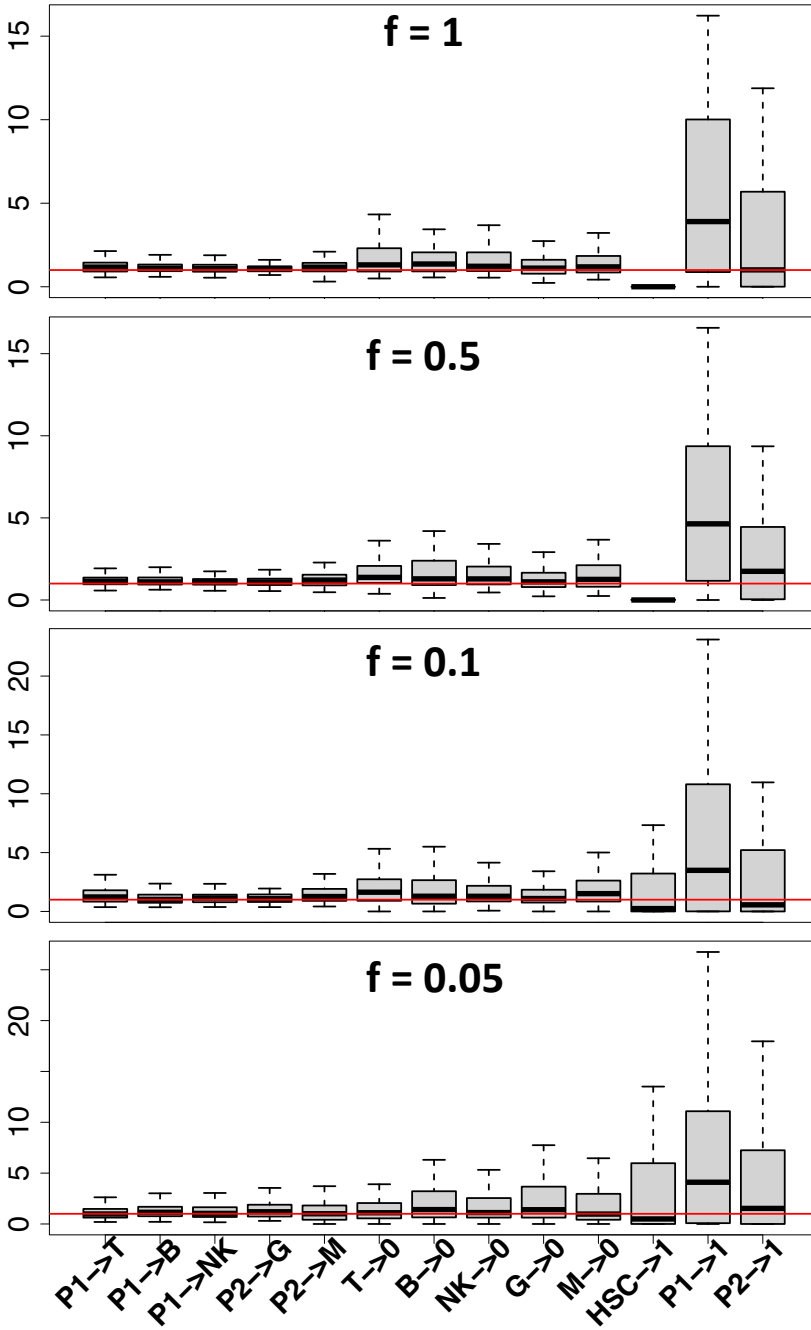
model from [33]. Finally, model (D) assumes that while lymphoid T/B and myeloid G/M develop in parallel through separate branches from different progenitors, there is a third developmental branch for the NK cells which is separated/independent from the first two branches.



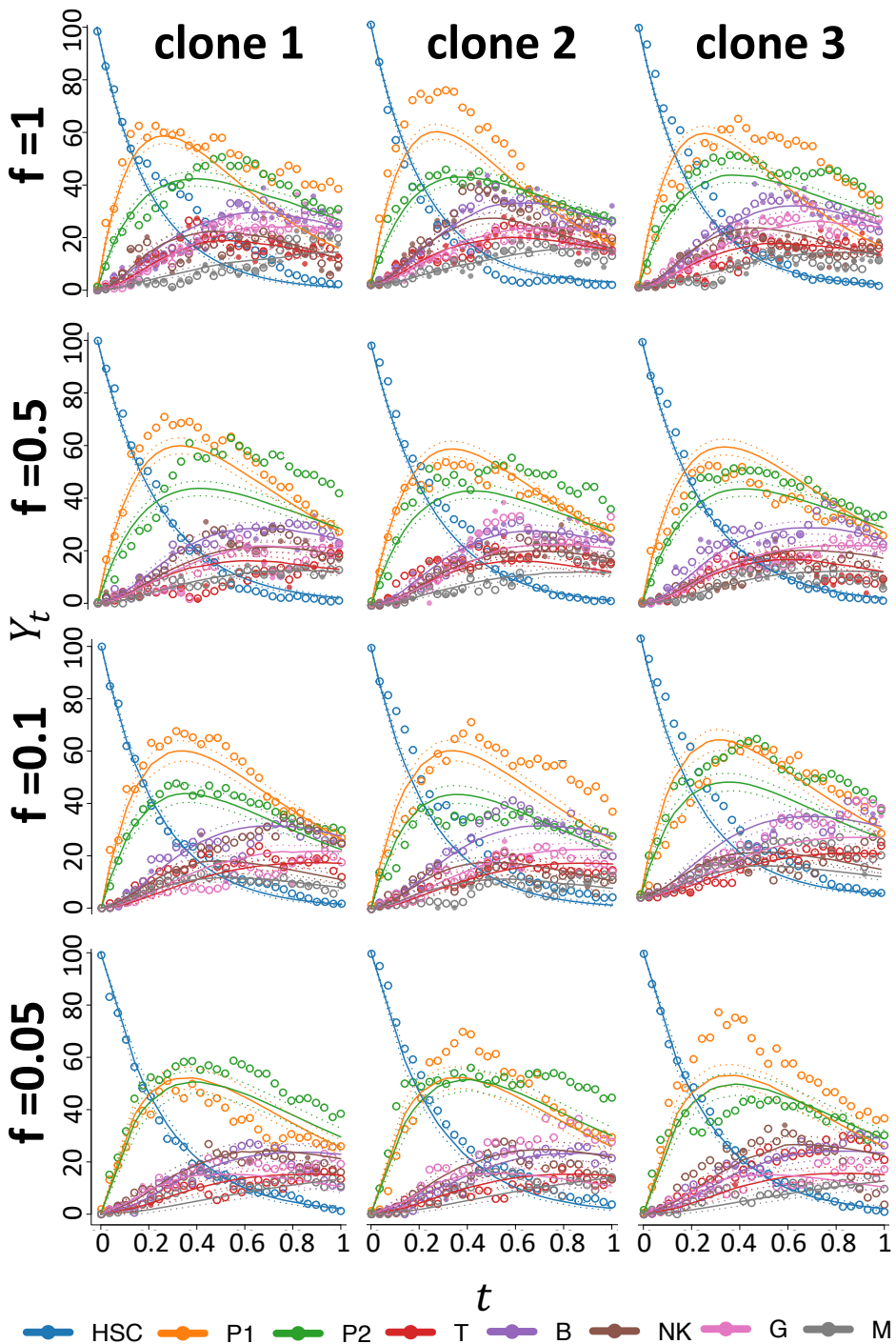
**Figure 2.G.1** | Varying  $T$  (rows): Boxplots of the estimated parameters over 100 independent simulations.



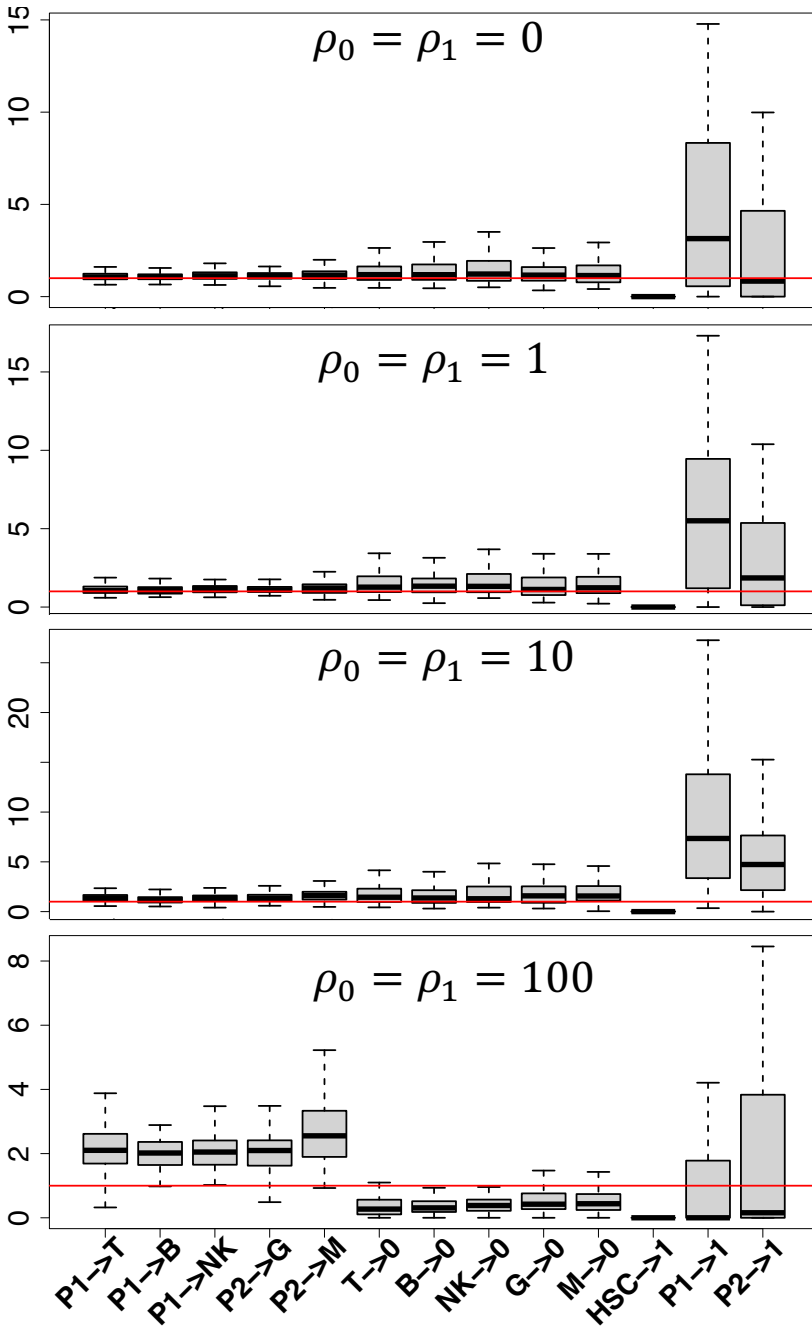
**Figure 2.G.2** | Varying  $T$  (rows): The simulated process  $\{x_t\}_t$  (empty dots), the noise-corrupted measurements  $\{y_k\}_k$  (full dots), and the estimated smoothing moments  $m_{k|K}^s$  and  $P_{k|K}^s$  for each cell type (colors) and clone (columns).



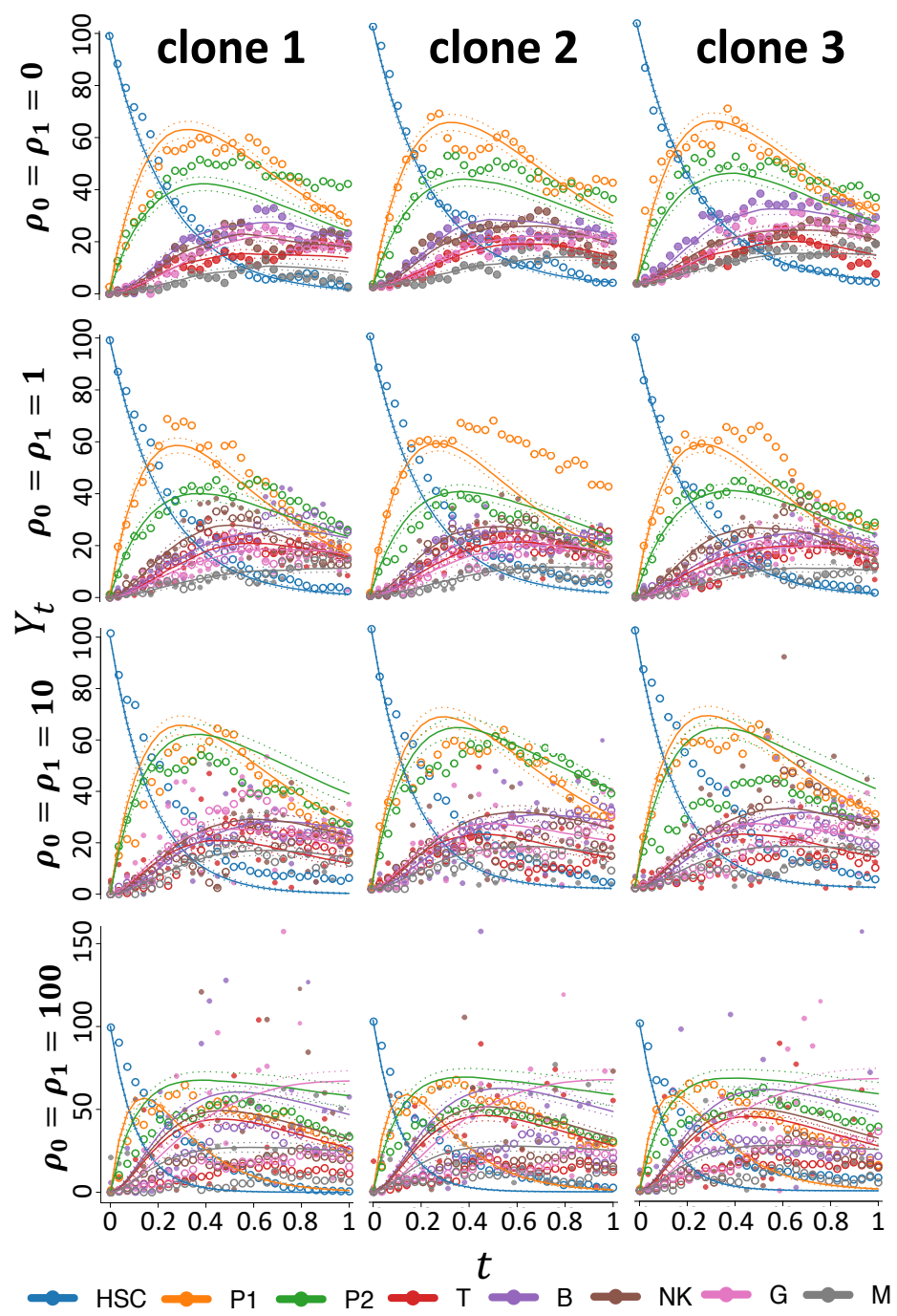
**Figure 2.G.3** | Varying  $f$  (rows): Boxplots of the estimated parameters over 100 independent simulations.



**Figure 2.G.4** | Varying  $f$  (rows): The simulated process  $\{x_t\}_t$  (empty dots), the noise-corrupted measurements  $\{y_k\}_k$  (full dots), and the estimated smoothing moments  $m_{k|K}^s$  and  $P_{k|K}^s$  for each cell type (colors) and clone (columns).

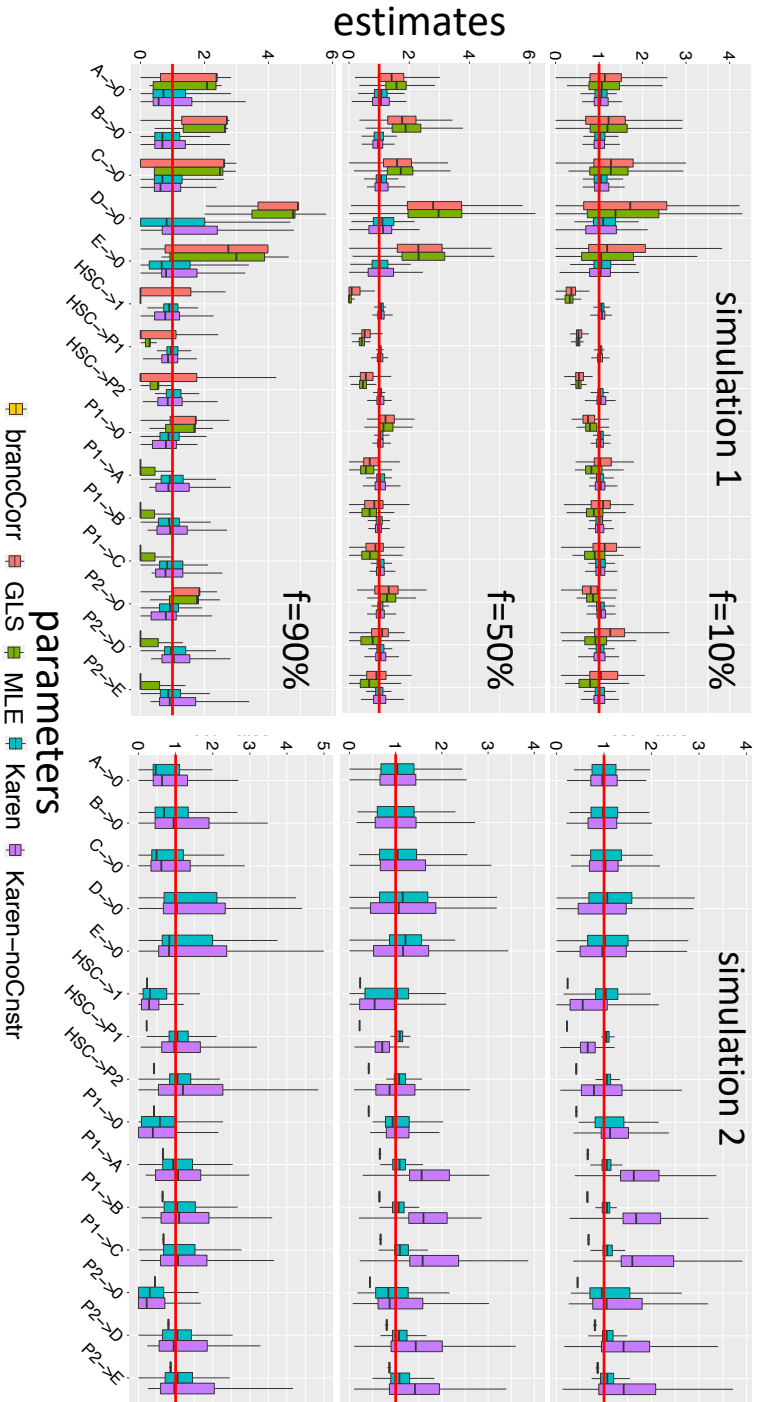


**Figure 2.G.5** | Varying  $\rho_0$  and  $\rho_1$  (rows): Boxplots of the estimated parameters over 100 independent simulations.

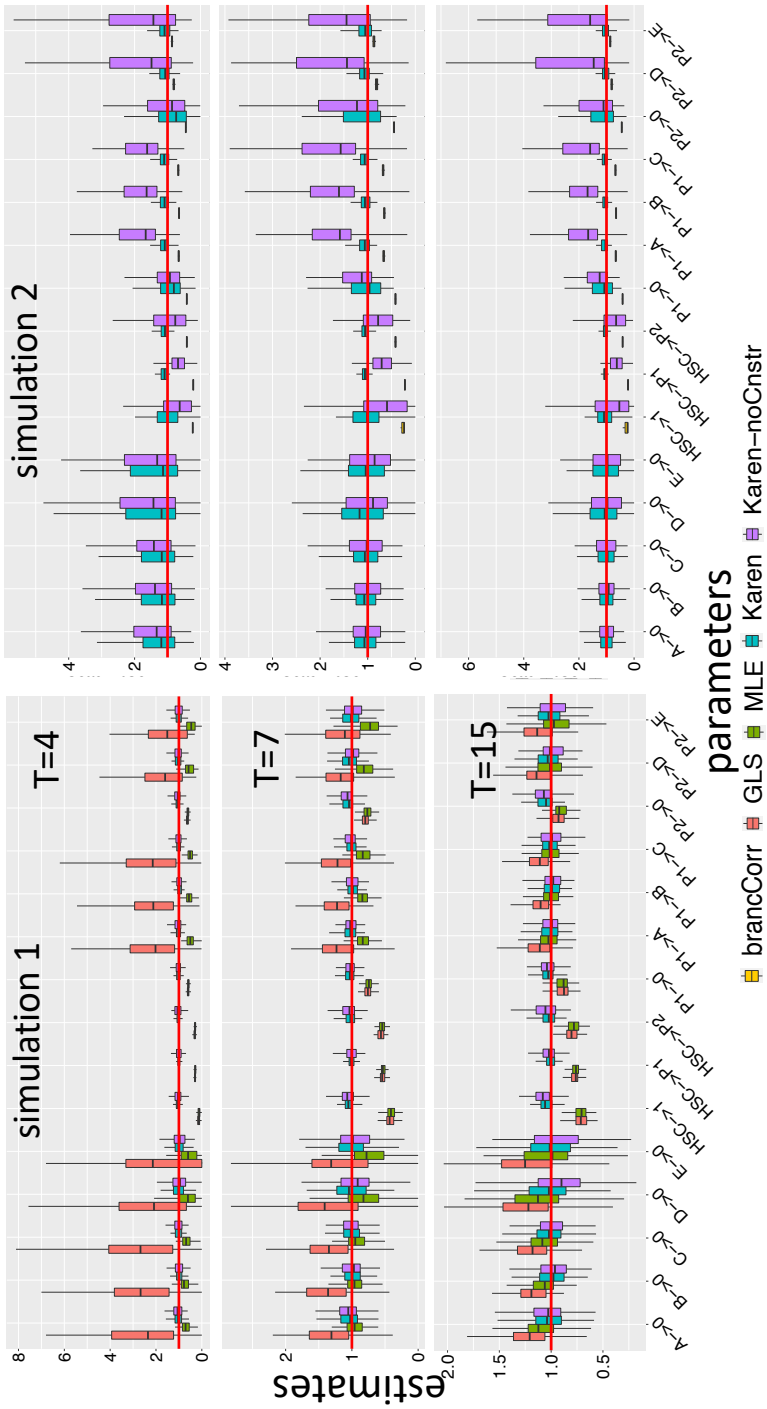


**Figure 2.G.6** | Varying  $\rho_0$  and  $\rho_1$  (rows): The simulated process  $\{x_t\}_t$  (empty dots), the noise-corrupted measurements  $\{y_k\}_k$  (full dots), and the estimated smoothing moments  $m_{k|K}^s$  and  $P_{k|K}^s$  for each cell type (colors) and clone (columns).

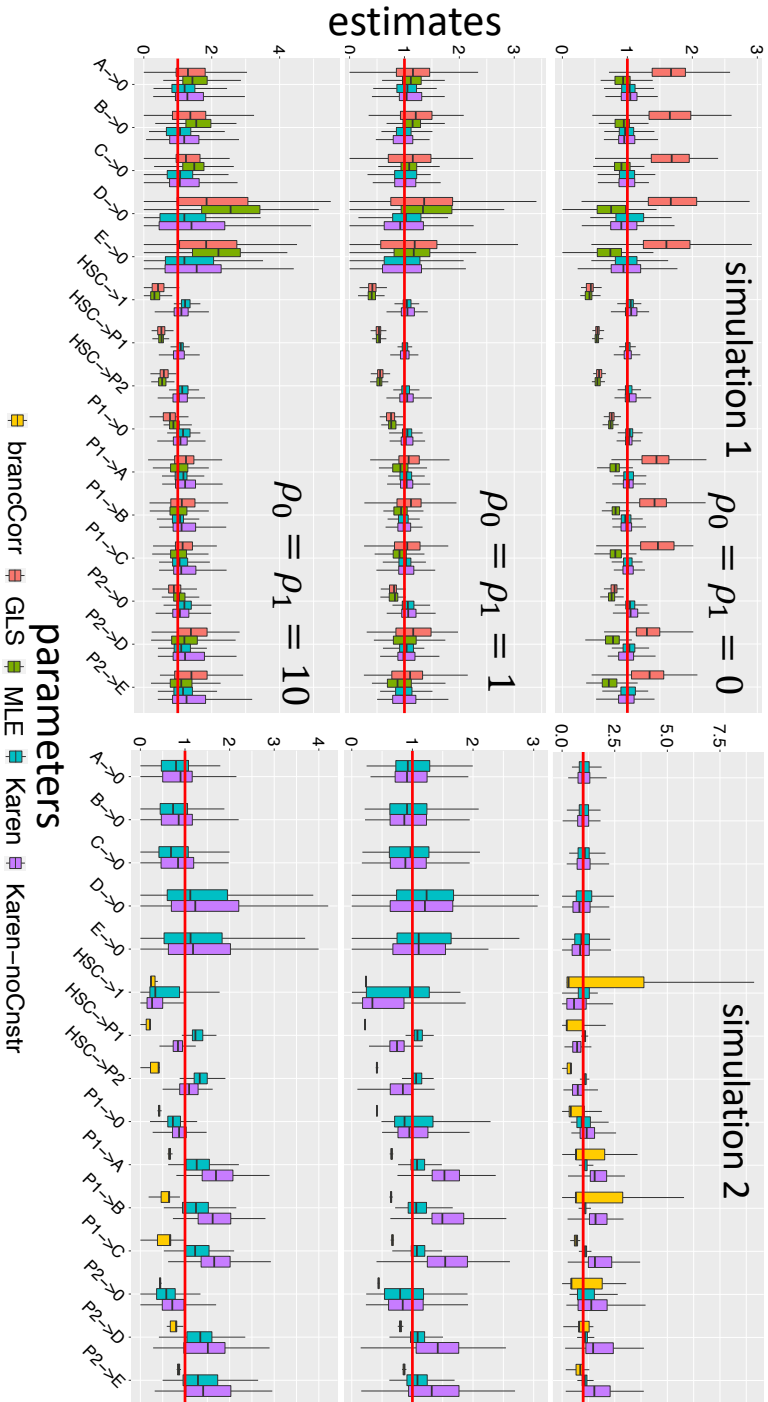




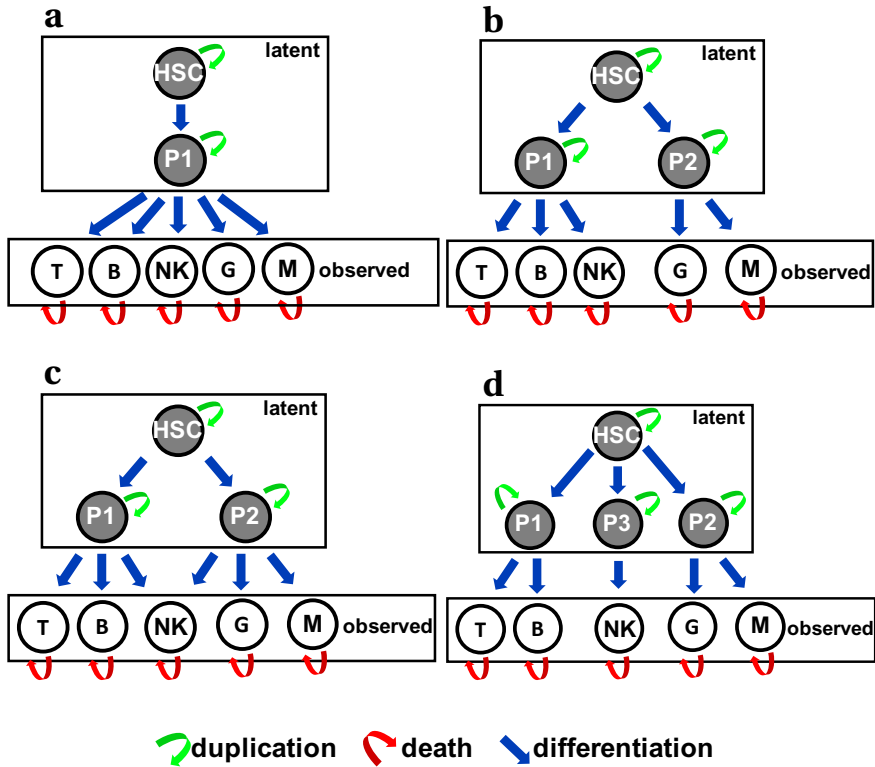
**Figure 2.H.1** | For each simulation with observed (left) and systematically missing (right) progenitors HSC, P1 and P2: boxplots (y-axis) of the estimated parameters divided by the true ones for each reaction rate (x-axis) obtained from each method (colors) under different values for the fraction  $f$  of false negative errors (rows).



**Figure 2.H.2** | For each simulation 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 (colors) under different values for the number of time points  $T$  available (rows).



**Figure 2.H.3** | For each simulation 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 (colors) under different magnitudes of the measurement noise parameters  $\rho_0$  and  $\rho_1$  (rows).



**Figure 2.K.1** | Graphical representation of the candidate models: 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.

## REFERENCES

- [1] L. Del Core, D. Pellin, M. A. Grzegorzczuk, and E. C. Wit, “Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data,” *bioRxiv*, 2022.
- [2] G. M. Cooper, R. E. Hausman, and R. E. Hausman, *The cell: a molecular approach*, vol. 4. ASM press Washington, DC, 2007.
- [3] 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.
- [4] J. E. Till, E. A. McCulloch, and L. Siminovitch, “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.
- [5] C. Di Serio, S. Scala, and P. Vicard, “Bayesian networks for cell differentiation process assessment,” *Stat*, vol. 9, no. 1, p. e287, 2020. e287 STAT-20-0009.R1.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] S. N. Catlin, J. L. Abkowitz, and P. Gutterp, “Statistical inference in a two-compartment model for hematopoiesis,” *Biometrics*, vol. 57, no. 2, pp. 546–553, 2001.
- [11] J. Xu, S. Koelle, P. Gutterp, 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] 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.
- [13] D. Bobo, M. Lipatov, J. Rodriguez-Flores, A. Auton, and B. Henn, "False negatives are a significant feature of next generation sequencing callsets," 2016.
- [14] 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.
- [15] L. Del Core, M. A. Grzegorzczuk, and E. C. Wit, "Stochastic inference of clonal dominance in gene therapy studies," *bioRxiv*, 2022.
- [16] L. Del Core, D. Cesana, P. Gallina, Y. N. S. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. A. Grzegorzczuk, "Normalization of clonal diversity in gene therapy studies using shape constrained splines," *Scientific Reports*, vol. 12, p. 3836, Mar. 2022.
- [17] 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.
- [18] 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.
- [19] J.-A. Ribeil, S. Hacein-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.
- [20] 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  $\beta$ -thalassemia," *New England Journal of Medicine*, vol. 378, no. 16, pp. 1479–1493, 2018.
- [21] 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.

- [22] G. H. W. Gebhardt, A. Kupcsik, and G. Neumann, “The kernel kalman rule,” *Machine Learning*, vol. 108, pp. 2113–2157, Dec 2019.
- [23] P. Sjöberg, P. Lötstedt, and J. Elf, “Fokker–planck approximation of the master equation in molecular biology,” *Computing and Visualization in Science*, vol. 12, no. 1, pp. 37–50, 2009.
- [24] P. Érdi and J. Tóth, *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models*. Manchester University Press, 1989.
- [25] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2011.
- [26] M. Behr, P. Benner, and J. Heiland, “Solution formulas for differential sylvester and lyapunov equations,” *Calcolo*, vol. 56, no. 4, pp. 1–33, 2019.
- [27] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.
- [28] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.
- [29] A. Gelb, J. Kasper, R. Nash, C. Price, and A. Sutherland, “Applied optimal estimation, a. gelb, ed,” 1974.
- [30] S. Särkkä *et al.*, *Recursive Bayesian inference on stochastic differential equations*. Helsinki University of Technology, 2006.
- [31] I. S. Mbalawata, S. Särkkä, and H. Haario, “Parameter estimation in stochastic differential equations with markov chain monte carlo and non-linear kalman filtering,” *Computational Statistics*, vol. 28, no. 3, pp. 1195–1223, 2013.
- [32] P. J. Davis and P. Rabinowitz, *Methods of numerical integration*. Courier Corporation, 2007.
- [33] H. Kawamoto and Y. Katsura, “A new paradigm for hematopoietic cell lineages: revision of the classical concept of the myeloid–lymphoid dichotomy,” *Trends in Immunology*, vol. 30, no. 5, pp. 193–200, 2009.

# 3

## **MIXED-EFFECTS REACTION NETWORKS OF CLONAL DOMINANCE**


---

Parts of this chapter have been published in “Stochastic inference of clonal dominance in gene therapy studies” [1].



## ABSTRACT

3


*Mathematical models of haematopoiesis can provide insights on abnormal cell expansions (clonal dominance), and in turn can guide safety monitoring in gene therapy clinical applications. Clonal tracking is a recent high-throughput technology that can be used to quantify cells arising from a single haematopoietic stem cell ancestor after a gene therapy treatment. Thus, clonal tracking data can be used to calibrate the stochastic differential equations describing clonal population dynamics and hierarchical relationships in vivo. In this work we propose a random-effects stochastic framework that allows to investigate the presence of events of clonal dominance from high-dimensional clonal tracking data. Our framework is based on the combination between stochastic reaction networks and mixed-effects generalized linear models. Starting from the Kramers-Moyal approximated Master equation, the dynamics of cells duplication, death and differentiation at clonal level, can be described by a local linear approximation. The parameters of this formulation, which are inferred using a maximum likelihood approach, are assumed to be shared across the clones and are not sufficient to describe situation in which clones exhibit heterogeneity in their fitness that can lead to clonal dominance. In order to overcome this limitation, we extend the base model by introducing random effects for the clonal parameters. This extended formulation is calibrated to the clonal data using a tailor-made expectation maximization algorithm. We also provide the companion  package RestoreNet, publicly available for download at <https://cran.r-project.org/package=RestoreNet>. Simulations studies show that our proposed method outperforms the state-of-the-art. The application of our method in two in-vivo studies unveils the dynamics of clonal dominance. Our tool can provide statistical support to biologists in gene therapy safety analyses.*

### 3.1. BACKGROUND

In gene therapy the correction of the defective gene(s) underlying the disease is, in principle, sufficient for inducing disease remission or even full recovery [2]. Since the blood system possesses a hierarchical structure with haematopoietic stem cells (HSCs) at its root [3], correction of large numbers of HSCs might be sufficient to eradicate a genetic disease [4, 5]. But genetic modification of large numbers of cells is associated with the higher probability of unintentional vector insertions near proto oncogenes, that may lead to insertional mutagenesis [6–8]. Insertional mutagenesis causes a significant change in clone fitness that can lead to the clones' abnormal expansion and to an unbalanced contribution of different clones to blood cells production. Clonal dominance, characterised by the outgrowth of a small subset of clones (oligoclonality) or one clone (monoclonality) in the most extreme cases, poses serious concerns in the context of gene therapy clinical trials because they might represent the initial stage of a leukemic transformation and are in general considered negative predictors of long term therapeutic benefit.

Clonal dominance in malignant haematopoiesis has been previously identified as a consequence of a clonal competition that is corrupted by disease progression [9, 10]. However, clonal dominance has also been observed in normal haematopoiesis, even in the case of truly neutral clonal markers [11–13]. Indeed, on the basis of various mathematical models, progression of monoclonality has been discussed also for normal (non-leukaemic) stem cell systems [14–18]. While there is strong evidence for clonal selection inducing monoclonal systems in the crypts of the small intestine [19–22], such a process has not been demonstrated for the haematopoietic system yet. There are several high-throughput systems that allow to quantitatively investigate those mechanisms. In gene therapy applications, clonal tracking is performed by using permanent molecular identifier integrated in the host cell genome. In pre-clinical animal studies, these are short fragments of random or semi-random DNA stretches called barcodes, whereas in clinical setting vector integration sites are in general used. After transplantation, all the progeny deriving through cell differentiation inherits the original labels, thus allowing computational modelling to unveil population dynamics and hierarchical relationships

in vivo [23–26].

Here we extend the work by [27, 28] and propose a random-effects cell differentiation network to detect the dynamics of clonal expansion from high dimensional clonal tracking data. In particular, starting from the definition of the master equation [29], a set of Ito-type stochastic differential equations is derived to describe the first two-order moments of the process. We estimate the parameters of the Ito system from its Euler-Maruyama local linear approximation (LLA) [30] using a maximum likelihood approach. Although the base LLA model formulation has been shown to be effective in modelling cell differentiation [28], it has some limitations as it considers all clone trajectories to be iid realizations of the same underlying stochastic process, and does not take into account possible heterogeneous behaviour across the clones. Therefore, the base LLA formulation cannot be used to model clonal dominance. In this work we further increase the flexibility of the base LLA model to take into account for potential heterogeneity in clones' behaviour in both duplication and differentiation rates. To this end we introduce random effects for the clones inside the LLA formulation, providing a mixed-effects LLA model. Then, we use the inferred mixed model to identify which clones are mainly expanding and in which cell compartments. Parameter inference in the mixed-effects formulation is performed by means of an expectation-maximization algorithm, for which we developed an efficient implementation in the  package RestoreNet. Our random-effects LLA formulation describes a stochastic process of clonal dominance on a network of cell lineages. We tested and validated our method in simulation studies, including a direct comparison with the state-of-the-art method GLS [28]. Subsequently, is applied to investigating the dynamics of clonal expansion in a in-vivo model of rhesus macaque haematopoiesis [31]. Finally, by analysing an in-vivo model of tumor prone mice, our method identifies the expected impact of vector genotoxicity on clonal dynamics [32].

### 3

## 3.2. METHODS

An outline of our proposed stochastic framework is as follows. RestoreNet takes a clonal tracking dataset as input, along with a set of reactions coding for cellular duplication, death and differentiation. The system of stochas-

tic differential equations describing the clonal dynamics are translated into a generalized linear model formulation, that possibly includes clone-specific random effects on the dynamics parameters. Subsequently, the parameters are inferred and, if an event of clonal dominance is detected, a pie-chart shows the clones that are expanding and in which cell lineage. A graphical representation of the framework is provided in Figure 3.2.1. This section contains a concise description of the stochastic formulation of clonal dominance and the corresponding inference method. A more detailed description of the stochastic model can be found in Section 3.D.

### 3.2.1. A STOCHASTIC MODEL FOR CELL DIFFERENTIATION

Consistently with the definition of a stochastic quasi-reaction network of Section 2.A, we consider a Markov process

$$\mathbf{x}_t = (x_{1t}, \dots, x_{nt}), \quad (3.2.1)$$

for a single clone and  $n$  cell types ( $i = 1, \dots, n$ ) that evolve, in a time interval  $(t, t + \Delta t)$ , according to a set of net-effect vectors  $\{\mathbf{v}_{i_k}\}_{k=1}^{K_i}$  and hazard functions  $\{h_{i_k}(\mathbf{x}_t, \boldsymbol{\theta})\}_{k=1}^{K_i}$  defined as

$$\mathbf{v}_{i_k} = \begin{cases} (\dots 1 \dots)' \\ (\dots -\frac{1}{i} \dots)' \\ (\dots -\frac{1}{i} \dots \frac{2}{\mathcal{O}(i)} \dots)' \end{cases} \quad h_{i_k}(\mathbf{x}_t, \boldsymbol{\theta}) = \begin{cases} x_{it} \alpha_i \\ x_{it}^2 \delta_i \\ x_{it} \lambda_{i\mathcal{O}(i)} \end{cases} \quad (3.2.2)$$

where

$$\mathcal{O}(i) = \{j | \lambda_{ij} > 0\} \quad (3.2.3)$$

is the offspring set of cell type  $i$ , and  $K_i$  is the total number of reactions that involve cell type  $i$  and its offspring set  $\mathcal{O}(i)$ . The definitions of the hazard functions and the net-effects follow from the law of mass action, consistently with Eq. (2.A.6) of Section 2.A. The hazard functions include a linear growth term  $x_{it} \alpha_i$  for cell lineage  $i$  with a duplication rate parameter  $\alpha_i > 0$ , a quadratic term  $x_{it}^2 \delta_i$  for cell death of lineage  $i$  with a death rate parameter  $\delta_i > 0$ , and a linear term  $x_{it} \lambda_{ij}$  describing cell differentiation from lineage  $i$  to any lineage  $j \in \mathcal{O}(i)$  with a differentiation rate  $\lambda_{ij} > 0$ .

The vector parameter

$$\boldsymbol{\theta} = \left( \alpha_1, \dots, \alpha_n, \delta_1, \dots, \delta_n, \boldsymbol{\lambda}'_{1\mathcal{O}(1)}, \dots, \boldsymbol{\lambda}'_{n\mathcal{O}(n)} \right)', \quad (3.2.4)$$

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

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_{1_1} \cdots \mathbf{v}_{1_{K_1}} \cdots \mathbf{v}_{n_1} \cdots \mathbf{v}_{n_{K_n}}] \in \mathbb{Z}^{n \times K}, \\ \mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}) &= \left( h_{1_1}(\mathbf{x}_t, \boldsymbol{\theta}), \dots, h_{1_{K_1}}(\mathbf{x}_t, \boldsymbol{\theta}) \cdots \cdots h_{n_1}(\mathbf{x}_t, \boldsymbol{\theta}), \dots, h_{n_{K_n}}(\mathbf{x}_t, \boldsymbol{\theta}) \right)', \end{aligned} \quad (3.2.5)$$

where  $K = \sum_{i=1}^n K_i$  is the total number of reactions involved in the network.

### 3.2.2. LLA FORMULATION OF CLONAL DOMINANCE

Let  $\mathbf{y}_t = (y_{1t}, \dots, y_{nt})'$  be the vector of the measurements collected at time  $t$  for a  $n$ -dimensional counting process  $\mathbf{x}_t = (x_{1t}, \dots, x_{nt})'$  obeying to a network of stochastic biochemical reactions defined by a net-effect matrix  $\mathbf{V} \in \mathbb{Z}^{n \times K}$ , a vector parameter  $\boldsymbol{\theta} \in \mathbb{R}^K$  and an hazard vector  $\mathbf{h}(\mathbf{x}, \boldsymbol{\theta}) = (h_1(\mathbf{x}, \boldsymbol{\theta}), \dots, h_K(\mathbf{x}, \boldsymbol{\theta}))'$  and let

$$\underbrace{\begin{bmatrix} \Delta y_{t_0} \\ \vdots \\ \Delta y_{t_{T-1}} \end{bmatrix}}_{\Delta \mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{M}_{t_0} \\ \vdots \\ \mathbf{M}_{t_{T-1}} \end{bmatrix}}_{\mathbf{M}} \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_{nT}(\mathbf{0}, \underbrace{\begin{bmatrix} \Sigma(\boldsymbol{\theta}, \sigma^2) \\ \mathbf{W}_{t_0}(\boldsymbol{\theta}) \\ \ddots \\ \mathbf{W}_{t_{T-1}}(\boldsymbol{\theta}) \end{bmatrix}}_{\mathbf{W}(\boldsymbol{\theta})} + \sigma^2 \mathbf{I}_{nT}), \quad (3.2.6)$$

be the local linear approximation of the Kramers-Moyal approximated Master equation (see Section 2.B for details) where

$$\begin{aligned} \Delta \mathbf{y}_t &= \mathbf{M}_t \boldsymbol{\theta} + (\mathbf{W}_t(\boldsymbol{\theta}) + \sigma^2 \mathbf{I}_n)^{1/2} \Delta \boldsymbol{\varepsilon}_t, \quad \Delta \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n), \\ \mathbf{M}_t &= \mathbf{V} \begin{bmatrix} \prod_{i=1}^n \binom{y_{it}}{r_{1i}} & & \\ & \ddots & \\ & & \prod_{i=1}^n \binom{y_{it}}{r_{Ki}} \end{bmatrix} \Delta t, \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_K)', \\ \mathbf{W}_t(\boldsymbol{\theta}) &= \mathbf{V} \begin{bmatrix} h_1(\mathbf{y}_t, \boldsymbol{\theta}) & & \\ & \ddots & \\ & & h_1(\mathbf{y}_t, \boldsymbol{\theta}) \end{bmatrix} \mathbf{V}' \Delta t, \end{aligned} \quad (3.2.7)$$

with  $\sigma^2$  being the measurement noise variance,  $\mathbf{M}_t\boldsymbol{\theta}$  the mean drift,  $\mathbf{W}_t(\boldsymbol{\theta})$  the diffusion matrix, and  $\Delta\mathbf{y}_t = \mathbf{y}_{t+\Delta t} - \mathbf{y}_t$  is a finite-time increment of  $\mathbf{y}$  in the time interval  $\Delta t$ . From Eq. (3.2.6) it can be seen that all clones share the same vector parameter  $\boldsymbol{\theta}$ . To infer the parameters of Eqs. (3.2.6)-(3.2.7) we developed a maximum likelihood algorithm which is fully described in Section 3.C.

In some cases it may happen that the clones being analysed are drawn from a hierarchy of  $J$  different populations that possibly behave differently in terms of dynamics. In this case it might be of interest to quantify the population-average  $\boldsymbol{\theta}$  and the clonal-specific effects  $\mathbf{u}$  around the average  $\boldsymbol{\theta}$  for the description of clone-specific dynamics. For achieving this goal, we extend the LLA formulation of Eq. (3.B.3) with a mixed-effects model [33] by introducing random effects  $\mathbf{u}$  for the  $J$  distinct clones on the vector parameter  $\boldsymbol{\theta}$ , leading to a random-effects stochastic reaction network (RestoreNet). The extended random-effects formulation becomes

$$\Delta\mathbf{y} = \underbrace{\begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ & \ddots \\ \mathbf{0} & \mathbf{M}_J \end{bmatrix}}_{\mathbf{M} \in \mathbb{R}^{nT \times Jp}} \mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathbf{u} \sim \mathcal{N}_{Jp} \left( \underbrace{\mathbf{1}_J \otimes \boldsymbol{\theta}}_{\boldsymbol{\theta}_u}, I_J \otimes \underbrace{\begin{bmatrix} \tau_1^2 & \mathbf{0} \\ & \ddots \\ \mathbf{0} & \tau_p^2 \end{bmatrix}}_{\Delta_u} \right), \quad (3.2.8)$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_{nT}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2)),$$

where  $\mathbf{M}$  is the block-diagonal design matrix for the random effects  $\mathbf{u}$  centered in  $\boldsymbol{\theta}$ , each block  $\mathbf{M}_j$  is clone-specific, and  $\otimes$  is the Kronecker product. As in the case of the null model of Eq. (3.B.3), we estimate  $\sigma^2$  based on data. From Section 3.D, the conditional distribution of the random effects  $\mathbf{u}$  given the data  $\Delta\mathbf{y}$  is

$$\mathbf{u} | \Delta\mathbf{y} \sim \mathcal{N}_{Jp}(E_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}[\mathbf{u}], V_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}(\mathbf{u})), \quad (3.2.9)$$

where

$$\begin{aligned} E_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}[\mathbf{u}] &= V_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}(\mathbf{u}) (\mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\Delta\mathbf{y} + \Delta_u^{-1}\boldsymbol{\theta}_u), \\ V_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}(\mathbf{u}) &= (\mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\mathbf{M} + \Delta_u^{-1})^{-1}, \end{aligned} \quad (3.2.10)$$

and  $\boldsymbol{\psi} = (\boldsymbol{\theta}', \sigma^2, \tau_1^2, \dots, \tau_p^2)'$  is the vector of all the unknown parameters. Once the parameters are estimated (see next section for inference details),

the conditional expectations  $E_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}}[\mathbf{u}]$  can then be used as a proxy for the clone-specific dynamic parameters. This method allows to infer clone-specific dynamics by extremely reducing the problem dimensionality from  $J \cdot p$  to  $2 \cdot p + 1$  ( $J \gg 2$ ).

### 3.2.3. INFERENCE PROCEDURE


3

In order to infer the maximum likelihood estimator  $\hat{\boldsymbol{\psi}}$  for

$$\boldsymbol{\psi} = (\boldsymbol{\theta}, \sigma^2, \tau_1^2, \dots, \tau_p^2), \quad (3.2.11)$$

we have developed an efficient expectation-maximization (E-M) algorithm where the collected cell increments  $\Delta\mathbf{y}$  and the random effects  $\mathbf{u}$  take the roles of the observed and latent states respectively. The full analytical expression of the E-step function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*) = E_{\mathbf{u}|\Delta\mathbf{y};\boldsymbol{\psi}^*}[\ell(\Delta\mathbf{y}, \mathbf{u}; \boldsymbol{\psi})]$  and its partial derivatives  $\frac{\partial}{\partial \psi_j} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  are available (see Section 3.D). In the E-M algorithm we iteratively update the E-function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  using the current estimate  $\boldsymbol{\psi}^*$  of  $\boldsymbol{\psi}$  and then we minimize the  $-Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  w.r.t.  $\boldsymbol{\psi}$ . As the E-step function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  is non-linear and the parameters are box-constrained, we used the L-BFGS-B algorithm from the `optim()` base R function for optimization, to which we provided the objective function, along with its gradient  $\nabla_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$ , as input. The E-M algorithm is iterated until a convergence criterion is met, that is when the relative errors of the E-step function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  and the parameters  $\boldsymbol{\psi}^*$  are lower than a predefined tolerance.

Once we get the E-M estimate  $\hat{\boldsymbol{\psi}}$  for the parameters we evaluate the goodness-of-fit of the mixed-model according to the conditional Akaike Information Criterion [34]. As every E-M algorithm, the choice of the starting point  $\boldsymbol{\psi}_s$  is very important from a computational point of view. We chose  $\boldsymbol{\psi}_s = (\boldsymbol{\theta}_s, \sigma_s^2, \tau_1^2 = 0, \dots, \tau_p^2 = 0)$  as a starting point where  $(\boldsymbol{\theta}_s, \sigma_s^2)$  is the optimum found in the fixed-effects LLA formulation of Eq. (3.B.3). This is a reasonable choice since we want to quantify how the dynamics  $E_{\mathbf{u}|\Delta\mathbf{y};\hat{\boldsymbol{\psi}}}[\mathbf{u}]_j$  of each clone  $j$  departs from the average dynamics  $\boldsymbol{\theta}_s$ . With the help of simulation studies (see Results section), we empirically proved that this choice always led to a conditional expectation  $E_{\mathbf{u}|\Delta\mathbf{y};\hat{\boldsymbol{\psi}}}[\mathbf{u}]$  consistent with the true clone-specific dynamic parameters  $\boldsymbol{\theta}$ . Computational details can be found in Section 3.D. The pseudocode of the E-M algorithm is provided

in Algorithm 5 of Section 3.D. The maximum likelihood inference for the basal model and the expectation maximization algorithm for the random-effects model are implemented in the  package RestoreNet, available for download at <https://cran.r-project.org/package=RestoreNet>.

### 3.2.4. MODEL SELECTION

The fixed-effects model  $\mathcal{M}_0$  is scored according to the corrected Akaike Information Criterion (AIC) [35] defined as

$$AIC(\mathcal{M}_0) = -2\ell_{\mathcal{M}_0}(\boldsymbol{\theta}, \sigma^2 | \Delta \mathbf{y}) + \frac{2dp_{\mathcal{M}_0}}{d - p_{\mathcal{M}_0} - 1}, \quad (3.2.12)$$

where  $\ell_{\mathcal{M}_0}$  is the log-likelihood of the null model  $\mathcal{M}_0$ ,  $d = nT$  is the size of  $\Delta \mathbf{y}$ , and  $p_{\mathcal{M}_0}$  the corresponding number of parameters. The random-effects model  $\mathcal{M}_1$  is ranked with the conditional akaike information criterion (cAIC) [34] defined as

$$cAIC(\mathcal{M}_1) = -2\ell(\Delta \mathbf{y} | \mathbf{u}; \boldsymbol{\psi}) + 2(\rho + 1), \quad (3.2.13)$$

where  $\ell(\Delta \mathbf{y} | \mathbf{u}; \boldsymbol{\psi})$  is the conditional log-likelihood of the response measurements  $\Delta \mathbf{y}$  given the random effects  $\mathbf{u}$ ,  $\boldsymbol{\psi}$  is the vector of all the unknown parameters, and  $\rho$  is the effective degrees of freedom of  $\mathcal{M}_1$  [36] defined as the trace  $\rho = \text{tr}(\mathbf{H})$  of the hat matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{M} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\mathbf{M} & \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\mathbf{M} \\ \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\mathbf{M} & \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)\mathbf{M} + \Delta_{\mathbf{u}}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \\ \mathbf{M}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \end{bmatrix}. \quad (3.2.14)$$

To measure the distance of the fixed-effects model  $\mathcal{M}_0$  from the mixed-effects model  $\mathcal{M}_1$  we use the the Kullback-Leibler (KL) divergence [37]

$$\begin{aligned} KL_{div}(\mathcal{M}_0 \| \mathcal{M}_1) &= \int p(\Delta \mathbf{y}) \log \frac{p(\Delta \mathbf{y})}{q(\Delta \mathbf{y})} d(\Delta \mathbf{y}) \\ &= \frac{1}{2} \left\{ \text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - d + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \log \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} \right\}, \end{aligned} \quad (3.2.15)$$

where  $p$  and  $q$  are the multivariate Gaussian density functions of Eqs. (3.B.3) and (3.D.1), whose mean vector and covariance matrix are given by

$$\begin{aligned} \boldsymbol{\mu}_0 &= \mathbf{M}\hat{\boldsymbol{\theta}}_0, & \boldsymbol{\Sigma}_0 &= \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0, \hat{\sigma}_0^2), \\ \boldsymbol{\mu}_1 &= \mathbf{M}\hat{\boldsymbol{\theta}}_1 + \mathbf{M}E_{\mathbf{u}|\Delta \mathbf{y}; \hat{\boldsymbol{\psi}}}[\mathbf{u}], & \boldsymbol{\Sigma}_1 &= \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_1, \hat{\sigma}_1^2), \end{aligned} \quad (3.2.16)$$



	$\alpha_A$	$\alpha_B$	$\alpha_C$	$\alpha_D$	$\delta_A$	$\delta_B$	$\delta_C$	$\delta_D$	$\lambda_{A \rightarrow B}$	$\lambda_{A \rightarrow C}$	$\lambda_{C \rightarrow D}$
$c_1$	0.2	0.15	0.17	0.45	0.001	0.007	0.004	0.002	0.13	0.15	0.08
$c_2$	0.2	0.15	0.17	0.09	0.001	0.007	0.004	0.002	0.13	0.15	0.08
$c_3$	0.2	0.15	0.51	0.09	0.001	0.007	0.004	0.002	0.13	0.15	0.08

**Table 3.3.1** | For each synthetic clone (row) the parameter values (columns) used for inference.

3

where  $(\hat{\theta}_0, \hat{\sigma}_0^2)$  and  $(\hat{\theta}_1, \hat{\sigma}_1^2)$  are the parameter estimates for  $\mathcal{M}_0$  and  $\mathcal{M}_1$ . To make model divergences comparable across different sized samples, we use the rescaled KL divergence  $KL_{div}(\mathcal{M}_0 \parallel \mathcal{M}_1) / d$ .

### 3.3. RESULTS

#### 3.3.1. IN SILICO VALIDATION STUDY

We simulated the dynamics of  $J = 3$  distinct clones in four synthetic cell types A, B, C, D following the differentiation network structure of Figure 3.3.1. The net-effect matrix  $\mathbf{V}$  and the hazard vector  $h(\mathbf{x}, \boldsymbol{\theta})$  were derived from Eq. (3.2.2). To simulate the clonal tracking data we used the  $\tau$ -leaping Algorithm 3 of Section 3.A, with a time lag  $\tau = 1$ , that has been run independently for each clone. We designed each simulation so that the first clone dominates lineage D and the third clone dominates lineage C with a sampling frequency  $T = 100$ . The values that were used for the reaction parameters are reported in Table 3.3.1.

We first ran a single simulation under different magnitudes for the noise variance  $\sigma^2$ . Then we fit the random-effects model of Eq. (3.D.1) to the simulated data using Algorithm 5 of Section 3.D. We reported in Figures 3.3.2-3.3.4 the simulated trajectories and a scatterplot of the estimated conditional expectation  $E_{\mathbf{u}|\Delta\mathbf{y};\hat{\boldsymbol{\psi}}}[\mathbf{u}]$  for the random-effects model against the true clone-specific parameters. In the same figure we also reported a piechart where each clone  $k$  is identified with a pie whose slices are lineage-specific and weighted with  $w_k^l$ , defined as the difference between the conditional expectations of the duplication and death parameters, that is

$$w_k^l = E_{\mathbf{u}|\Delta\mathbf{y};\hat{\boldsymbol{\psi}}}[u_{\alpha_l}^k] - E_{\mathbf{u}|\Delta\mathbf{y};\hat{\boldsymbol{\psi}}}[u_{\delta_l}^k], \quad (3.3.1)$$

where  $u_{\alpha_l}^k$  and  $u_{\delta_l}^k$  are the random-effects for duplication and death of clone  $k$  in cell lineage  $l$ . The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of

$$\mathbf{w}_k = (w_k^{l_1}, \dots, w_k^{l_n}), \quad (3.3.2)$$

where  $n$  is the number of cell types. Therefore, the larger the diameter, the more the corresponding clone expanded into the lineage associated to the largest slice. The values of the estimated conditional expectations are reported in Table 3.3.2. The scatterplot of Figures 3.3.2-3.3.4 clearly indicates a strong agreement between the true parameters and the conditional expectations  $E_{\mathbf{u}|\Delta\mathbf{y};\hat{\psi}}[\mathbf{u}]$ . In particular, as expected, as the noise variance  $\sigma^2$  increased, the parameter estimates gradually moved away from the diagonal, so that the precision decreased. Also, our model correctly detected the dominance of clones 1 and 3 in lineages D and C respectively, even for large values of  $\sigma^2$ , as suggested by the pie-charts of Figure 3.3.4 and by the values of Table 3.3.2.

Subsequently, to check goodness-of-fit, we ran 100 independent simulations separately for each noise variance setting. After fitting both the base model of Eq. (3.B.3) and the random-effects model of Eq. (3.D.1), using Algorithms 4 and 5 of Sections 3.C-3.D, the latter always reached a significantly lower AIC compared to the null model, as suggested by the boxplots of Figure 3.3.5. This result clearly indicates that our proposed random-effects stochastic reaction network was able to measure variation between clones in terms of differentiation dynamics and to detect events of clonal dominance.

### 3.3.2. COMPARISON WITH GLS METHOD

We compared our proposed method with the state-of-the-art method GLS [28]. To this end, we have designed two different simulation studies. In the first simulation study all the clones shared the same vector parameter, while in the second study we induced the same clonal expansions of previous section. In both studies we used the differentiation network structure of Figure 3.3.1 as the true generative model from which we simulated clonal trajectories, using the  $\tau$ -leaping Algorithm 3 of Section 3.A, with a time lag  $\tau = 1$ . The net-effect matrix  $\mathbf{V}$  and the hazard vector  $h(\mathbf{x}, \boldsymbol{\theta})$

	$\sigma^2 = .1$			$\sigma^2 = 1$			$\sigma^2 = 10$		
	$c_1$	$c_2$	$c_3$	$c_1$	$c_2$	$c_3$	$c_1$	$c_2$	$c_3$
$\alpha_A$	0.198	0.198	0.199	0.183	0.191	0.198	0.151	0.139	0.127
$\alpha_B$	0.151	0.152	0.148	0.146	0.148	0.145	0.163	0.148	0.137
$\alpha_C$	0.171	0.168	0.509	0.163	0.168	0.518	0.166	0.175	0.649
$\alpha_D$	0.446	0.094	0.098	0.450	0.100	0.121	0.479	0.199	0.319
$\delta_A$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.001
$\delta_B$	0.007	0.007	0.007	0.007	0.007	0.007	0.008	0.007	0.007
$\delta_C$	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.005
$\delta_D$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.004
$\delta_{A \rightarrow B}$	0.129	0.130	0.130	0.129	0.130	0.133	0.127	0.126	0.110
$\delta_{A \rightarrow C}$	0.149	0.150	0.148	0.148	0.149	0.151	0.154	0.155	0.153
$\delta_{C \rightarrow D}$	0.081	0.079	0.079	0.079	0.080	0.078	0.082	0.079	0.058

**Table 3.3.2** | Conditional expectations  $E_{\mathbf{u}|\Delta\mathbf{y},\hat{\boldsymbol{\psi}}}[\mathbf{u}]$  of the random-effects obtained from the estimated parameters  $\hat{\boldsymbol{\psi}}$  for each reaction rate (rows) under different magnitudes of the noise variance  $\sigma^2$  (outer columns) for each clone (inner columns).

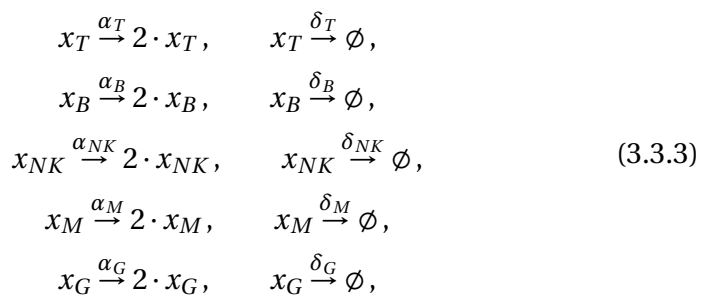
were derived from Eq. (3.2.2). For each study, we ran 100 independent simulations under different noise variance settings ( $\sigma^2 \in \{0.1, 1, 10\}$ ). Subsequently we fit both our proposed method RestoreNet and the competitor method GLS. We reported the results in Figure 3.3.6, showing boxplots of the relative errors between the true parameters and the estimated parameters provided by each method.

Figure 3.3.6 clearly indicates that our proposed inference method RestoreNet overall outperformed the competitor method GLS. Indeed, while in the first simulation study (no clonal dominance) both methods provided similar parameter estimates, in the second simulation study (with clonal dominance) our proposed method RestoreNet provided better parameter estimates compared to GLS. This result suggests that our proposed method RestoreNet was able to infer a cell differentiation network with clone-specific parameters. In conclusion, results from this synthetic study show that our method outperformed the competitor one for the identification of clonal dominance.

### 3.3.3. CLONAL DYNAMICS IN RHESUS MACAQUES

We analysed clonal tracking data collected from an established hematopoietic stem cell model, previously used to investigate hematopoietic reconstitution in Rhesus Macaques [31]. Mobilized peripheral blood (MPB) CD34+ cells from three macaques were transduced with barcoded vectors and, following engraftment, myeloid Granulocytes (G), Monocytes (M), and lymphoid T, B, and Natural Killer (NK) cells were flow sorted for 9.5 months (ZH33), 6.5 months (ZH17), and 4.5 months (ZG66) [38]. The total numbers of clones collected are 1165 (ZH33), 1280 (ZH17), and 1291 (ZG66). Further details on transduction protocols and culture conditions can be found in the original study.

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 total number of reads (see Table 2.J.1 of Section 2.J). This discrepancy made all the samples not directly comparable. Therefore we rescaled the barcode counts according to Eq. (2.J.1) of Section 2.J before analysis. We compared the base and random-effects models on the rhesus macaques clonal tracking data. Since the CD34+ cells were not collected, we only estimated the duplication parameters  $\alpha_T, \alpha_B, \alpha_{NK}, \alpha_M, \alpha_G$  and the death parameters  $\delta_T, \delta_B, \delta_{NK}, \delta_M, \delta_G$  of the lymphoid (T, B, NK) and myeloid (M, G) cells. Therefore the differentiation parameters were not considered in our model, and the net-effect matrix and the hazard vector were obtained from Eqs. (3.2.2)-(3.2.5) accordingly. Thus, the biochemical reactions were defined as



where the left and right columns list the duplication and death reactions, respectively. The corresponding model became effectively a birth/death model including 10 dynamic parameters, one duplication and death rate

		$p$	AIC	$KL_{div}(\mathcal{M}_0\ \mathcal{M}_1)$	$KL_{div}(\mathcal{M}_0\ \mathcal{M}_1)/d$
ZH33	$\mathcal{M}_0$	11.00	81377.27		
	$\mathcal{M}_1$	434.16	38160.15	21062.95	1.87
ZH17	$\mathcal{M}_0$	11.00	336752.11		
	$\mathcal{M}_1$	478.43	29478.05	291854802.44	114228.89
ZG66	$\mathcal{M}_0$	11.00	31194.60		
	$\mathcal{M}_1$	410.92	21384.85	232030.37	83.77

**Table 3.3.3** | Comparison between fixed-effects  $\mathcal{M}_0$  and mixed-effects  $\mathcal{M}_1$  models: Number of parameters ( $p$ ), AIC, KL divergence  $KL_{div}(\mathcal{M}_0\|\mathcal{M}_1)$  and rescaled KL divergence  $KL_{div}(\mathcal{M}_0\|\mathcal{M}_1)/d$  in each rhesus macaque.

for each lineage. We fit both the fixed-effect model of Eq. (3.B.3) and the mixed-effects model of Eq. (3.D.1) separately to the data of each animal. To further remove bias, we focused our analyses on the clones that were recaptured at least 5 times. This resulted in a number of clones  $J$  equal to 481 (ZH33), 139 (ZH17), and 202 (ZG66), and in 6 (ZH33), 5 (ZH17), and 4 (ZG66) time points.

We reported the results on model selection in Table 3.3.3, and the estimated parameters  $\hat{\psi}$  in Table 3.3.4. Using the estimated parameters  $\hat{\psi}$ , following Eq. (3.D.5), we computed the net conditional expectations of Eq. (3.3.1), which we used as a proxy for the clone-specific net-duplication  $\alpha_l - \delta_l$  in each cell lineage  $l$ . The resulting values are reported in Figure 3.3.7 in a box-plot fashion. Subsequently, in Figure 3.3.8 we proposed to use a weighted pie chart to visualize our findings at clonal level. Consistently with previous section, each pie, corresponding to a particular clone, was weighted by its net conditional expectations, as defined in Eq. (3.3.1).

As a result, according to the AIC values, in each animal the mixed-effects model ( $\mathcal{M}_1$ ) outperformed the fixed-effects one ( $\mathcal{M}_0$ ). This means that the clones did not follow the same average dynamics for the birth-death process. Instead, the dynamic of some clones departed from the average dynamics with a significant (random) effect. In particular, the conditional net-duplication rates  $E_{\mathbf{u}|\Delta\mathbf{y};\hat{\psi}}[u_{\alpha_l}^k] - E_{\mathbf{u}|\Delta\mathbf{y};\hat{\psi}}[u_{\delta_l}^k]$  of Figures 3.3.7 - 3.3.8 suggest events of clonal dominance in specific cell lineages. As an example, for the animals ZH33 and ZG66 we observed clonal expansions into NK cells. Whereas, for the animal ZH17 we observed clonal expansions into G and B cell lineages. Finally, for the animal ZG66 we also observed

	ZH33		ZH17		ZG66	
	$\theta$	$\tau^2$	$\theta$	$\tau^2$	$\theta$	$\tau^2$
$\alpha_T$	0.813	1.176	2.246	1.051	1.081	2.702
$\alpha_B$	0.193	0.597	6.503	4.648	0.055	0.876
$\alpha_{NK}$	0.758	2.253	2.435	2.364	1.095	1.943
$\alpha_G$	0.197	0.403	10.931	53.216	0.847	1.318
$\alpha_M$	0.360	0.547	3.298	4.256	2.198	1.800
$\delta_T$	0.155	0.074	0.172	0.741	0.039	0.059
$\delta_B$	0.102	0.059	2.159	36.268	0.006	0.051
$\delta_{NK}$	0.228	0.089	0.223	0.406	0.098	0.100
$\delta_G$	0.039	0.029	13.211	70.756	0.018	0.017
$\delta_M$	0.100	0.059	0.012	0.018	0.035	0.019

**Table 3.3.4** | Parameter estimated for the proposed mixed effects model: Fixed effects ( $\theta$ ) and variance ( $\tau^2$ ) of the random effects for both the duplication  $\alpha$  and death  $\delta$  parameters for each cell lineage and each rhesus macaque.

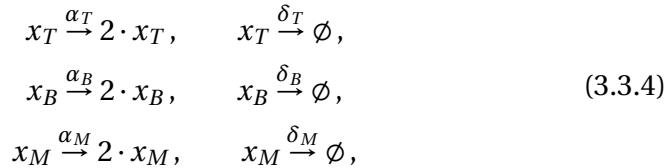
events of clonal dominance into M and T cell lineages. Furthermore, the weighted pie charts from Figure 3.3.8 revealed different gradients of clonal dominance between the three rhesus macaques. As an example, by looking at the size of the pies, it is possible to observe an higher clonal dominance of NK cells in ZH33, and of G cells in ZH17, compared to the expansions of M, NK and T cells detected in ZG66, where the diameters of the clone-specific pies are rather similar. Not only the proposed mixed-effects model detected clonal dominance in certain cell types, it also detected which clones were responsible.

### 3.3.4. GENOTOXIC EFFECTS ON CLONAL DYNAMICS

We analyzed an in-vivo clonal tracking dataset previously used in [32] to investigate clonal diversity in tumor-prone mice under two different treatment conditions. *Cdkn2a*<sup>-/-</sup> tumor prone *Lin*<sup>-</sup> cells were 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, from myeloid cells (CD11b+) and unsorted blood cells (total MNC). Clonal tracking samples were collected under heterogeneous technical conditions as reported in Table 2.I.1 of Section 2.I. These confounding effects made the samples not directly comparable. Therefore we rescaled the samples following the description in Section 2.I before analysis. The total number of distinct clones collected were 45186 and 20471 for the PGK and SFV treatments respectively. To further remove bias, we focused our analyses on the top  $J = 1000$  most recaptured clones across lineages and time. The number of time-points  $T$  was equal to 7 (PGK) and 6 (SFV).

Next, we compared the fixed-effects model of Eq. (3.B.3) and the mixed-effects model of Eq. (3.D.1) on the rescaled clonal tracking data, so as to compare the dynamics of clonal dominance under the two viral vector conditions. Since the HSCs were not collected, we only estimated the duplication parameters  $\alpha_T, \alpha_B, \alpha_M$  and the death parameters  $\delta_T, \delta_B, \delta_M$  of the lymphoid (T, B) and myeloid (M) cells. Therefore, by analogy to previous section the differentiation parameters were not considered in our model, and the net-effect matrix and the hazard vector were obtained from Eqs. (3.2.2)-(3.2.5) accordingly. Therefore, the biochemical reactions were defined as



where the left and right columns list the duplication and death reactions, respectively. We fit both the fixed-effects model of Eq. (3.B.3) and the mixed-effects model of Eq. (3.D.1) separately to the data of each vector treatment. Both models included six dynamic parameters, that is one scalar value for each combination of cell type with duplication and death reactions.

We reported the results on model selection in Table 3.3.5, and the estimated parameters  $\hat{\psi}$  in Table 3.3.6. Then, from the estimated parameters  $\hat{\psi}$  we computed the conditional expectations of Eq. (3.3.1), which we used as a proxy for the clone-specific net-duplication  $\alpha_l - \delta_l$  in each cell lineage  $l$ . By analogy to previous section, the resulting values are reported in Fig-

		$p$	AIC	$KL_{div}(\mathcal{M}_0\ \mathcal{M}_1)$	$KL_{div}(\mathcal{M}_0\ \mathcal{M}_1)/d$
PGK	$\mathcal{M}_0$	7.00	115997.43		
	$\mathcal{M}_1$	471.40	65083.07	17098.71	1.29
SFV	$\mathcal{M}_0$	7.00	63520.89		
	$\mathcal{M}_1$	842.00	30147.56	52431.53	6.51

**Table 3.3.5** | Comparison between fixed-effects  $\mathcal{M}_0$  and mixed-effects  $\mathcal{M}_1$  models: Number of parameters ( $p$ ), AIC, KL divergence  $KL_{div}(\mathcal{M}_0\|\mathcal{M}_1)$  and rescaled KL divergence  $KL_{div}(\mathcal{M}_0\|\mathcal{M}_1)/d$  in each treatment group.

	PGK		SFV	
	$\theta$	$\tau^2$	$\theta$	$\tau^2$
$\alpha_M$	0.058	1.014	1.287	5.781
$\alpha_B$	0.092	0.872	0.024	0.408
$\alpha_T$	0.632	2.625	3.367	2.824
$\delta_M$	0.095	0.041	0.232	0.085
$\delta_B$	0.079	0.028	0.156	0.080
$\delta_T$	0.127	0.044	0.437	0.193

**Table 3.3.6** | Parameter estimated for the proposed mixed effects model: Fixed effects ( $\theta$ ) and variance ( $\tau^2$ ) of the random effects for both the duplication  $\alpha$  and death  $\delta$  parameters for each cell lineage and each treatment group.

ure 3.3.9 in a box-plot fashion, while in Figure 3.3.10 we proposed to use a weighted pie chart to visualize our findings at clonal level.

As a result, according to the AIC values, under each treatment the mixed-effects model ( $\mathcal{M}_1$ ) outperformed the fixed-effects one ( $\mathcal{M}_0$ ). This means that the clones exhibited heterogeneity in their dynamics for the birth/death process. The dynamics of some clones departed from the average dynamics with a significant (random) effect. In particular, the conditional net-duplication rates of Eq. (3.3.1) from Figures 3.3.9 - 3.3.10 suggest events of clonal dominance in specific cell lineages. For example, under the PGK treatment we observed clonal expansions into T cells. Whereas, under the SFV treatment we observed clonal expansions into M and T cell lineages with even higher conditional rates compared to PGK. Furthermore, the Kullback-Leibler divergence from Table 3.3.5 revealed a different gradient of clonal dominance between the two treatments, suggesting that the clonal expansions identified in the SFV case were more




significant compared to PGK.

### 3.4. DISCUSSION AND CONCLUSION

In this work we proposed a random-effects cell differentiation network which takes into account heterogeneity in the dynamics across the clones. Our framework extends the clone neutral local linear approximation of a stochastic quasi-reaction network, written in the Ito formulation, by introducing random-effects for the clones on the dynamics parameters to allow for clonal dominance. We used a maximum likelihood approach to infer the parameters of the base (fixed-effects only) model that are then used as initial values for the estimation of the random-effects model by means of an E-M algorithm. We tested our framework with a  $\tau$ -leaping simulation study, showing accurate performance of the method in the identification of a clonal expansion and in the inference of the true parameters. Then, by means of an additional in-silico study, we have shown that our method outperforms the state-of-the-art method GLS [28]. Subsequently, the application of our proposed method on a rhesus macaque clonal tracking study revealed significant clonal dominance for specific cell types. Particularly interesting is that the NK clonal expansions detected by our model were already observed by former studies [31, 39, 40], and therefore our findings are consistent with those previously obtained. Indeed [39] described the oligoclonal expansions of NK cells and the long-term persistence of HSPCs and immature NK cells. Finally, our proposed method allowed to detect the expected impact of vector genotoxicity on clonal dynamics in a tumor-prone mice model of haematopoiesis, as already observed in a previous study [32].

The main approximation, in both the basal and random-effects formulations, is the piece-wise linearity of the process. In both cases we consider first a local linear approximation of the Kramers-Moyal approximated Master equation, which is then used to infer the process parameters either with or without random-effects. Although the linearity assumption makes all the computations easier, this approximation becomes poor as the time lag increments (the  $\Delta t$ s) of the collected data increase. This can be addressed by introducing in the likelihood higher-order approximation terms than the ones considered by the Euler-Maruyama method. The Milstein approx-

imation is a possible choice [41]. Another, completely different, approach is to employ extended Kalman filtering (EKF) which is suitable for non-linear state space formulations [42]. Also, our framework cannot consider false-negative errors or missing values of clonal tracking data. Also for this issue, an EKF formulation could be a possible extension. The frequentist-based inference step of our proposed E-M algorithm may be replaced by Bayesian alternatives. For example, the E-step function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  could be replaced by a Metropolis-Hasting step [43, 44]. Alternatively, a variational Bayes method could be employed, where the unknown vector parameter  $\boldsymbol{\psi}$  is treated as an additional latent variable [45]. Our future work will aim to extend the  package RestoreNet by including other types of reactions (besides cell duplication, cell death and cell differentiation).


Our tool can be considered as complementary to the classical Shannon entropy index [32] in detecting fast and uncontrolled growing of clones after a gene therapy treatment. Indeed, while the Shannon entropy measures the diversity of a population of clones as a whole, RestoreNet provides a clone-specific quantification of dominance in terms of conditional mean and variance of the expansion rates. Our proposed method provides a prototype model of clonal haematopoiesis whose parameters are calibrated to fit high-dimensional clonal tracking data. Our data-driven model can be integrated with those obtained with alternative approaches, where the unknown parameters are either set to experimentally-derived quantities, computed from the steady states, or based on independent studies [46, 47].

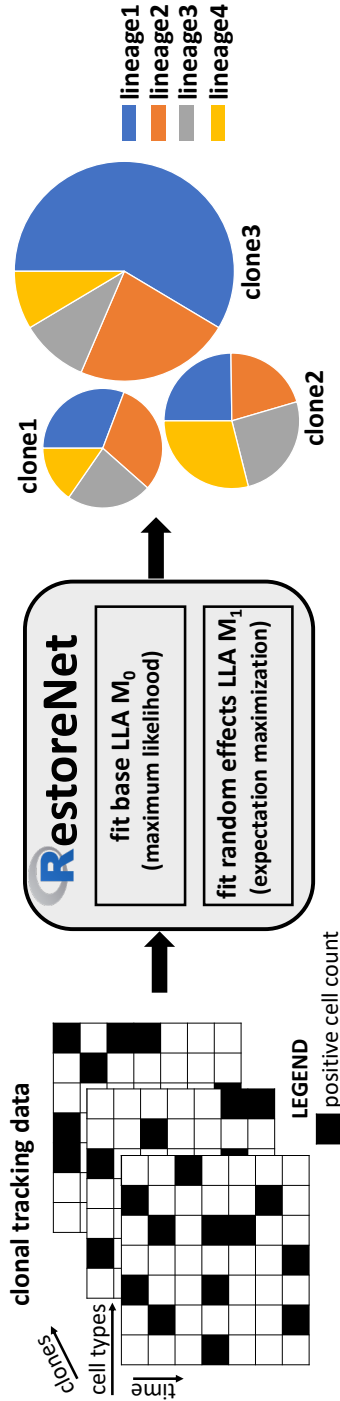
In conclusion, our proposed stochastic framework is able to detect deviant clonal behaviour relative to the average dynamics of haematopoiesis. This is an important aspect for gene therapy applications where it is crucial to quickly detect any adverse event that may be related to clonal dominance. Therefore our tool can provide statistical support in gene therapy surveillance analyses. Our proposed method also has potential applications in other biomedical longitudinal studies with subject-specific dynamics, such as population infection dynamics [48, 49], population analysis of tumor development [50], and genetic regulatory networks [51]. Moreover, our proposed mixed-effects formulation of stochastic quasi-reaction networks can potentially be applied to more general, non-Markovian, classes of network models, such as stochastic hybrid systems with memory (SHSM). This more general class of models suits history-dependent bio-

logical systems, such as neural dynamics and immune responses [52, 53]. A mixed-effects formulation of dynamical systems may find room also in optimal investments problems, such as stochastic games in a continuous-time Markov regime-switching environment [54]. Indeed, if such models can be written in a Ito-type formulation, mixed-effects on sensible subjects (e.g. groups of investors in a market) can be incorporated.

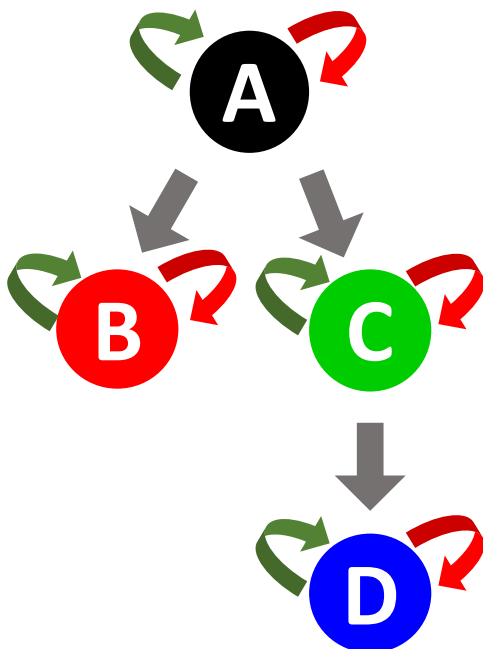
**3**

### **3.5. AVAILABILITY OF DATA AND MATERIALS**

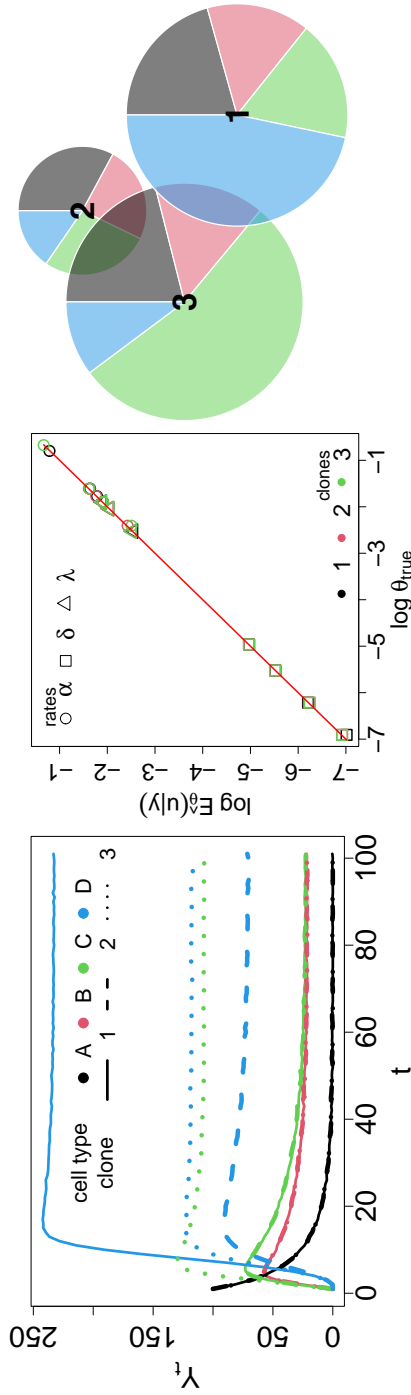
The code that supports the findings of this study is openly available at <https://github.com/delcore-luca/ClonalDominance>. The  package RestoreNet is publicly available for download from CRAN.



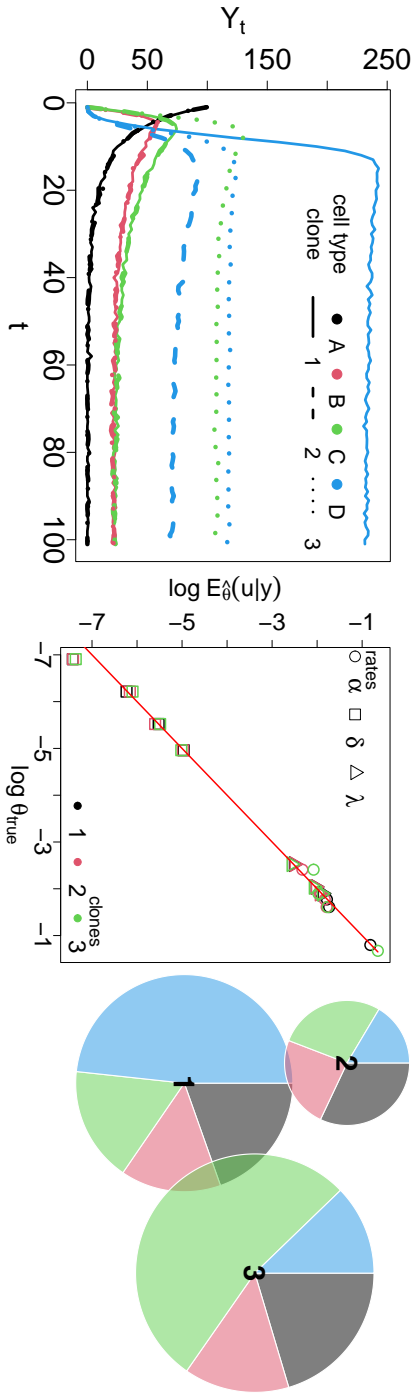
**Figure 3.2.1** | Schematic representation of the analysis: A three-dimensional clonal tracking dataset (left) is received as input from our proposed stochastic framework RestoreNet (middle). It mainly consists in two parts, such as a maximum likelihood step to infer the base LLA model, and an expectation-maximization step to infer the random-effects LLA formulation. Finally, a clonal piechart is returned, where each clone is identified by a pie whose slices are lineage specific and proportional to their expansion rates (right).



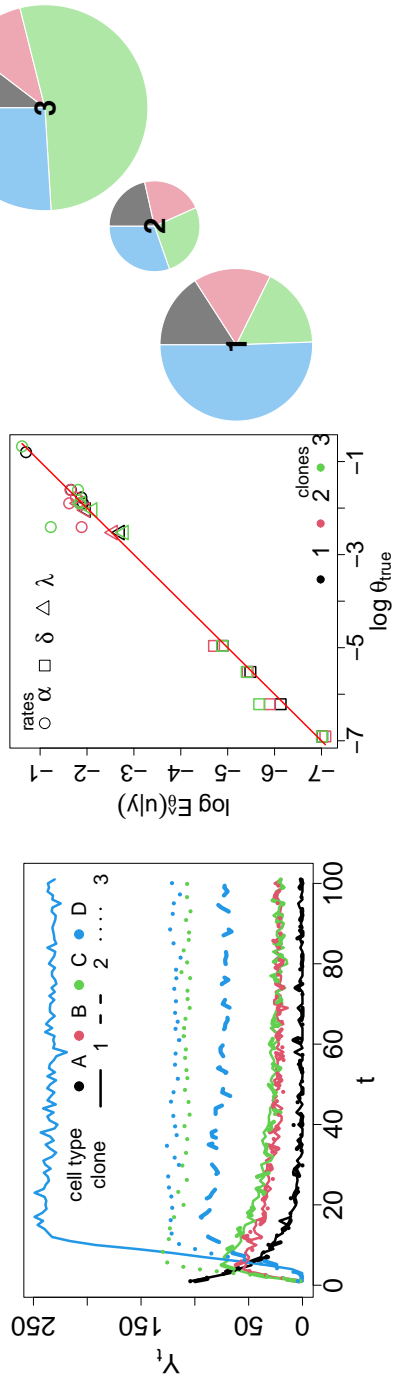
**Figure 3.3.1** | Differentiation structure of four synthetic cell types A, B, C, D. Cell duplication, death and differentiation are indicated with green, red and grey arrows.



**Figure 3.3.2** | Simulation results - part 1: (left): Simulated trajectories. (center): Scatterplot between the clone-specific true parameters  $\theta_{true}$  and the conditional expectation  $E_{u|\Delta y, \psi}[u]$ . (right): Clonal pie-charts where each clone  $k$  is identified with a pie whose slices are lineage-specific and weighted according to Eq. (3.3.1). The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of  $w_k$ , as defined in Eq. (3.3.2). The synthetic noise variance has been set to  $\sigma^2 = 0.1$ .

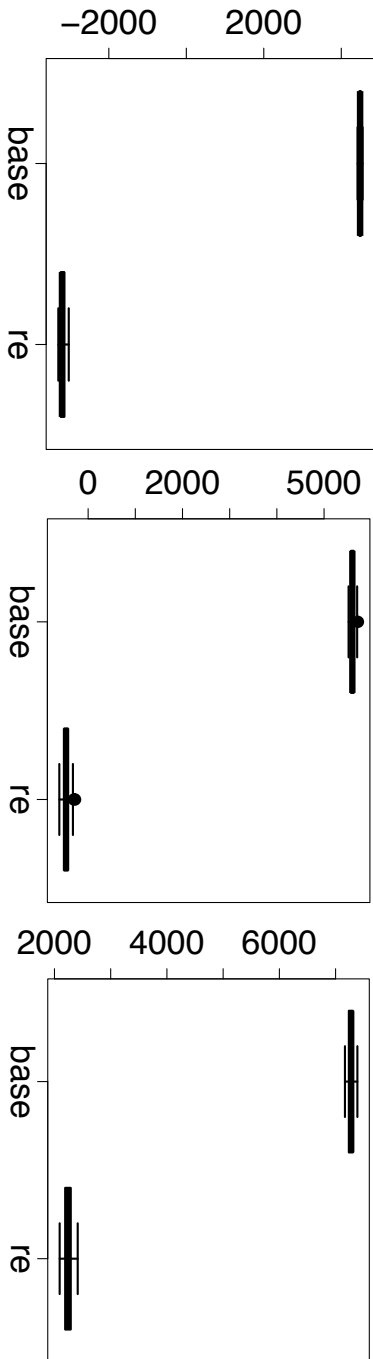


**Figure 3.3.3** | Simulation results - part 2: (left): Simulated trajectories. (center): Scatterplot between the clone-specific true parameters  $\theta_{true}$  and the conditional expectation  $E_{\mathbf{u}|\Delta y, \hat{\psi}}[\mathbf{u}]$ . (right): Clonal pie-charts where each clone  $k$  is identified with a pie whose slices are lineage-specific and weighted according to Eq. (3.3.1). The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of  $\mathbf{w}_k$ , as defined in Eq. (3.3.2). The synthetic noise variance has been set to  $\sigma^2 = 1$ .

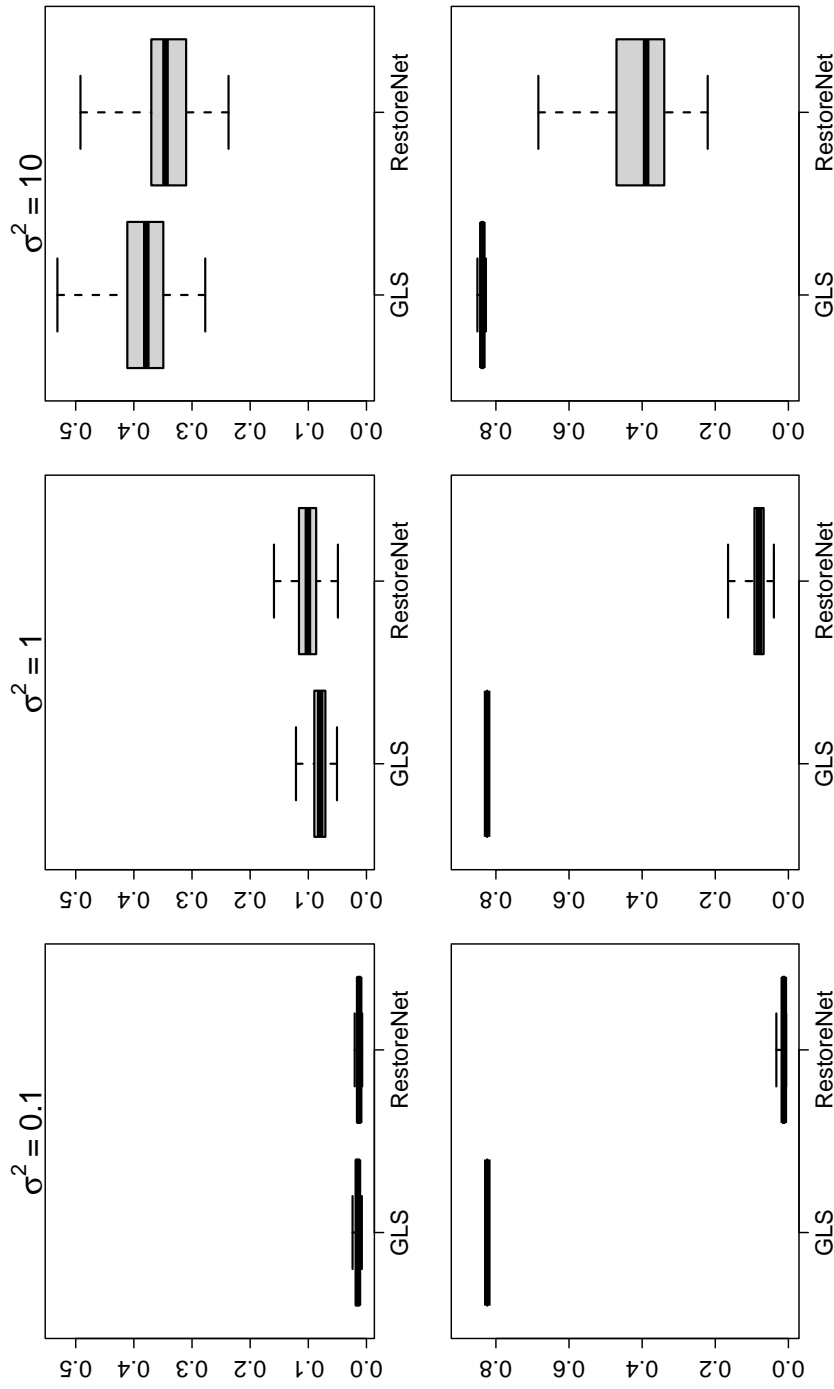


**Figure 3.3.4** | Simulation results - part 3: (left): Simulated trajectories. (center): Scatterplot between the clone-specific true parameters  $\theta_{true}$  and the conditional expectation  $E_{\mathbf{u}|\Delta y; \hat{\psi}}[\mathbf{u}]$ . (right): Clonal pie-charts where each clone  $k$  is identified with a pie whose slices are lineage-specific and weighted according to Eq. (3.3.1). The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of  $\mathbf{w}_k$ , as defined in Eq. (3.3.2). The synthetic noise variance has been set to  $\sigma^2 = 10$ .



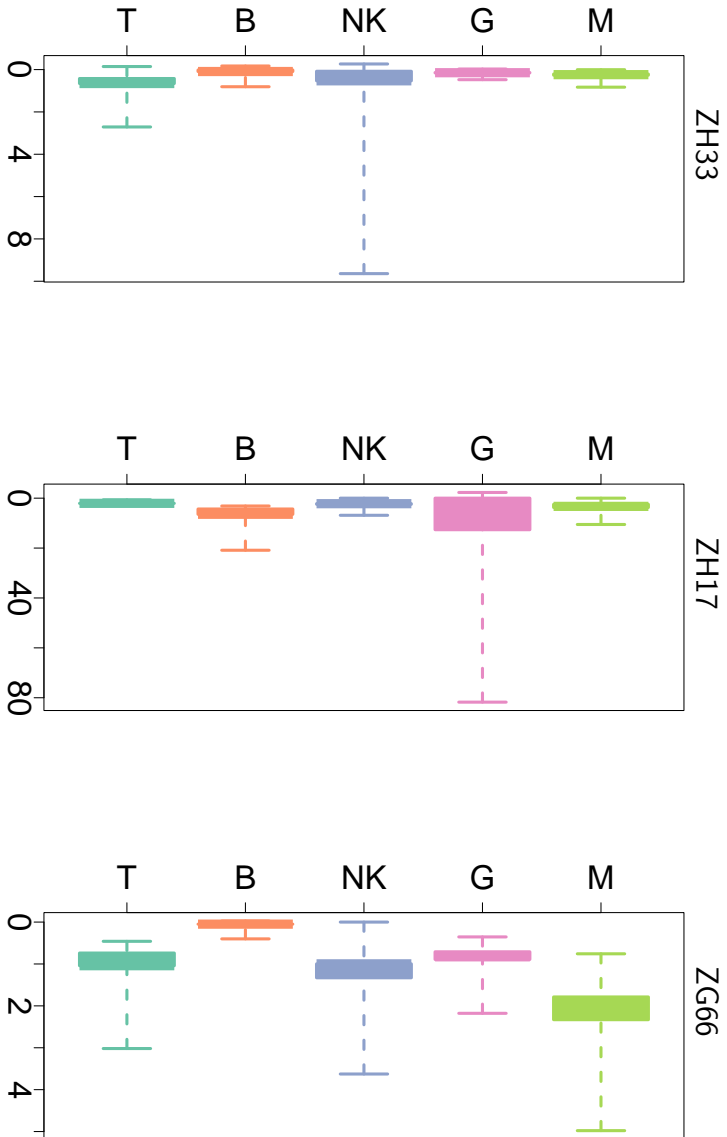


**Figure 3.3.5** | Boxplot of the AICs of the fixed-effects (base) and random-effects (re) models under a measurement noise level equal to 0.1 (left), 1 (center) and 10 (right).

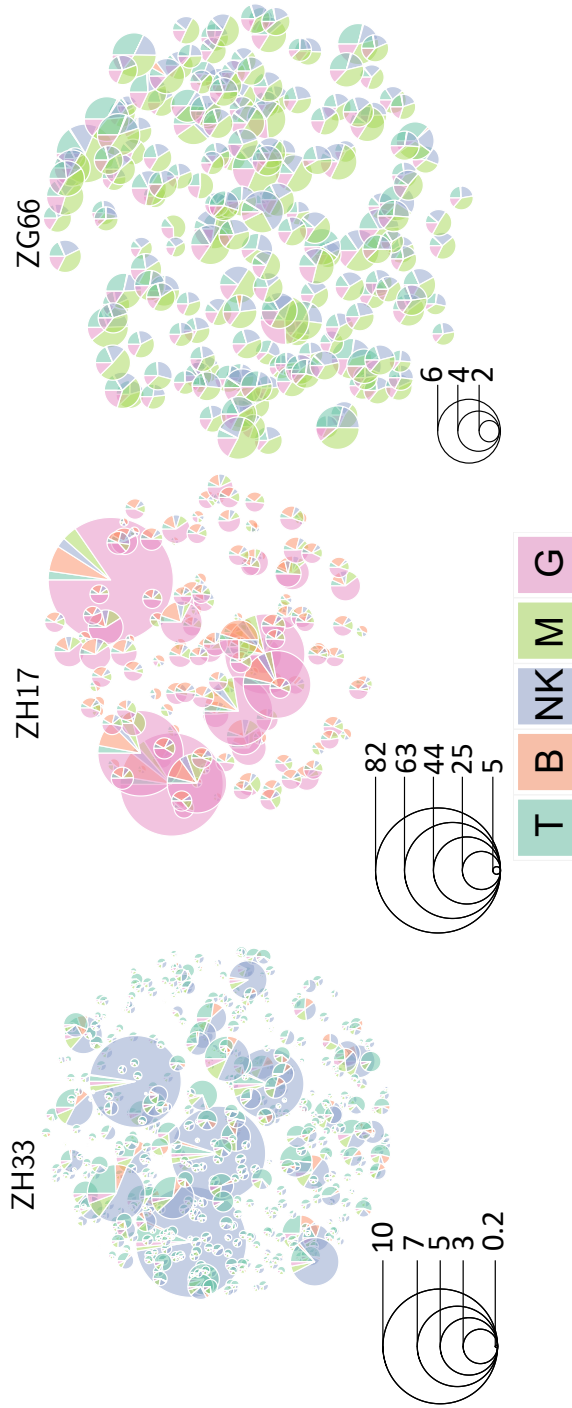


**Figure 3.3.6** | Boxplots of the relative errors between the true parameters and the estimated parameters provided by each candidate method (x-axis) for simulation study 1 (top) and 2 (bottom) under each noise variance setting (columns).

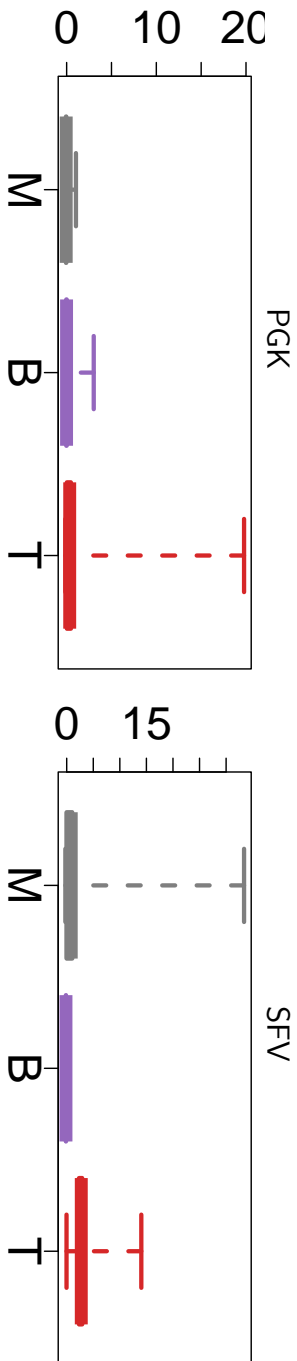




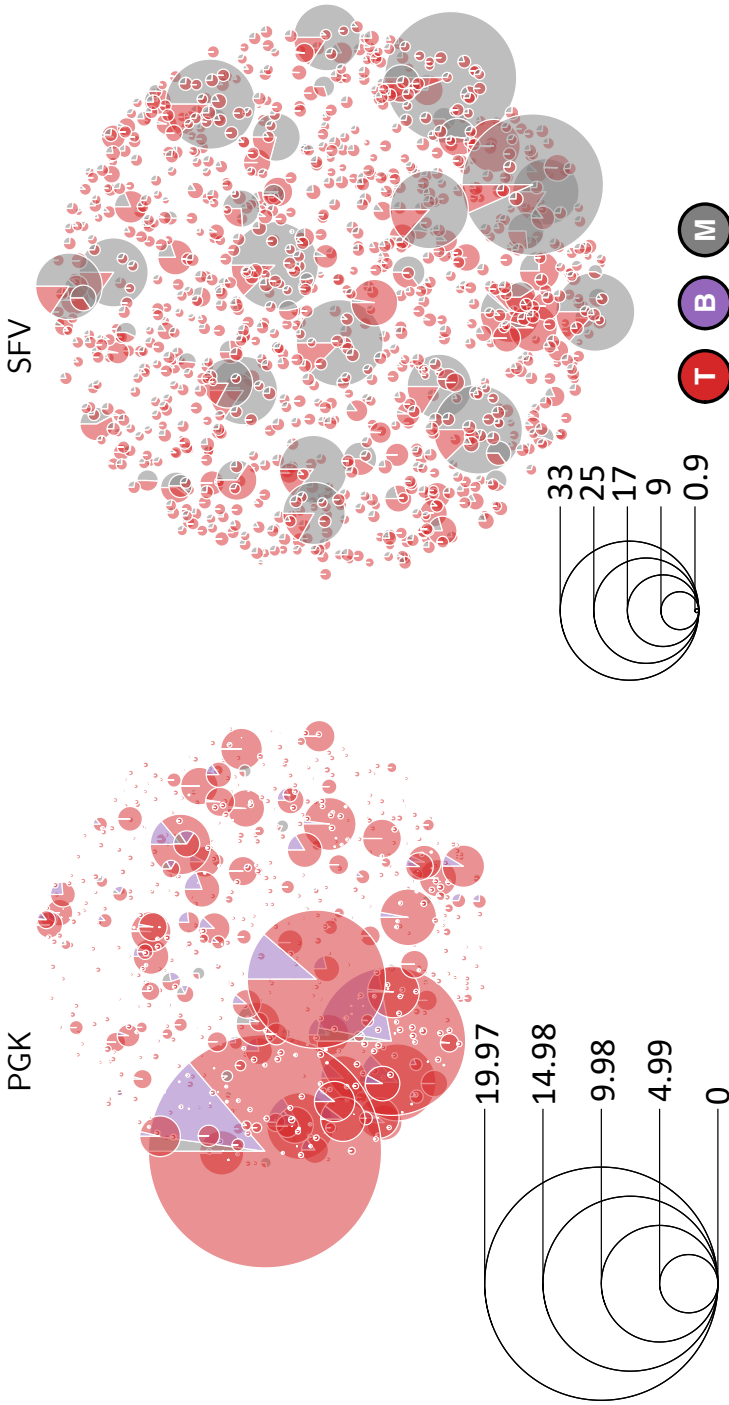
**Figure 3.3.7** | For each animal analyzed (columns), the boxplots of the conditional expectations  $E_{\mathbf{u}|\Delta\mathbf{y},\psi} [u_{\alpha_l}^k] - E_{\mathbf{u}|\Delta\mathbf{y},\psi} [u_{\delta_l}^k]$  computed from the estimated parameters  $\hat{\psi}$  for the clone-specific net-duplication  $\alpha_l - \delta_l$  in each cell lineage  $l$  (different colors). The whiskers extend to the data extremes.



**Figure 3.3.8** | Estimated clonal pie-charts for the rhesus macaques ZH33 (left), ZH17 (center) and ZG66 (right): Each  $k$ -th clone is identified with a pie whose slices are lineage-specific and weighted according to Eq. (3.3.1). The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of  $w_k$ , as defined in Eq. (3.3.2). The legend scales are different across the three plot panels.



**Figure 3.3.9** | For each treatment group (columns), the boxplots of the conditional expectations of Eq. (3.3.1) computed from the estimated parameters  $\hat{\psi}$  for the clone-specific net-duplication  $\alpha_l - \delta_l$  in each cell lineage  $l$  (different colors). The whiskers extend to the data extremes.



**Figure 3.3.10** | Estimated clonal pie-charts under the vector treatments PGK (left) and SFV (right): Each  $k$ -th clone is identified with a pie whose slices are lineage-specific and weighted according to Eq. (3.3.1). The diameter of the  $k$ -th pie is proportional to the euclidean 2-norm of  $\mathbf{w}_k$ , as defined in Eq. (3.3.2). The legend scales are different across the two plot panels.





# APPENDIX

## 3.A. $\tau$ -LEAPING ALGORITHM

A  $\tau$ -leaping algorithm is an alternative method to a Gillespie algorithm for simulating triggering-chain events. Instead of simulating a waiting time for the first reaction to occur and selecting the corresponding winner reaction, a  $\tau$ -leaping algorithm simulates the number of occurrences of each possible event after a time-lag equal to  $\tau$  elapsed. Formally, let  $\{N_r(t)\}_{t \geq 0}$  be an inhomogeneous Poisson point process representing the number of reactions of type  $r$  that took place up to (and including) time  $t$ . Therefore

$$N_r(t) \sim \text{Poisson} \left( \int_0^t \theta_r(s) ds \right), \quad (3.A.1)$$

$$E[N_r(t + \Delta t) - N_r(t)] = \int_t^{t+\Delta t} \theta_r(s) ds \hat{=} \Theta_t^r. \quad (3.A.2)$$

The last equation gives an estimate of the expected number of reactions of type  $r$  that took place within the time interval  $[t, t + \Delta t[$ . Therefore, the expected number of molecules  $\mathbf{y}_{t+\Delta t}$  at time  $t + \Delta t$  given the current number of molecules  $\mathbf{y}_t$  can be easily obtained by adding to  $\mathbf{y}_t$  the product between the expected number of events  $E[N_r(t + \Delta t) - N_r(t)]$  that have happened in the time interval  $[t, t + \Delta t[$ , the corresponding net-effect, and the number of ways that reaction can occur, leading to

$$\mathbf{y}_{t+\Delta t} = \mathbf{y}_t + \mathbf{V} \begin{bmatrix} \Theta_t^1 \prod_{i=1}^n \binom{y_{it}}{r_{1i}} \\ \vdots \\ \Theta_t^K \prod_{i=1}^n \binom{y_{it}}{r_{Ki}} \end{bmatrix} \quad (3.A.3)$$

The pseudocode of the  $\tau$ -leaping algorithm is reported in Algorithm 3.



**Input:**  $S$  (no. simulations),  $\mathbf{y}_0$  (initial state),  
 $\tau$  (time lag),  $\theta(t)$  (reaction rates)  
**Output:**  $\{\mathbf{y}_t\}_t$   
 $t \leftarrow 0$ ;  
 $\mathbf{y}_t \leftarrow \mathbf{y}_0$ ;  
**for**  $s = 1 : S$  **do**  
    **for**  $r = 1 : K$  **do**  
         $\Theta_t^r = \int_t^{t+\Delta t} \theta_r(s) ds$ ;  
    **end**  
     $\mathbf{y}_{t+\Delta t} \leftarrow \mathbf{y}_t + \mathbf{V} \begin{bmatrix} \Theta_t^1 \prod_{i=1}^n (y_{it}) \\ \vdots \\ \Theta_t^K \prod_{i=1}^n (y_{it}) \end{bmatrix}$ ;  
     $t \leftarrow t + \tau$ ;  
**end**

3

**Algorithm 3:**  $\tau$ -leaping algorithm

### 3.B. EULER-MARUYAMA APPROXIMATION

**Remark 3.B.1. (Generalized Linear Model (GLM) formulation)**

Using previous results and some linear algebra, the approximated Ito equation (2.B.3) can be further approximated as

$$\Delta \mathbf{y}_t = \mathbf{V} \begin{bmatrix} \prod_{i=1}^n (y_{it}) \\ \vdots \\ \prod_{i=1}^n (y_{it}) \end{bmatrix} \Delta t \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\boldsymbol{\theta}} + \left( \mathbf{V} \underbrace{\begin{bmatrix} h_1(\mathbf{y}_t, \boldsymbol{\theta}) \\ \vdots \\ h_1(\mathbf{y}_t, \boldsymbol{\theta}) \end{bmatrix}}_{\mathbf{W}_t(\boldsymbol{\theta})} \mathbf{V}' \Delta t + \sigma^2 \mathbf{I}_n \right)^{1/2} \Delta \boldsymbol{\varepsilon}_t, \quad (3.B.1)$$

$\Delta \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n),$

or more compactly

$$\Delta \mathbf{y}_t = \mathbf{M}_t \boldsymbol{\theta} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_N(\mathbf{0}, \mathbf{W}_t(\boldsymbol{\theta}) + \sigma^2 \mathbf{I}_n), \quad (3.B.2)$$

where we included the term  $\sigma^2 \mathbf{I}_N$  so as to prevent singularity of the diffusion term, and to additionally explain noise variance. In practice, since we collect only discrete-time increments  $\Delta \mathbf{y}_t = \mathbf{y}_{t+\Delta t} - \mathbf{y}_t$ , we consider an Euler-Maruyama local linear approximation (LLA) of the approximated Ito

equation. Indeed we also replaced the infinitesimal increments  $dt$  and  $d\mathbf{y}_t$  with the discrete increments  $\Delta t$  and  $\Delta\mathbf{y}_t$ . Then, all the time-specific blocks can be stacked together obtaining the full generalized linear model (GLM) formulation

$$\underbrace{\begin{bmatrix} \Delta\mathbf{y}_{t_0} \\ \vdots \\ \Delta\mathbf{y}_{t_{T-1}} \end{bmatrix}}_{\Delta\mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{M}_{t_0} \\ \vdots \\ \mathbf{M}_{t_{T-1}} \end{bmatrix}}_{\mathbf{M}} \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N} \left( \mathbf{0}, \underbrace{\begin{bmatrix} \Sigma(\boldsymbol{\theta}, \sigma^2) \\ \mathbf{W}_{t_0}(\boldsymbol{\theta}) & & \\ & \ddots & \\ & & \mathbf{W}_{t_{T-1}}(\boldsymbol{\theta}) \end{bmatrix}}_{\mathbf{W}(\boldsymbol{\theta})} + \sigma^2 \mathbf{I}_{nT} \right), \quad (3.B.3)$$

which is convenient for parameters inference.

### 3.C. INFERENCE OF THE BASE GLM MODEL

We infer the parameters  $(\boldsymbol{\theta}, \sigma^2)$  of (3.B.3) with a maximum likelihood approach, that is we solve the following constrained optimization problem

$$\hat{\boldsymbol{\theta}}_{ML} \leftarrow \underset{\boldsymbol{\theta} \geq \mathbf{0}; \sigma^2 \geq 0}{\operatorname{argmin}} f(\boldsymbol{\theta}, \sigma^2), \quad (3.C.1)$$

where the objective function is

$$f(\boldsymbol{\theta}, \sigma^2) = \log(|\mathbf{W}_*|) + (\mathbf{d}\mathbf{y} - \mathbf{M}\boldsymbol{\theta})' \mathbf{W}_*^{-1} (\mathbf{d}\mathbf{y} - \mathbf{M}\boldsymbol{\theta}), \quad (3.C.2)$$

and we compactly write the diffusion matrix  $\mathbf{W}_* = \mathbf{W}(\boldsymbol{\theta}, \sigma^2)$  as a function of the free parameters. Using the rules of matrix calculus [55], the partial derivatives of  $f$  w.r.t.  $\boldsymbol{\theta}$  and  $\sigma^2$  can be written as

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \sigma^2) &= \nabla_{\boldsymbol{\theta}} \log(|\mathbf{W}_*|) + \mathbf{d}\mathbf{y}' \nabla_{\boldsymbol{\theta}} \mathbf{W}_*^{-1} \mathbf{d}\mathbf{y} + 2\boldsymbol{\theta}' \mathbf{M}' \mathbf{W}_*^{-1} \mathbf{M} + \\ &\quad - 2(\mathbf{M}' \mathbf{W}_*^{-1} + \boldsymbol{\theta}' \mathbf{M}' \nabla_{\boldsymbol{\theta}} \mathbf{W}_*^{-1}) \mathbf{d}\mathbf{y} + \boldsymbol{\theta}' \mathbf{M}' \nabla_{\boldsymbol{\theta}} \mathbf{W}_*^{-1} \mathbf{M}\boldsymbol{\theta}, \end{aligned} \quad (3.C.3)$$

$$\begin{aligned} \nabla_{\sigma^2} f(\boldsymbol{\theta}, \sigma^2) &= \nabla_{\sigma^2} \log(|\mathbf{W}_*|) + \mathbf{d}\mathbf{y}' \nabla_{\sigma^2} \mathbf{W}_*^{-1} \mathbf{d}\mathbf{y} + \\ &\quad - 2\boldsymbol{\theta}' \mathbf{M}' \nabla_{\sigma^2} \mathbf{W}_*^{-1} \mathbf{d}\mathbf{y} + \boldsymbol{\theta}' \mathbf{M}' \nabla_{\sigma^2} \mathbf{W}_*^{-1} \mathbf{M}\boldsymbol{\theta} \\ &\quad \operatorname{tr}(\mathbf{W}_*^{-1}) - (\mathbf{d}\mathbf{y} - \mathbf{M}\boldsymbol{\theta})' \mathbf{W}_*^{-1} \mathbf{W}_*^{-1} (\mathbf{d}\mathbf{y} - \mathbf{M}\boldsymbol{\theta}), \end{aligned} \quad (3.C.4)$$

**Input:**  $M, \Delta y$


**Output:**  $\hat{\theta}_{ML}^p$

$$\hat{\theta}_{ML}^p \leftarrow \underset{\theta_k \geq 0; \sigma^2 \geq 0}{\operatorname{argmin}} \{ \log(|W_*|) + (\Delta y - M\theta)' W_*^{-1} (\Delta y - M\theta) \}$$

**Algorithm 4:** Maximum Likelihood inference for the base model.

where

$$\begin{aligned} \frac{\partial}{\partial \theta_j} W_*^{-1} &= -W_*^{-1} \frac{\partial}{\partial \theta_j} W_* W_*^{-1}, & \frac{\partial}{\partial \theta_j} W_* &= W((\dots, 1, \dots), 0), \\ \frac{\partial}{\partial \sigma^2} W_*^{-1} &= -W_*^{-1} W_*^{-1}, & \frac{\partial}{\partial \theta_j} \log|W_*| &= \operatorname{tr} \left( W_*^{-1} \frac{\partial}{\partial \theta_j} W_* \right), \\ & & \frac{\partial}{\partial \sigma^2} \log|W_*| &= \operatorname{tr} (W_*^{-1}). \end{aligned} \quad (3.C.5)$$

Then, we solve the optimization problem (3.C.1) by using the objective function (3.C.2) and its gradients (3.C.3)-(3.C.4) inside the L-BFGS-B optimization algorithm from the `optim()` function of the `stats`  package. The inference procedure is summarised in Algorithm 4.

### 3.D. RANDOM-EFFECTS REACTION NETWORKS

From Eq. (3.B.3) it can be seen that all the molecules  $y_1, \dots, y_n$  share the same parameter vector  $\theta$ . In some cases it may happen that the molecules being analysed are drawn from a hierarchy of  $J$  different populations having different properties. In this case it might be of interest to quantify the population-average  $\theta$  and the subject-specific effects  $u$  around the average  $\theta$  for the description of the subject-specific dynamics. Therefore, to quantify the contribution of each subject  $j = 1, \dots, J$  on the process's dynamics we extended the LLA formulation of Eq. (3.B.3) by introducing random effects  $u$  for the  $J$  distinct subjects on the parameter vector  $\theta$ , leading to the following mixed-effects [33] formulation

$$\Delta \mathbf{y} = \underbrace{\begin{bmatrix} \mathbf{M}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{M}_J \end{bmatrix}}_{\mathbf{M} \in \mathbb{R}^{n \times Jp}} \mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathbf{u} \sim \mathcal{N}_{Jp} \left( \underbrace{\mathbf{1}_J \otimes \boldsymbol{\theta}, \mathbf{I}_J \otimes}_{\boldsymbol{\theta}_u} \underbrace{\begin{bmatrix} \tau_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \tau_p^2 \end{bmatrix}}_{\Delta_u} \right), \quad (3.D.1)$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2)),$$

where  $\mathbf{M}$  is the block-diagonal design matrix for the random effects  $\mathbf{u}$  centered in  $\boldsymbol{\theta}$ , and each block  $\mathbf{M}_j$  is subject-specific. As in the case of the null model of Eq. (3.B.3), to explain additional noise of the data and to avoid singularity of the stochastic covariance matrix  $\mathbf{W}(\boldsymbol{\theta})$  we added to its diagonal a small unknown quantity  $\sigma^2$  which we infer from the data. In order to infer the maximum likelihood estimator  $\hat{\boldsymbol{\psi}}$  for

$$\boldsymbol{\psi} = \left( \boldsymbol{\theta}, \sigma^2, \tau_1^2, \dots, \tau_p^2 \right), \quad (3.D.2)$$

we developed an efficient expectation-maximization E-M algorithm where  $\Delta \mathbf{y}$  and  $\mathbf{u}$  take the roles of the observed and latent states respectively. Under this framework

$$\begin{aligned} p(\mathbf{u} | \Delta \mathbf{y}) &\propto_{\mathbf{u}} p(\Delta \mathbf{y} | \mathbf{u}) p(\mathbf{u}) \\ &\propto_{\mathbf{u}} \exp \left( -\frac{1}{2} \mathbf{u}' (\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \mathbf{M} + \Delta_u^{-1}) \mathbf{u} \right. \\ &\quad \left. + \mathbf{u}' (\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \Delta_u^{-1} \boldsymbol{\theta}_u) \right), \end{aligned} \quad (3.D.3)$$

and therefore

$$\mathbf{u} | \Delta \mathbf{y} \sim \mathcal{N}_{Jp}(E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}}[\mathbf{u}], V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}}(\mathbf{u})), \quad (3.D.4)$$

where

$$\begin{aligned} E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}}[\mathbf{u}] &= V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}}(\mathbf{u}) (\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \Delta_u^{-1} \boldsymbol{\theta}_u), \\ V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}}(\mathbf{u}) &= (\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \mathbf{M} + \Delta_u^{-1})^{-1}. \end{aligned} \quad (3.D.5)$$

Also, the joint log-likelihood of  $\Delta \mathbf{y}$  and  $\mathbf{u}$  is given by

$$\begin{aligned} l(\Delta \mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) &\propto_{\boldsymbol{\psi}} l(\Delta \mathbf{y} | \mathbf{u}; \boldsymbol{\psi}) + l(\mathbf{u}; \boldsymbol{\psi}) \\ &\propto_{\boldsymbol{\psi}} -\frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2)| - \frac{1}{2} (\Delta \mathbf{y} - \mathbf{M}\mathbf{u})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) (\Delta \mathbf{y} - \mathbf{M}\mathbf{u}) + \\ &\quad -\frac{1}{2} \log |\Delta \mathbf{u}| - \frac{1}{2} (\mathbf{u} - \boldsymbol{\theta}_u)' \Delta \mathbf{u}^{-1} (\mathbf{u} - \boldsymbol{\theta}_u), \end{aligned} \quad (3.D.6)$$

3

which only depends on  $\mathbf{u}$  linearly via its first two-order conditional moments of Eq. (3.D.5). Therefore, it follows for the E-step function that

$$\begin{aligned} Q(\boldsymbol{\psi} | \boldsymbol{\psi}^*) &= E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*} [l(\Delta \mathbf{y}, \mathbf{u}; \boldsymbol{\psi})] = -\frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2)| \\ &\quad -\frac{1}{2} \{ \Delta \mathbf{y}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} - 2 E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*} [\mathbf{u}]' \mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \\ &\quad + \text{tr}(\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \mathbf{M} [V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u}) + E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}] E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}']]) \} + \\ &\quad -\frac{1}{2} \log |\Delta \mathbf{u}| - \frac{1}{2} \text{tr}(\Delta \mathbf{u}^{-1} [V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u}) + E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}] E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}']]) + \\ &\quad + E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \Delta \mathbf{u}^{-1} \boldsymbol{\theta}_u - \frac{1}{2} \boldsymbol{\theta}_u' \Delta \mathbf{u}^{-1} \boldsymbol{\theta}_u. \end{aligned} \quad (3.D.7)$$

The gradient of  $Q(\boldsymbol{\psi} | \boldsymbol{\psi}^*)$  is defined by the following partial derivatives

$$\begin{aligned} \frac{\partial}{\partial \theta_j} Q(\boldsymbol{\psi} | \boldsymbol{\psi}^*) &= -\frac{1}{2} \text{tr} \left( \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \frac{\partial}{\partial \theta_j} \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2) \right) + \\ &\quad -\frac{1}{2} \left\{ -\Delta \mathbf{y}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \frac{\partial}{\partial \theta_j} \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \right. \\ &\quad + 2 E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \frac{\partial}{\partial \theta_j} \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \\ &\quad + \text{tr} \left( -\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \frac{\partial}{\partial \theta_j} \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \right) \mathbf{M} \left[ V_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u}) \right. \\ &\quad \left. + E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}] E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}'] \right] \left. \right\} + \\ &\quad + E_{\mathbf{u} | \Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \Delta \mathbf{u}^{-1} \frac{\partial}{\partial \theta_j} \boldsymbol{\theta}_u - \boldsymbol{\theta}_u' \Delta \mathbf{u}^{-1} \frac{\partial}{\partial \theta_j} \boldsymbol{\theta}_u, \end{aligned} \quad (3.D.8)$$

$$\begin{aligned}
\frac{\partial}{\partial \tau_j} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*) &= -\frac{1}{2} \text{tr} \left( \Delta_u^{-1} \frac{\partial}{\partial \tau_j} \Delta_u^{-1} \right) + \\
-\frac{1}{2} \text{tr} \left( \Delta_u^{-1} \frac{\partial}{\partial \tau_j} \Delta_u^{-1} \Delta_u^{-1} \left[ V_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u}) + E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}] E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \right] \right) &+ \quad (3.D.9) \\
-E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \Delta_u^{-1} \frac{\partial}{\partial \tau_j} \Delta_u^{-1} \Delta_u^{-1} \boldsymbol{\theta}_u + \frac{1}{2} \boldsymbol{\theta}'_u \Delta_u^{-1} \frac{\partial}{\partial \tau_j} \Delta_u^{-1} \Delta_u^{-1} \boldsymbol{\theta}_u,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*) &= -\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2)) \\
-\frac{1}{2} \left\{ -\Delta \mathbf{y}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + \right. & \\
+ 2 E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \Delta \mathbf{y} + & \quad (3.D.10) \\
+ \text{tr} \left( -\mathbf{M}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}, \sigma^2) \mathbf{M} \left[ V_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u}) \right. \right. & \\
\left. \left. + E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}] E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]' \right] \right) \left. \right\}. &
\end{aligned}$$

In the E-M algorithm we iteratively update the E-function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  using the current estimate  $\boldsymbol{\psi}^*$  of  $\boldsymbol{\psi}$  and then we minimize the  $-Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  w.r.t.  $\boldsymbol{\psi}$ . The E-M algorithm is run until a convergence criterion is met, that is when the relative errors of both the E-step function  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  and the vector parameter  $\boldsymbol{\psi}$  are lower than a predefined tolerance. Once we get the E-M estimate  $\hat{\boldsymbol{\psi}}$  for the parameters we evaluate the goodness-of-fit of the mixed-model according to the conditional Akaike Information Criterion [34]. As every E-M algorithm, the choice of the starting point  $\boldsymbol{\psi}_s$  is very important from a computational point of view. We chose as a starting point  $\boldsymbol{\psi}_s = (\hat{\boldsymbol{\theta}}_{ML}, \hat{\sigma}_{ML}^2, \tau_1^2 = 0, \dots, \tau_p^2 = 0)$  where  $(\hat{\boldsymbol{\theta}}_{ML}, \hat{\sigma}_{ML}^2)$  is the optimum found in the fixed-effects LLA formulation of Eq. (3.B.3). This is a reasonable choice since we want to quantify how the dynamics  $E_{\mathbf{u}|\Delta \mathbf{y}; \hat{\boldsymbol{\psi}}}[\mathbf{u}]_j$  of each subject  $j$  departs from the average dynamics  $\hat{\boldsymbol{\theta}}_{ML}$ . The E-M pseudocode is given in Algorithm 5.

**Input:**  $\boldsymbol{\psi}^* = (\hat{\boldsymbol{\theta}}_{ML}, \hat{\sigma}_{ML}^2, \tau_1^2 = 0, \dots, \tau_p^2 = 0)$ ,  $\mathbf{M}$ ,  $\Delta \mathbf{y}$

**Output:**  $\hat{\boldsymbol{\psi}}_{EM}$

chose a small tolerance  $tol$  and set  $\epsilon = +\infty$ ;

**while**  $\epsilon > tol$  **do**

    update  $E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]$  and  $V_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}(\mathbf{u})$  as defined in Eq. (3.D.5);

    set to zero the negative elements of  $E_{\mathbf{u}|\Delta \mathbf{y}; \boldsymbol{\psi}^*}[\mathbf{u}]$ ;

    update  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  and  $\nabla_{\boldsymbol{\psi}^*} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$  according to  
    Eqs. (3.D.8)-(3.D.10);

    set  $\boldsymbol{\psi}_{old} \leftarrow \boldsymbol{\psi}^*$ ;

    update  $\boldsymbol{\psi}^* \leftarrow \underset{\boldsymbol{\psi} \geq \mathbf{0}}{\operatorname{argmin}} - Q(\boldsymbol{\psi}|\boldsymbol{\psi}^*)$ ;

    update  $\epsilon = |Q(\boldsymbol{\psi}_{old}|\boldsymbol{\psi}_{old}) - Q(\boldsymbol{\psi}^*|\boldsymbol{\psi}^*)|$ ;

**end**

$\hat{\boldsymbol{\psi}}_{EM} = \boldsymbol{\psi}^*$

**Algorithm 5:** E-M inference algorithm for the mixed-effects model.

## REFERENCES

- [1] L. Del Core, M. A. Grzegorzczuk, and E. C. Wit, “Stochastic inference of clonal dominance in gene therapy studies,” *bioRxiv*, 2022.
- [2] T. Friedmann and R. Roblin, “Gene therapy for human genetic disease?,” *Science*, vol. 175, no. 4025, pp. 949–955, 1972.
- [3] D. Bryder, D. J. Rossi, and I. L. Weissman, “Hematopoietic stem cells: the paradigmatic tissue-specific stem cell,” *The American journal of pathology*, vol. 169, no. 2, pp. 338–346, 2006.
- [4] O. S. Kustikova, A. Wahlers, K. Kühlcke, B. Stähle, A. R. Zander, C. Baum, and B. Fehse, “Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population,” *Blood*, vol. 102, no. 12, pp. 3934–3937, 2003.
- [5] B. Fehse, O. Kustikova, M. Bubenheim, and C. Baum, “Pois (s) on—it’s a question of dose. . .,” *Gene therapy*, vol. 11, no. 11, pp. 879–881, 2004.
- [6] C. Baum, J. Düllmann, Z. Li, B. Fehse, J. Meyer, D. A. Williams, and C. Von Kalle, “Side effects of retroviral gene transfer into hematopoietic stem cells,” *Blood, The Journal of the American Society of Hematology*, vol. 101, no. 6, pp. 2099–2113, 2003.
- [7] U. Modlich, O. S. Kustikova, M. Schmidt, C. Rudolph, J. Meyer, Z. Li, K. Kamino, N. Von Neuhoff, B. Schlegelberger, K. Kuehlcke, *et al.*, “Leukemias following retroviral transfer of multidrug resistance 1 (mdr1) are driven by combinatorial insertional mutagenesis,” *Blood*, vol. 105, no. 11, pp. 4235–4246, 2005.
- [8] C. Baum, O. Kustikova, U. Modlich, Z. Li, and B. Fehse, “Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors,” *Human gene therapy*, vol. 17, no. 3, pp. 253–263, 2006.
- [9] S. N. Catlin, P. Gutterp, and J. L. Abkowitz, “The kinetics of clonal dominance in myeloproliferative disorders,” *Blood*, vol. 106, no. 8, pp. 2688–2692, 2005.
- [10] I. Roeder, M. Horn, I. Glauche, A. Hochhaus, M. C. Mueller, and M. Loeffler, “Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications,” *Nature medicine*, vol. 12, no. 10, pp. 1181–1184, 2006.
- [11] C. E. Müller-Sieburg, R. H. Cho, M. Thoman, B. Adkins, and H. B. Sieburg, “Deterministic regulation of hematopoietic stem cell self-renewal and differentiation,” *Blood, The Journal of the American Society of Hematology*, vol. 100, no. 4, pp. 1302–1309, 2002.



- [12] 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, no. 2, pp. 609–616, 2005.
- [13] H. B. Sieburg, R. H. Cho, B. Dykstra, N. Uchida, C. J. Eaves, and C. E. Muller-Sieburg, "The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets," *Blood*, vol. 107, no. 6, pp. 2311–2316, 2006.
- [14] M. Loeffler, A. Birke, D. Winton, and C. Potten, "Somatic mutation, monoclonality and stochastic models of stem cell organization in the intestinal crypt," *Journal of theoretical biology*, vol. 160, no. 4, pp. 471–491, 1993.
- [15] M. Loeffler, T. Bratke, U. Paulus, Y. Li, and C. Potten, "Clonality and life cycles of intestinal crypts explained by a state dependent stochastic model of epithelial stem cell organization," *Journal of Theoretical Biology*, vol. 186, no. 1, pp. 41–54, 1997.
- [16] M. Loeffler and I. Roeder, "Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models—a conceptual approach," *Cells Tissues Organs*, vol. 171, no. 1, pp. 8–26, 2002.
- [17] F. A. Meineke, C. S. Potten, and M. Loeffler, "Cell migration and organization in the intestinal crypt using a lattice-free model," *Cell proliferation*, vol. 34, no. 4, pp. 253–266, 2001.
- [18] I. Roeder, K. Braesel, R. Lorenz, and M. Loeffler, "Stem cell fate analysis revisited: interpretation of individual clone dynamics in the light of a new paradigm of stem cell organization," *Journal of biomedicine and biotechnology*, vol. 2007, 2007.
- [19] D. Winton, M. Blount, and B. Ponder, "A clonal marker induced by mutation in mouse intestinal epithelium," *Nature*, vol. 333, no. 6172, pp. 463–466, 1988.
- [20] H.-S. Park, R. A. Goodlad, and N. A. Wright, "Crypt fission in the small intestine and colon. a mechanism for the emergence of g6pd locus-mutated crypts after treatment with mutagens," *The American journal of pathology*, vol. 147, no. 5, p. 1416, 1995.
- [21] M. Bjerknes and H. Cheng, "Modulation of specific intestinal epithelial progenitors by enteric neurons," *Proceedings of the National Academy of Sciences*, vol. 98, no. 22, pp. 12497–12502, 2001.
- [22] C. S. Potten, C. Booth, and D. M. Pritchard, "The intestinal epithelial stem cell: the mucosal governor," *International journal of experimental pathology*, vol. 78, no. 4, pp. 219–243, 1997.
- [23] L. Biasco, D. Pellin, S. Scala, F. Dionisio, L. Basso-Ricci, L. Leonardelli, S. Scaramuzza, C. Baricordi, F. Ferrua, M. Cicalese, S. Giannelli, V. Neduva, D. Dow, M. Schmidt, C. Von Kalle, M. Roncarolo, F. Ciceri, P. Vicard, E. Wit, C. Di Serio, L. Naldini, and

- A. Aiuti, "In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases," *Cell Stem Cell*, vol. 19, no. 1, pp. 107–119, 2016.
- [24] C. Wu, B. Li, R. Lu, S. Koelle, Y. Yang, A. Jares, A. Krouse, M. Metzger, F. Liang, K. Loré, C. Wu, R. Donahue, I. Chen, I. Weissman, and C. Dunbar, "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.
- [25] F. Mazurier, O. I. Gan, J. L. McKenzie, M. Doedens, and J. E. Dick, "Lentivector-mediated clonal tracking reveals intrinsic heterogeneity in the human hematopoietic stem cell compartment and culture-induced stem cell impairment," *Blood*, vol. 103, no. 2, pp. 545–552, 2004.
- [26] L. Biasco, M. Rothe, J. W. Schott, and A. Schambach, "Integrating vectors for gene therapy and clonal tracking of engineered hematopoiesis," *Hematology/Oncology Clinics*, vol. 31, pp. 737–752, 2020/04/07 2017.
- [27] D. Pellin, *Stochastic modelling of dynamical systems in biology [PhD thesis]*. PhD thesis, University of Groningen, 2017.
- [28] 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.
- [29] N. Bailey, *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley Classics Library, Wiley, 1990.
- [30] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2011.
- [31] 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.
- [32] L. Del Core, D. Cesana, P. Gallina, Y. N. S. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. A. Grzegorzcyk, "Normalization of clonal diversity in gene therapy studies using shape constrained splines," *Scientific Reports*, vol. 12, p. 3836, Mar. 2022.
- [33] A. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science, CRC Press, 2018.
- [34] F. Vaida and S. Blanchard, "Conditional Akaike Information for mixed-effects models," *Biometrika*, vol. 92, no. 2, pp. 351–370, 2005.

- [35] 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.
- [36] S. Müller, J. L. Scealy, and A. H. Welsh, “Model Selection in Linear Mixed Models,” *Statistical Science*, vol. 28, no. 2, pp. 135 – 167, 2013.
- [37] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [38] 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.
- [39] C. Wu, D. A. Espinoza, S. J. Koelle, D. Yang, L. Truitt, H. Schlums, B. A. Lafont, J. K. Davidson-Moncada, R. Lu, A. Kaur, *et al.*, “Clonal expansion and compartmentalized maintenance of rhesus macaque nk cell subsets,” *Science immunology*, vol. 3, no. 29, p. eaat9781, 2018.
- [40] C. Wu, R. D. Mortlock, T. Shin, S. Cordes, X. Fan, J. Brenchley, D. A. Allan, S. G. Hong, and C. E. Dunbar, “Tissue-resident clonal expansions of rhesus macaque nk cells,” *Blood*, vol. 138, p. 998, 2021.
- [41] G. N. Mil'shtejn, “Approximate integration of stochastic differential equations,” *Theory of Probability & Its Applications*, vol. 19, no. 3, pp. 557–562, 1975.
- [42] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.
- [43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.
- [44] W. K. Hastings, “Monte carlo sampling methods using markov chains and their applications,” 1970.
- [45] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [46] P. Ashcroft, M. G. Manz, and S. Bonhoeffer, “Clonal dominance and transplantation dynamics in hematopoietic stem cell compartments,” *PLOS Computational Biology*, vol. 13, pp. 1–20, 10 2017.
- [47] R. K. Pedersen, M. Andersen, T. Stiehl, and J. T. Ottesen, “Mathematical modelling of the hematopoietic stem cell-niche system: Clonal dominance based on stem cell fitness.,” *Journal of Theoretical Biology*, vol. 518, p. 110620, 2021.

- [48] D. Liu, T. Lu, X.-F. Niu, and H. Wu, "Mixed-effects state-space models for analysis of longitudinal dynamic systems," *Biometrics*, vol. 67, no. 2, pp. 476–485, 2011.
- [49] M. A. Nowak and C. R. M. Bangham, "Population dynamics of immune responses to persistent viruses," *Science*, vol. 272, no. 5258, pp. 74–79, 1996.
- [50] B. Ribba, N. Holford, P. Magni, I. Trocóniz, I. Gueorguieva, P. Girard, C. Sarr, M. El-ishmereni, C. Kloft, and L. Friberg, "A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 3, no. 5, p. 113, 2014.
- [51] T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling," *BMC Bioinformatics*, vol. 8, pp. 1–22, 2007.
- [52] N. GOKGOZ and H. ÖKTEM, "Modeling of tumor-immune system interaction with stochastic hybrid systems with memory: A piecewise linear approach," *Advances in the Theory of Nonlinear Analysis and its Application*, vol. 5, no. 1, pp. 25–38, 2021.
- [53] G.-W. Weber, O. Ugur, P. Taylan, and A. Tezel, "On optimization, dynamics and uncertainty: A tutorial for gene-environment networks," *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2494–2513, 2009. *Networks in Computational Biology*.
- [54] E. Savku and G.-W. Weber, "Stochastic differential games for optimal investment problems in a markov regime-switching jump-diffusion market," *Annals of Operations Research*, vol. 312, no. 2, pp. 1171–1196, 2022.
- [55] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," nov 2012. Version 20121115.



# 4

## A NORMALIZED MEASURE OF CLONAL DIVERSITY

---

Parts of this chapter have been published in “Normalization of clonal diversity in gene therapy studies using shape constrained splines” [1].

## ABSTRACT

*Viral vectors are used to insert genetic material into semirandom genomic positions of hematopoietic stem cells which, after reinfusion into patients, regenerate the entire hematopoietic system. Hematopoietic cells originating from genetically modified stem cells will harbor insertions in specific genomic positions called integration sites, which represent unique genetic marks of clonal identity. Therefore, the analysis of vector integration sites present in the genomic DNA of circulating cells allows to determine the number of clones in the blood ecosystem. Shannon diversity index is adopted to evaluate the heterogeneity of the transduced population of gene corrected cells. However, this measure can be affected by several technical variables such as the DNA amount used and the sequencing depth of the library analyzed and therefore the comparison across samples may be affected by these confounding factors. We developed an advanced spline-regression approach that leverages on confounding effects to provide a normalized entropy index. Our proposed method was first validated and compared with two state of the art approaches in a specifically designed in vitro assay. Subsequently our approach allowed to observe the expected impact of vector genotoxicity on entropy level decay in an in vivo model of hematopoietic stem cell gene therapy based on tumor prone mice.*

### 4.1. INTRODUCTION

Gamma retroviral and Lentiviral Vectors (LVs) are widely adopted in Gene Therapy (GT) thanks to their ability to insert therapeutic transgenes in the host cell genome of hematopoietic stem/progenitor cell (HSPC). After transplantation into the patient the HSPCs reconstitute the entire hematopoietic system and correct the genetic defect. Therefore, vector integration ensures the maintenance of gene correction during self-renewal of HSPCs as well as its transmission to their cell progeny[2]. These vectors

integrate semi-randomly within the human genome, and then each transduced cell harbours a vector integration in a distinct genomic position (integration site, IS) that can be adopted as a genetic mark to distinguish each engrafted clone. The retrieval of IS from transduced cells can be done by using PCR protocols that allow to specifically amplify the vector/genome junctions from their genomic DNA. Sequencing and mapping on the target cell reference genome allow to identify IS that can be univocally used for clonal identity. Therefore, the analysis of vector IS from DNA of blood cells harvested at specific time points after transplant from GT patients provide information on number of hundreds to thousands of clones present in circulation and their relative abundance. For this reason, IS studies are required for safety and long-term efficacy assessment in preclinical and clinical studies [3–9].

The Shannon entropy index, a well-established measure of species diversity in ecology[10], has become one of the most widely used measure of IS diversity in HSC-GT applications [11]. This measure has been positively correlated to high levels of genetic modification and engraftment of genetically modified cells while low levels of entropy were associated to poor levels of genetic modification, or oligoclonality due to poor engraftment or even the appearance of highly dominant clones resulting from malignant transformation [12]. Indeed, the complexity of a given DNA sample is computed considering both the total number of different IS obtained and their relative abundance. Thus, highly polyclonal samples characterized by large number of IS whose abundance is evenly distributed will have a higher Shannon diversity index than oligoclonal samples with a relatively smaller number of IS and/or characterised by the presence of highly dominant clones. However, the Shannon diversity index does not consider variations in sample size (amounts of DNA analyzed) or the efficiency in species retrieval (different PCR protocols for IS retrieval, sequencing platforms and sequencing depth) of complex ecosystems, such as the population of vector integrations sites in the genome [13].

Thus, while Shannon diversity index provides an objective measure of the clonal complexity of any given IS sample, these confounding factors should be taken in account when the clonal complexity of different samples is compared. Since longitudinal studies of GT patients for IS monitoring could require the analysis of several samples collected under



heterogeneous technical conditions, a method aimed at removing confounding effects in diversity index is needed. To remove confounding factors in the estimations of ecosystem diversity, several methods have been applied. Random subsampling without replacement, called “rarefying” [14], is among the most popular methods for the normalization of species count data in ecology as well as for next generation sequencing (NGS) data in microbiology. Given a predefined sequence depth (total count, SD), a subsample from each library is generated by randomly picking reads without replacement, until the selected total number of counts is reached. Although rarefying has become the state-of-the-art tool for NGS data analysis [15], some limitations have been recognized. Indeed, [16] demonstrated that rarefying is statistically inadmissible and should be avoided. Furthermore, in [17] it was highlighted that estimates of species diversity in sites/habitats at local scale, namely the  $\alpha$ -diversity [18], for rarefied microbiome count data may be strongly biased. This is mainly due to the rare species which may be over- (or under-) represented in the samples that have been normalized to a smaller depth by rarefaction. An alternative normalization to rarefying is scaling, which adjusts the size of all samples by scaling their counts to the same total amount. Scaling preserves the relative frequencies of the species and keeps the species richness unchanged. Therefore, simple scaling does not remove the effect provided from the library depth neither on species richness nor on species diversity. [19] introduces a novel normalization method for species count data called scaling with ranked subsampling (SRS) and the authors demonstrate its suitability for the analysis of microbial communities.

The growing number of normalization and scaling approaches highlights that a robust method has not been developed yet. In this work we show that all proposed methods have limitations. In particular they miss of a precise quantification of the effect of each confounding variable on the Shannon entropy. Furthermore, we also show that the rescaled Shannon entropy index obtained by either rarefying or scaling with ranked subsampling still suffers from the effect of the confounders. We propose a spline-regression approach aimed to explain and remove those effects from the diversity indexes. The effect of the confounders is measured using a B-spline term whose shape is restricted according to a biological-sustained hypothesis. We test our framework by analysing a novel in-vitro

dataset properly designed to simulate the same clonality state under different combinations of technical conditions. We also compare our method with the previously proposed methods from the literature in terms of efficiency according to hypothesis testing. That is, we consider a rescaling method to be more efficient if there is more evidence for the corresponding rescaled measure being independent from the effect of the candidate confounders. Finally, our rescaling approach allowed to unmask the expected impact of vector genotoxicity on entropy level decay in an *in vivo* model of hematopoietic stem cell gene therapy based on tumor prone mice [20, 21].

## 4.2. POTENTIAL ARTEFACTS IN CLONAL TRACKING

There are several high-throughput systems capable to quantitatively track cell types repopulation from an individual stem cell after a gene therapy treatment [22–24]. Tracking cells by random labeling is one of the most sensitive systems [25]. In HSC-GT applications, haematopoietic stem cells (HSCs) are sorted from the bone marrow of the treated subject and uniquely labeled by the random insertion of a viral vector inside its genome. Each label, called clone, or integration site (IS), is defined as the genomic coordinates where the viral vector integrates. After transplantation, all the progeny deriving through cell differentiation inherits the original labels. During follow-up, the labels are collected from tissues and blood samples using Next Generation Sequencing (NGS) [26–29]. NGS is a recent approach for DNA and RNA sequencing, which consists of a complex interplay of chemistry, hardware, optical sensors and software [30–33]. In gene therapy applications NGS does allow identifying, quantifying and tracking clones arising from the same HSC ancestor. Over the past decades, clonal tracking has proven to be a cutting-edge analysis capable to unveil population dynamics and hierarchical relationships *in vivo* [34–37]. Clonal diversity, measuring how many distinct clones are collected and how they distribute, can address some of these aspects. Loosely speaking, the less distinct clones the lower the clonal diversity and in turn the less the system is being repopulated in that particular cell compartment. Furthermore, under the same number of different clones collected, the more their distribution is far from the uniform, the lower the clonal

diversity and the more the dominance of few clones, thus suggesting the possible occurrence of an adverse event. The Shannon entropy index [38], a well-established measure of population diversity in ecology studies [10], nowadays is hugely used as a proxy of clonal diversity in gene therapy applications [11]. Following [38], the Shannon entropy index is defined as

$$h(\mathbf{x}) = - \sum_{i=1}^n P(x_i) \log P(x_i), \quad (4.2.1)$$

where  $\mathbf{x}$  has possible realisations  $x_1, \dots, x_n$  which occur with probabilities  $P(x_1), \dots, P(x_n)$ . Therefore, Shannon entropy is a special case (up to a change in sign and a multiplicative factor  $1/n$ ) of the Kullback-Leibler divergence [39]

$$D_{KL}(P\|Q) = \sum_{i=1}^n P(x_i) \log \frac{P(x_i)}{Q(x_i)}, \quad (4.2.2)$$

when the reference  $Q$  is the uniform distribution on  $x_1, \dots, x_n$ .

Potential limitations of entropy-based measures in gene therapy applications are related to the heterogeneous nature of NGS data [40–44]. Indeed due to sampling and technical conditions, such as the amount of the host DNA being sequenced and the PCR being adopted, the number of reads obtained per library can span orders of magnitudes [13] which may affect the cellular counts and in turn their Shannon entropy. These differences in magnitude of library size/depth mainly depends on unequal pooling of PCR products before sequencing. In order to pool PCR products from individual samples in equimolar amounts [45], DNA concentrations are commonly determined by ultraviolet-visible (UV) or fluorescence spectroscopy, real-time PCR or digital PCR [46]. Although these methods are very effective [47], an identical library size across samples is difficult to achieve. Nonetheless, if we define the multiplicity of infection (MOI) as the average ratio between the number of virus particles and the number of target cells present in a defined space, then the actual number of viruses that will integrate on any given cell can be described by a stochastic process, such as some cells may absorb more than one infectious agent while others may not absorb any of them. Typically, the probability  $P(n|m)$  that a cell will absorb  $n$  virus particles when inoculated with an MOI of  $m$  can

be modelled as a Poisson variable with rate  $m$ ,

$$P(n|m) = \frac{m^n e^{-m}}{n!}. \quad (4.2.3)$$

Therefore, by definition, it is possible to increase the expected number of vector copies per cell (VCN) by properly tuning the MOI in the design of the experiment/treatment. As a result, the VCN may affect the number of IS collected and in turn the Shannon entropy.

In Figure 4.2.1 we show the behaviour of the Shannon entropy index as a function of the DNA amount, the VCN and the sequencing depth (SD) in the case of an in-vitro assay described in section 4.4.1. Figure 4.2.1 suggests that the Shannon entropy index strongly depends on the quantitative confounders, until it reaches a steady-state. These features motivate us to use shape constrained splines (SCS) in order to model the effect of the candidate confounders on the entropy measurements.

## 4.3. METHODS: SHAPE CONSTRAINED SPLINES

### 4.3.1. DEFINITION OF THE MODEL

Shape-constrained splines (SCS) for fitting, smoothing and interpolation have been explored and proposed in various works, such as [48–53]. In this work we follow the cone-projection approach [52, 53]. We model the logarithmic observed entropies  $h_i$ 's, for  $i = 1, \dots, n$ , as a function of a SCS-bases  $\mathbf{C}_i^k$  for every potential confounder  $k = 1, \dots, K$  plus a term  $\mathbf{F}_i^j$  for any other additional feature of interest  $j = 1, \dots, J$ , so that

$$\log(h_i) = \beta_0 + \sum_{k=1}^K \mathbf{C}_i^k \boldsymbol{\beta}_c^k + \sum_{j=1}^J \mathbf{F}_i^j \boldsymbol{\beta}_f^j + \varepsilon_i, \quad i = 1, \dots, n, \quad (4.3.1)$$

where  $\beta_0$  is the intercept,  $\mathbf{C}_i^k$  is the basis of a quadratic spline for the  $k$ -th confounder used for observation  $i$  for which we assume a saturation state at the right boundary knot and a monotone increasing concave shape. For our applications, the boundary knots of a spline basis associated to a variable  $x$  are defined as the minimal and maximal value of  $x$ . The term  $\mathbf{F}_i^j$  corresponds to the  $i$ -th observation of a basis describing the  $j$ -th additional component, such as the time or the cell type. The corresponding

parameter vectors are  $\beta_c^k$  and  $\beta_f^j$  respectively. Finally we assume for the noise variable:

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n. \quad (4.3.2)$$

Therefore, the model can be compactly written as

$$\underbrace{\log \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}}_{\log(\mathbf{h})} = \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{1}} \beta_0 + \underbrace{\begin{bmatrix} C_1^1 & \dots & C_1^K \\ \vdots & & \vdots \\ C_n^1 & \dots & C_n^K \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \beta_c^1 \\ \vdots \\ \beta_c^K \end{bmatrix}}_{\beta_c} + \underbrace{\begin{bmatrix} F_1^1 & \dots & F_1^J \\ \vdots & & \vdots \\ F_n^1 & \dots & F_n^J \end{bmatrix}}_{\mathbf{F}} \underbrace{\begin{bmatrix} \beta_f^1 \\ \vdots \\ \beta_f^J \end{bmatrix}}_{\beta_f} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}, \quad (4.3.3)$$

4

where the number of features  $K$  and  $J$  depend on the project and/or the specific research questions that must be addressed. Quadratic splines are characterized by the discontinuity of the second-order derivative, which makes their treatments harder than cubic splines. This applies already to unconstrained spline fitting and interpolation. In particular, the definition of quadratic penalised/smoothing splines is not straightforward. Therefore, in general, cubic splines should be preferred over quadratic splines. Despite this, in our in-vitro assay (VA) application only three distinct values of both the DNA amount and the vector copy number (VCN) are available. Therefore, in order to get a good trade-off between bias and variance as well as in order to obtain a full-rank design matrix, we chose quadratic splines with one interior knot for both the DNA amount and the vector copy number (VCN), and a quadratic spline with two interior knots for the sequencing depth. In Section 4.B we compare the fits of quadratic and cubic splines. For consistency, we also use quadratic splines in the mice study on genotoxicity, where our goal is to model the relationship between the entropy and several confounders; cf. Section 4.4.2.

### 4.3.2. SHAPE-CONSTRAINED SPLINES NORMALIZATION

For simplicity, we set  $K = 1$  and  $J = 0$  in Eq. (4.3.3), namely we consider only one confounder and no additional factors of interest. The general case can be obtained straightforwardly. Here we follow [54] and we represent

an  $(r + 1)$ -th order B-spline as

$$m(x) = \sum_{j=1}^q \beta_j B_j^r(x) \quad (4.3.4)$$

where, for  $j = 1, \dots, q$ , the bases are iteratively computed as

$$B_j^r(x) = \frac{x - k_j}{k_{j+r+1} - k_j} B_j^{r-1}(x) + \frac{k_{j+r+2} - x}{k_{j+r+2} - k_{j+1}} B_{j+1}^{r-1}(x), \quad (4.3.5)$$

$$B_j^{-1}(x) = \begin{cases} 1, & k_j \leq x \leq k_{j+1} \\ 0, & \text{otherwise} \end{cases}, \quad (4.3.6)$$

for a given sequence of evenly spaced knots  $\xi_1 \leq \xi_2 \leq \dots \leq \xi_{q+r+2}$ , where  $q$  is the number of basis functions and  $\beta_j$ 's are the corresponding coefficients. The first order derivative of Eq. (4.3.4) can be written as

$$m'(x) = \frac{1}{\delta} \sum_{j=2}^q B_j^{r-1}(x) (\beta_j - \beta_{j-1}), \quad (4.3.7)$$

where  $\delta$  is the distance between two adjacent knots. Since all B-spline basis functions are nonnegative by definition, a sufficient condition for  $m'(x) \geq 0$ , and in turn for the monotone-increasing shape of  $m(x)$ , is

$$\beta_j - \beta_{j-1} \geq 0 \quad j = 2, \dots, q. \quad (4.3.8)$$

Furthermore, the second order derivative of Eq. (4.3.4) can be written as

$$m''(x) = \frac{1}{\delta^2} \sum_{j=3}^q B_j^{r-2}(x) (\beta_j - 2\beta_{j-1} + \beta_{j-2}). \quad (4.3.9)$$

Then a sufficient condition for  $m''(x) \leq 0$  and in turn for the concavity of the spline in Eq. (4.3.4) is

$$\beta_j - 2\beta_{j-1} + \beta_{j-2} \leq 0, \quad j = 3, \dots, q. \quad (4.3.10)$$

The monotonicity and concavity constraints can be written respectively as

$$\mathbf{C}_1 \boldsymbol{\beta} \geq 0, \quad \mathbf{C}_2 \boldsymbol{\beta} \geq 0, \quad \boldsymbol{\beta} = [\beta_1 \cdots \beta_q]'. \quad (4.3.11)$$

where

$$\mathbf{C}_1 = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(q-1) \times q}, \quad \mathbf{C}_2 = \begin{bmatrix} -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(q-2) \times q}. \quad (4.3.12)$$

If both monotonicity and concavity constraints must be satisfied, the first  $q-2$  constraints/rows of  $\mathbf{C}_1$  are redundant, as stated by the following Lemma.

**Lemma 4.3.1.** *If  $\beta_j - 2\beta_{j-1} + \beta_{j-2} \leq 0 \quad \forall j = 3, \dots, q$  and  $\beta_q - \beta_{q-1} \geq 0$ , then  $\beta_j - \beta_{j-1} \geq 0 \quad \forall j = 2, \dots, q-1$ .*

4

*Proof.*  $\boxed{j = q-1}$ :  $\beta_q - 2\beta_{q-1} + \beta_{q-2} \leq 0$  and  $-\beta_q + \beta_{q-1} \leq 0$  hold, which together imply  $\beta_{q-1} - \beta_{q-2} \geq 0$ .

$\boxed{j = k+1 \Rightarrow j = k}$ :  $\beta_{k+1} - 2\beta_k + \beta_{k-1} \leq 0$  and  $-\beta_{k+1} + \beta_k \leq 0$  hold, which together imply  $\beta_k - \beta_{k-1} \geq 0$ . □

Therefore by Lemma 4.3.1, if both constraints  $\mathbf{C}_1$  and  $\mathbf{C}_2$  are applied, the whole matrix of constraints reduces to

$$\mathbf{C} = \begin{bmatrix} & & & -1 & 1 \\ -1 & 2 & -1 & -1 & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix}. \quad (4.3.13)$$

Furthermore, we need to consider that sampling might be characterised by a sequencing saturation level due to technical limitations. In this case, the saturation level can be included by considering a steady-state/stationary-point at the right boundary knot  $\xi_{q+r+2}$ , namely by setting


$$\partial^1 \mathbf{B}(\xi_{q+r+2}) \boldsymbol{\beta} = \mathbf{0}, \quad (4.3.14)$$

where

$$\partial^1 \mathbf{B}(x) = \left( \partial^1 B_1^r(x), \dots, \partial^1 B_q^r(x) \right) \quad (4.3.15)$$

is the first derivative of the spline basis

$$\mathbf{B}(x) = \left( B_1^r(x), \dots, B_q^r(x) \right) \quad (4.3.16)$$

evaluated at  $x$ . Our  code implementation allows to switch between the presence and absence of the saturation level by the additional logical input parameter SATURATION. By default this parameter is set to TRUE, but it can be switched to FALSE if the user prefers not to implement a saturation level (or a steady state) w.r.t. a particular predictor variable. In our case of quadratic degree ( $r = 1$ ), the constraint in Eq. (4.3.14) reduces to

$$\beta_q = -\frac{\partial B_{q-1}(\xi_{q+r+2})}{\partial B_q(\xi_{q+r+2})} \beta_{q-1}, \quad (4.3.17)$$

which can be written compactly using the following transformation

$$\mathcal{A} : \mathbb{R}^{n_X \times q} \rightarrow \mathbb{R}^{n_X \times (q-1)}, \quad \mathbf{X} \mapsto \mathbf{X}\mathbf{A} \quad (4.3.18)$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & 0 & & 1 \\ 0 & \cdots & 0 & -\frac{\partial B[x_n, q-1]}{\partial B[x_n, q]} \end{bmatrix} \in \mathbb{R}^{q \times (q-1)}, \quad (4.3.19)$$

and  $n_X$  is the number of rows of  $\mathbf{X}$ .

### 4.3.3. INFERENCE PROCEDURE

Given  $n$  observations of one predictor  $\mathbf{x} = (x_1, \dots, x_n)$  and a response  $\mathbf{y} = (y_1, \dots, y_n)$ , the restricted least squares estimate  $\hat{\boldsymbol{\beta}}_{RLS}$  of  $\boldsymbol{\beta}$  subject to the constraints in Eqs. (4.3.11) and (4.3.17) can be obtained as

$$\hat{\boldsymbol{\beta}}_{RLS} = \underset{\boldsymbol{\beta} \in S}{\operatorname{argmin}} (\mathbf{y} - \mathbf{B}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{B}\boldsymbol{\beta}), \quad (4.3.20)$$

where

$$S = \left\{ \boldsymbol{\beta} \in \mathbb{R}^q \mid \boldsymbol{\beta} \geq 0, \quad \mathbf{C}\boldsymbol{\beta} \geq 0, \quad \beta_q = -\frac{\partial B_{q-1}(\xi_{q+r+2})}{\partial B_q(\xi_{q+r+2})} \beta_{q-1} \right\}. \quad (4.3.21)$$

Therefore, using Eqs. (4.3.18) - (4.3.19) we can include the linear equality constraint (4.3.17) inside the objective function, and the optimisation



problem from Eq. (4.3.20) reduces to

$$\hat{\boldsymbol{\beta}}_{RLS}^* = \underset{\boldsymbol{\beta}^* \in S^*}{\operatorname{argmin}} \left\{ \underbrace{-2\boldsymbol{\beta}^* \mathbf{B} \mathbf{A} \mathbf{y} + \boldsymbol{\beta}^* (\mathbf{B} \mathbf{A})' \mathbf{B} \mathbf{A} \boldsymbol{\beta}^*}_{f} \right\}, \quad (4.3.22)$$

subject to only linear inequality constraints, where

$$\boldsymbol{\beta}^* = (\beta_1, \dots, \beta_{q-1})', \quad S^* = \{\boldsymbol{\beta}^* \geq 0; \mathbf{C} \mathbf{A} \boldsymbol{\beta}^* \geq 0\}. \quad (4.3.23)$$

Since  $f$  is a quadratic function, we solve Eq. (4.3.22) using quadratic optimization. To this end we use the function `solve.QP()` from the R package `quadprog`. Once the restricted least squares estimate  $\hat{\boldsymbol{\beta}}_{RLS}^*$  is obtained, we follow the cone projection approach [53] and we define a point-wise confidence interval (CI) with  $1 - \alpha/2$  coverage for  $\mathbf{B} \mathbf{A} \hat{\boldsymbol{\beta}}_{RLS}^*$  as

$$\mathbf{b}'_p \mathbf{A} \hat{\boldsymbol{\beta}}^* \mp z_{\alpha/2} \sqrt{\hat{\sigma}_{RLS}^2 (\mathbf{b}'_p \mathbf{A})' \hat{\mathbf{G}} \mathbf{b}'_p \mathbf{A}}, \quad (4.3.24)$$

where  $\mathbf{b}_p = \mathbf{B}(x_p)$  is the B-spline basis  $\mathbf{B}(x)$  evaluated at the prediction point  $x_p$  and the variance is estimated as

$$\hat{\sigma}_{RLS}^2 = \frac{(\mathbf{y} - \mathbf{B} \mathbf{A} \hat{\boldsymbol{\beta}}_{RLS}^*)' (\mathbf{y} - \mathbf{B} \mathbf{A} \hat{\boldsymbol{\beta}}_{RLS}^*)}{n - 1.5d}, \quad (4.3.25)$$

where  $d$  is the dimension of the cone's face where the projection of  $\mathbf{y}$  onto

$$\mathcal{F} = \{\boldsymbol{\eta} \in \mathbb{R}^n \mid \boldsymbol{\eta} = \mathbf{B} \mathbf{A} \boldsymbol{\beta}, \quad \mathbf{C} \mathbf{A} \boldsymbol{\beta} \geq 0\} \quad (4.3.26)$$

lands on. The dimension  $d$  of  $\mathcal{F}$  can be computed using the Hinge algorithm implemented in the R package `coneproject` [55]. The matrix  $\hat{\mathbf{G}}$  is computed as the following weighted average

$$\hat{\mathbf{G}} = \sum_{\mathcal{J} \subseteq \{1, \dots, q-1\}} \hat{\mathbf{G}}^{(\mathcal{J})} \hat{p}_{\mathcal{J}}, \quad (4.3.27)$$

with the  $(q-1) \times (q-1)$  matrix  $\hat{\mathbf{G}}^{(\mathcal{J})}$  defined as

$$\begin{aligned} \hat{\mathbf{G}}_{k \in \mathcal{J}, l \in \mathcal{J}}^{(\mathcal{J})} &= ((\mathbf{X}'_{\mathcal{J}} \mathbf{X}_{\mathcal{J}})^{-1})_{k \in \mathcal{J}, l \in \mathcal{J}} \\ \hat{\mathbf{G}}_{k \notin \mathcal{J}, l \in \mathcal{J}}^{(\mathcal{J})} &= \hat{\mathbf{G}}_{k \in \mathcal{J}, l \notin \mathcal{J}}^{(\mathcal{J})} = \hat{\mathbf{G}}_{k \notin \mathcal{J}, l \notin \mathcal{J}}^{(\mathcal{J})} = 0, \end{aligned} \quad (4.3.28)$$

where  $\mathbf{X}_{\mathcal{S}}$  are the columns of  $\mathbf{BA}$  indexed by  $\mathcal{S}$ . Each weight  $\hat{p}_{\mathcal{S}}$  represents the estimated probability that the projection of  $\mathbf{y}$  lands on the cone's face corresponding to  $\mathcal{S}$ . The probabilities  $\hat{p}_{\mathcal{S}}$  are obtained by simulating many normal random vectors with mean vector  $\hat{\mathbf{y}} = \mathbf{BA}\hat{\boldsymbol{\beta}}_{RLS}^*$  and covariance matrix  $\hat{\sigma}_{RLS}^2 \mathbf{I}_n$ , and recording the resulting sets  $\mathcal{S}$ 's, along with their frequencies. In case additional unconstrained components are present, the definition of Eq. (4.3.27) can be extended [53]. Furthermore, if we need to select from a set of candidate models featuring different covariates, we use information criteria [56]. For our analyses we use the corrected Akaike Information Criterion (AICc)

$$AIC(M) = -2\log(L(\hat{\boldsymbol{\beta}}_{RLS}^*|\mathbf{y})) + 2p + 2p \cdot \frac{p+1}{n-p} \quad (4.3.29)$$

for model selection, where  $L(\boldsymbol{\beta}|\mathbf{y})$  is the likelihood of model  $M$  and  $p$  the number of parameters of  $M$ , which is equal to  $d$  in our set-up. In case some models have similar AICc values, we follow Burnham et al. [56] and we average across all ones using the frequentist model average estimator

$$\hat{\boldsymbol{\beta}}_{FMA} = \sum_{l=1}^L \lambda_l \hat{\boldsymbol{\beta}}_l, \quad (4.3.30)$$

where  $\hat{\boldsymbol{\beta}}_l$  is the parameter vector estimated under the  $l$ -th candidate model, and  $\lambda_l$  the corresponding weight which can be computed as

$$\lambda_l = \frac{\exp(-BIC_l/2)}{\sum_{j=1}^L \exp(-BIC_j/2)}, \quad (4.3.31)$$

where  $BIC_j$  is the Bayesian Information Criterion (BIC) associated with the  $j$ -th model. In case of model averaging, the BIC is preferred over AIC/AICc, since it provides a better estimation of the marginal likelihood [56]. From a Bayesian perspective,  $\{\lambda_j\}_j$  can be interpreted as an estimator of the posterior probabilities

$$p(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)p(M_j)}{\sum_j p(\mathbf{y}|M_j)p(M_j)}, \quad j = 1, \dots, J, \quad (4.3.32)$$

of the candidate models under a uniform prior  $\{p(M_j)\}_j$ , where

$$p(\mathbf{y}|M_j) = \int_{\Theta_j} p(\mathbf{y}|M_j; \boldsymbol{\Theta}_j) p(\boldsymbol{\Theta}_j|M_j) d\boldsymbol{\Theta}_j \quad (4.3.33)$$

is the marginal likelihood [56].

#### 4.3.4. PSEUDOCODE

The SCS method for rescaling the observed Shannon entropies given a set of confounders and a set of features of interest is summarised in the pseudocode of Algorithm 6.

**Input:**  $\mathbf{F}$ (features of interest),  $\mathbf{C}$ (confounders),  
 $\mathbf{h}$ (observed entropies)

**Output:**  $\mathbf{h}^{res}$ (rescaled entropies)

1. Get the entire design matrix:

$$\mathbf{B} = [\mathbf{1} \quad \mathbf{C} \quad \mathbf{F}]$$

2. Compute the equality constraints matrix  $\mathbf{A}$

3. Estimate the restricted least squares parameters:

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}^* \in \{\boldsymbol{\beta}^* \geq 0, \mathbf{CA}\boldsymbol{\beta}^* \geq 0\}}{\operatorname{argmin}} \{-2\boldsymbol{\beta}^* \mathbf{B} \log(\mathbf{h}) + \boldsymbol{\beta}^* (\mathbf{B}\mathbf{A})' \mathbf{B} \mathbf{A} \boldsymbol{\beta}^*\}$$

4. Get the parameters  $\hat{\boldsymbol{\beta}}_c^*$  of the confounders

5. Compute the rescaled entropies:

$$\mathbf{h}^{res} = \exp(\log(\mathbf{h}) - \mathbf{C}\hat{\boldsymbol{\beta}}_c^*)$$

**Algorithm 6:** SCS pseudocode

## 4.4. APPLICATIONS OF SCS IN NGS DATA

### 4.4.1. IN-VITRO ASSAY

To evaluate the reliability and sensitivity of the SCS rescaling method, we generated an IS dataset originating from an EBV-transformed B cell line transduced with a LV at Multiplicity of Infection (MOI) of 0.1, 1 and 10 to obtain DNA samples with increasing levels of polyclonality. Therefore, by increasing the MOI at each transduction we expect an increase in the vector copy number (VCN). As expected, the different vector doses resulted in different VCNs (see Tables 4.A.2-4.A.3). Different amounts of DNA (5, 20 and 100 ng) were used for IS retrieval. LV ISs were retrieved by Sonication Linker-mediated (SLiM) - PCR [57]. Briefly, DNA material

was sheared by sonication, subjected to end-repair and adenylation and then split in 3 technical replicates. Each replicate was ligated to a different barcoded linker cassette and subjected to two rounds of PCR allowing the amplification of the cellular genomic portion close to the vector IS. The different barcoded PCR products from each sample were assembled in libraries and sequenced by using Illumina platform. After sequencing, reads were processed by a dedicated bioinformatic pipeline [58] to identify for each PCR/sample the different vector integration sites. For each IS the clonal abundance was determined by the R package SonicLength [59] using the corresponding fragment length distribution. A varying number of ISs, ranging from 22 to 40575, was obtained from each sample (see Table 4.A.1) and, as expected, the number of IS retrieved from each sample increased proportionally to the vector dose (see Table 4.A.2). The total number of sample's sequencing reads was used as proxy for the sample's sequencing depth (SD). The magnitude of VCN, DNA amount and SD affects the clonality so that the samples are incomparable. Indeed Figure 4.2.1 clearly shows a positive trend between the Shannon entropy index and the potential confounders. With the VA we are able to really understand the impact of the variables (confounding factors) to the entropy index, thus allowing a robust integrated analysis. We used the VA as “ground-truth” to compare our SCS-rescaling method with the competitor approaches (RAR and SRS). In our SCS method we took in consideration the DNA amount, VCN and SD as potential confounders.

In this case the number of candidate confounders is  $K = 3$  with no additional factors of interest ( $J = 0$ ) and, according to the general formulation of Eq. (4.3.3), the model was defined as

$$\log(\mathbf{h}) = \mathbf{1}\beta_0 + \underbrace{[C_{dna} \ C_{vcn} \ C_{sd}]}_{\mathbf{C}} \underbrace{\begin{bmatrix} \beta_{dna} \\ \beta_{vcn} \\ \beta_{sd} \end{bmatrix}}_{\beta_c} + \boldsymbol{\varepsilon}, \quad (4.4.1)$$

where we used two equidistant interior knots and the range of values as boundary knots for every SCS term in  $\mathbf{C}$ . We report the corresponding fitted surface in Figure 4.4.1 showing only the partial surface corresponding to the DNA and the VCN. In Figure 4.4.1 we also show the rescaled values, i.e. the residuals that remain after having adjusted for the confounders. That

is, according to the model definition of Eq. (4.3.3), we used the residuals

$$\mathbf{h}^{res} = \exp(\log(\mathbf{h}) - \mathbf{C}\hat{\boldsymbol{\beta}}_c) \quad (4.4.2)$$

as the rescaled values, where  $\hat{\boldsymbol{\beta}}_c$  is the vector of the fitted parameters. We compared our method with the two previously proposed in literature, such as the rarefaction (RAR) [14] and the ranked subsampling (SRS) [19] approaches. We assessed the efficiency of the rescaling methods by correlation  $p$ -values for the two-sided test problem:

$$H_0 : \rho(\mathbf{h}^*, \mathbf{c}_k) = 0 \text{ vs } H_1 : \rho(\mathbf{h}^*, \mathbf{c}_k) \neq 0, \quad (4.4.3)$$

4

where  $\rho(\cdot, \cdot)$  is the Spearman's rank correlation function,  $\mathbf{h}^*$  is the vector of the Shannon entropies either observed or rescaled with one of the candidate methods, and finally  $\mathbf{c}_k$  is the vector of the corresponding values collected for the  $k$ -th confounder. We preferred Spearman's rank correlation over Pearson correlation since we assumed that the relationships are monotonic and possibly non-linear. Low  $p$ -values give statistical evidence for dependencies and thus for unsolved confounding effects. For the comparison, the total amount of reads of the sample with the lowest SD has been chosen as rarefaction level with 1000 replications for both the standard rarefaction (RAR) and its ranked version (SRS). We report the results in Figure 4.4.2. These pictures show that our SCS method outperformed both RAR and SRS methods in terms of correlation test  $p$ -values between the rescaled entropy and every potential confounder. Indeed, for all three confounders our new approach yields high  $p$ -values (0.37, 0.15 and 0.31 for DNA, VCN and SD respectively), so that we have no indication to reject the null hypothesis that the rescaled entropies and the confounder values are still correlated. For each of the three competing approaches we got 2-3 very low  $p$ -values ( $\ll 0.01$ ), so that statistically significant amounts of correlations are left.

Subsequently, we also checked whether our SCS-rescaling method unveils comparable clonal levels among the samples. A proper rescaling method should return similar clonal diversities independently from the confounders. Indeed, Figures 4.4.1 and 4.4.3 show that our SCS-rescaling method drastically reduced the variability of the observed entropies due to the effect of the confounders and, in turn, that the clonal level of the

VA samples, measured by the SCS-rescaled Shannon entropy index, is approximately the same. It can be seen from Figure 4.4.3 that the competitor methods RAR and SRS are also able to reduce the variability of the observed entropies. However, unlike our new SCS-rescaling method, the competing methods did not remove the effect of the confounders, as confirmed by the  $p$ -values of Spearman's rank correlation tests provided in Figure 4.4.2. While the SCS-rescaling method made all (rank) correlations insignificant, the competing methods RAR and SRS left significant rank correlations (=dependencies) between confounders and the entropy. For more explanations and illustrations we refer to Section 4.B.2 and Figure 4.B.2.

#### 4.4.2. VIRAL VECTOR SAFETY IN A GENOTOXICITY STUDY

We analysed the IS data collected from an established hematopoietic stem cell gene therapy model previously used to demonstrate how the genotoxic impact of integrating vectors is strongly modulated by their designs [20, 21]. In this experimental setup  $Cdkn2a^{-/-}$  tumor prone  $Lin^{-}$  cells were ex-vivo transduced with two different LVs expressing GFP: the highly genotoxic LV vector, LV.SF.LTR (hereinafter refer as LTR) and the non-genotoxic SIN.LV.PGK.GFP.PRE (hereinafter referred as PGK). Transductions protocol and culture conditions were reported in [20, 21]. Twenty-four hours after transduction, vector- and mock- transduced cells ( $5-7.5 \times 10^5$  cells/mouse) were transplanted into lethally irradiated wild-type mice by tail vein injection (Mock-control, N=19; LV.SF.LTR, N=24 and SINLV.PGK, N=23). Six days after transduction the percentages of GFP+ cells were assessed by Fluorescence Activated Cell Sorting (FACS) analysis and ranged from 90 to 95% for all the vectors and conditions. Engraftment level of transduced cells was assessed by measuring the percentage of GFP-expressing cells in the peripheral blood at 8 weeks post transplantation and were  $80.8 \pm 2.9$  % and  $46.2 \pm 4.8$  % in the group of mice transplanted with PGK and LTR vector respectively. As expected, mice transplanted with  $Cdkn2a^{-/-}$   $Lin^{-}$  cells transduced with the the LTR vector developed tumors and died significantly earlier compared to mock-treated mice ( $p < 0.0001$ , Log-rank Mantel-Cox test, median survival time: 282 and 149.5 days for mock-control and LTR- transduced group respectively). Mice transplanted with

PGK-transduced cells did not show any acceleration of tumor onset compared to the mock-control group (median survival time: 289 days). All data are in agreement with the one previously published [20, 21]. For the retrieval of vector insertion sites (ISs), peripheral blood was collected on a monthly basis from transplanted animals receiving transduced cells. Lymphoid B and T cells as well as myeloid cells were isolated by fluorescence activated cell sorting. To recover enough DNA material, equal amounts of blood from two or three mice belonging to the same experimental group were pooled before the sorting procedure. The composition of pools was maintained constant during the whole experiment, so that each pool is composed by the same mice over time. ISs were then retrieved by SLiM-PCR[60] at different time points from sorted T (CD3+) and B (CD19+) lymphocytes, from myeloid cells (CD11b+) and unsorted blood cells (total MNC). From the DNA purified from all the different sorted samples, we also measured the VCN by ddPCR. Overall, a higher amount of ISs were retrieved from the group of mice transplanted with  $Lin^-$  cells transduced with PGK, reflecting the higher level of VCN observed in PGK versus LTR group of transplanted animals. Few statistics on the number of IS collected in each treatment/condition, along with the corresponding VCN, are reported in Table 4.4.1. The Shannon entropy index was then computed from each IS sample and the application of a simple spline without shape constraints and without considering any technical confounder yielded the results shown in Figure 4.4.4. From Figure 4.4.4 we cannot see a clear separation between the entropy profiles of the two vectors PGK and LTR. The prediction intervals overlap so that the differences in the profiles do not appear to be statistically significant. Henceforth, we cannot draw the conclusion that PGK is safer than LTR. However, the high variability of DNA amount (in nanograms), VCN, and the SD used for IS retrieval has a clear impact on the entropy measurements, as suggested by Figure 4.4.5. Furthermore, since some mice died faster, the size of each pool (PS) decreased over time, leading to variation in the cell counts and in turn in the Shannon diversity index calculations (see Figure 4.4.5). Few statistics of these quantities are reported in Table 4.4.1 separately for each vector treatment. This suggests that initial results of figure 6 might be biased by the presence of these confounding factors. The heterogeneity of these factors may affect the estimate of the cell counts and in turn the corresponding

	PGK					LTR				
	DNA	VCN	PS	SD	$n_{IS}$	DNA	VCN	PS	SD	$n_{IS}$
Min.	8.64	1.31	1	60	35	8.64	0.24	1	189	35
1st Qu.	106.56	10.9	2	1969	433	94.50	5.32	1	1130	217
Median	200.00	13.59	2	5881	720	200.00	6.3	2	2973	383
Mean	181.07	12.8	1.964	9351	989.3	222.88	6.219	2.104	4695	731.9
3rd Qu.	200.25	13.9	2	14055	1220	222.50	7.8	3	7390	873
Max.	973.00	27	3	49853	4324	973.00	10.5	7	15375	3213

**Table 4.4.1** | Mice study: Quartiles and range of the DNA amount, VCN, PS, SD and  $n_{IS}$  for the  $n = 242$  samples and separately for PGK (top) and LTR (bottom) treatment conditions.

Shannon entropies. We therefore applied our shape-constrained spline approach of Section 3, including the DNA amount, VCN, SD and PS as potential confounders.

We proceeded as follows: We used the general formulation of Eq. (4.3.3), including a shape constrained spline (SCS) term with two interior knots for every confounder, plus a spline term w.r.t. the time decay for every combination of cell lineage/marker ( $L$ ) and viral vector ( $V$ ) as additional factors of interest. In this way we described the entropy decays separately for each combination of cell marker and treatment (viral vector) while removing the bias provided by the potential confounders. We also set the vector specific intercept  $V$  to zero to make sure all the individuals have the same clonal diversity before the treatments. Therefore, following the general formulation of Eq. (4.3.3), the model it has been explicitly defined as

$$\log(\mathbf{h}) = \mathbf{1}\beta_0 + \underbrace{[\mathbf{C}_{dna} \ \mathbf{C}_{vcn} \ \mathbf{C}_{ps} \ \mathbf{C}_{sd}]}_{\mathbf{C}} \underbrace{\begin{bmatrix} \beta_{dna} \\ \beta_{vcn} \\ \beta_{ps} \\ \beta_{sd} \end{bmatrix}}_{\beta_c} + \underbrace{\begin{bmatrix} \mathbb{S}_t^{l_1} & & & \\ & \mathbb{S}_t^{l_2} & & \\ & & \mathbb{S}_t^{l_3} & \\ & & & \mathbb{S}_t^{l_4} \end{bmatrix}}_{\mathbf{F}} \underbrace{\begin{bmatrix} \beta_{l_1} \\ \beta_{l_2} \\ \beta_{l_3} \\ \beta_{l_4} \end{bmatrix}}_{\beta_f} + \boldsymbol{\varepsilon}, \quad (4.4.4)$$

where  $\mathbf{C}$  binds all the confounder's SCS bases and  $\beta_c$  is the vector with all the corresponding parameters stacked together. Alike,  $\mathbf{F}$  is a block-diagonal matrix where each block  $\mathbb{S}_t^l$  is defined as

$$\mathbb{S}_t^l = \begin{bmatrix} 1 & \mathbb{S}_t^{l,v_1} \\ & & & \\ & & & \mathbb{S}_t^{l,v_2} \\ & & & & & \end{bmatrix}, \quad (4.4.5)$$



and each sub-block  $\mathbb{S}_t^{l,\nu}$ , corresponding to the  $l$ -th cell lineage and the viral vector  $\nu$ , is the basis of a monotone decreasing quadratic spline w.r.t. the time  $t$  for which we assume a steady-state to the left of the second right boundary knot. Indeed, each mouse pool started with 2-7 mice, which then successively died, until no mouse was left, so that no measurements could be taken anymore and therefore we do not expect any further change in the entropy thereafter. For this purpose we use again the affine transformation defined in Eqs. (4.3.18) - (4.3.19). Finally, we refer to  $\boldsymbol{\beta}_f$  as the vector with all the corresponding parameters stacked together. Therefore in this case the number of confounders is  $K = 4$  and the number of additional factors of interest is  $J = 8$ , namely a spline basis for the time-decay for every combination of the two treatments and the four cell lineages.

In order to identify the most important confounders among the candidates, we have fitted our model for each of the  $2^4 - 1 = 15$  possible confounder subsets. Each candidate model included always  $\mathbf{F}$  as fixed term and featured at least one SCS term in  $\mathbf{C}$  for the confounders. Then we averaged across the most likely models according to the frequentist criterion defined in Eqs. (4.3.30) - (4.3.31) and we reported in Figure 4.4.6 the posterior distribution, along with the marginal inclusion probabilities of the four individual confounders. Results from model averaging suggest that the posterior distribution is mainly dominated by three models namely: PS + SD (4th model), VCN + SD (6th model), and SD (8th model). The remaining 12 models get substantially lower posterior probabilities and thus have only negligible effects on the model averaging estimator. Therefore, after computing the frequentist model averaging estimator

$$\hat{\boldsymbol{\beta}}_c^{fma} = \left[ \hat{\boldsymbol{\beta}}_0 \quad \left( \hat{\boldsymbol{\beta}}_c^{fma} \right)' \quad \hat{\boldsymbol{\beta}}_f' \right]' \quad (4.4.6)$$

of Eq. (4.3.30), we used the residuals

$$\mathbf{h}^{res} = \exp \left( \log(\mathbf{h}) - \mathbf{C} \hat{\boldsymbol{\beta}}_c^{fma} \right) \quad (4.4.7)$$

corresponding to the confounder terms as the rescaled values. Rescaled entropies are shown in Figure 4.4.7 together with the lineage  $\times$  vector-specific spline decays with a confidence interval of 0.95 coverage. Thanks to the SCS-rescaling approach, a significant difference in the entropy decay for

MNC, T-cells (CD3+) and Myeloid cells (CD11b) was observed depending on the genotoxicity level of the vector adopted. Whereas in the B-cell compartment no major differences in the entropy decay under the two vector treatments were observed. Indeed, consistently with the previous results [20, 21], the B-cell compartment is less affected by the genotoxicity of the LTR vector. Figures 4.B.5-4.B.7 from Section 4.B.2 clearly indicate that our proposed method SCS outperforms the competitor ones in detecting such differences.

## 4.5. DISCUSSION

We have shown that the Shannon entropy index, a widely used measure of genetic variability, is strongly affected by the variability of technical factors. We have introduced a shape-constrained splines (SCS) approach aimed to quantitatively measure and remove the effect of confounders from the target of interest. In particular we have shown that our approach can remove confounding effects from the Shannon entropy index. We also have shown that our SCS approach outperforms all the state of the art rarefaction approaches like the RAR [14] and its ranked-subsampling version [19]. That is, using a correlation test, we have found statistical evidence that the SCS-rescaled diversity measure does not significantly depend on the effect of the confounders anymore. Furthermore, our method is useful for genetic applications, as it provides an unbiased and more affordable measure of clonal diversity, and in turn it avoids drawing misleading results. As an example, the entropy decay of treatment-specific longitudinal studies may be erroneously interpreted if we do not take into account that the changes in the entropy increments may depend more on the confounders than on the biological treatments. Our method allows to discriminate between the two effects and to remove the one that comes from the confounders.

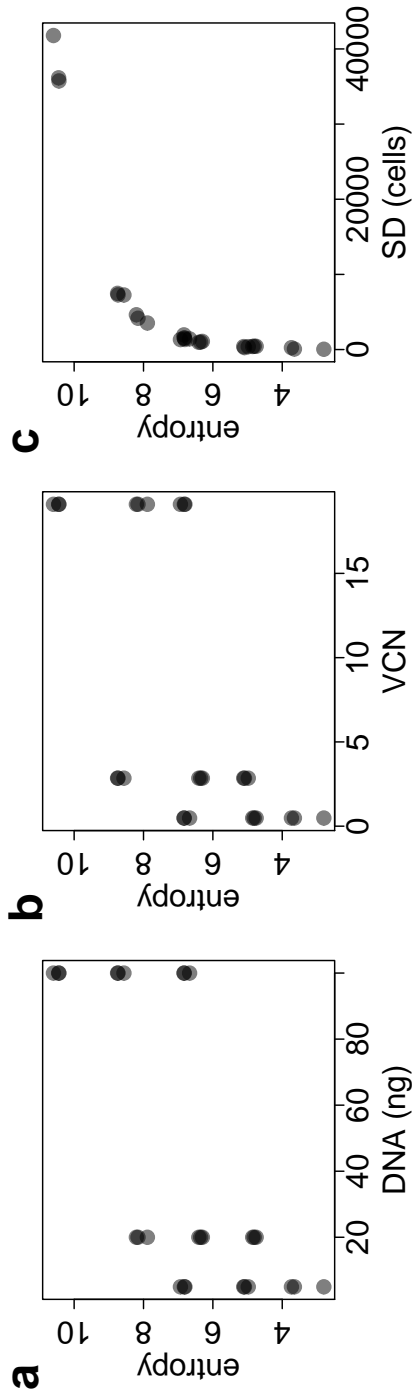
Since our approach is spline-regression based, its main limitation could be related to the available sample size. This is a potential problem when when the number  $n$  of libraries is too low to define a spline basis. For the same reason the degree of the splines and the number of its knots should be chosen carefully. In particular, for the case of only one library/sample it would be possible to rescale its diversity only using the

parameters inferred from an external controlled environment, like the VA explored in Section 4.4.1, with a sufficient library size  $n$ . This is the main price we pay if we switch from a rarefaction-based rescaling approach to a spline-regression based one. Our model averaging approach allows also to rank the impact of the confounders according to their approximated posterior probabilities. We perform model averaging by means of the Bayesian Information Criterion which allow us to get an estimate of the marginal likelihood of each candidate model, and in turn, of the corresponding marginal confounder inclusion posterior probabilities. One possible methodological extension of this framework could be the implementation of a more precise method to estimate the marginal likelihood. This could, for example, either be done by Laplace Integration or by Bayesian thermodynamic integration.

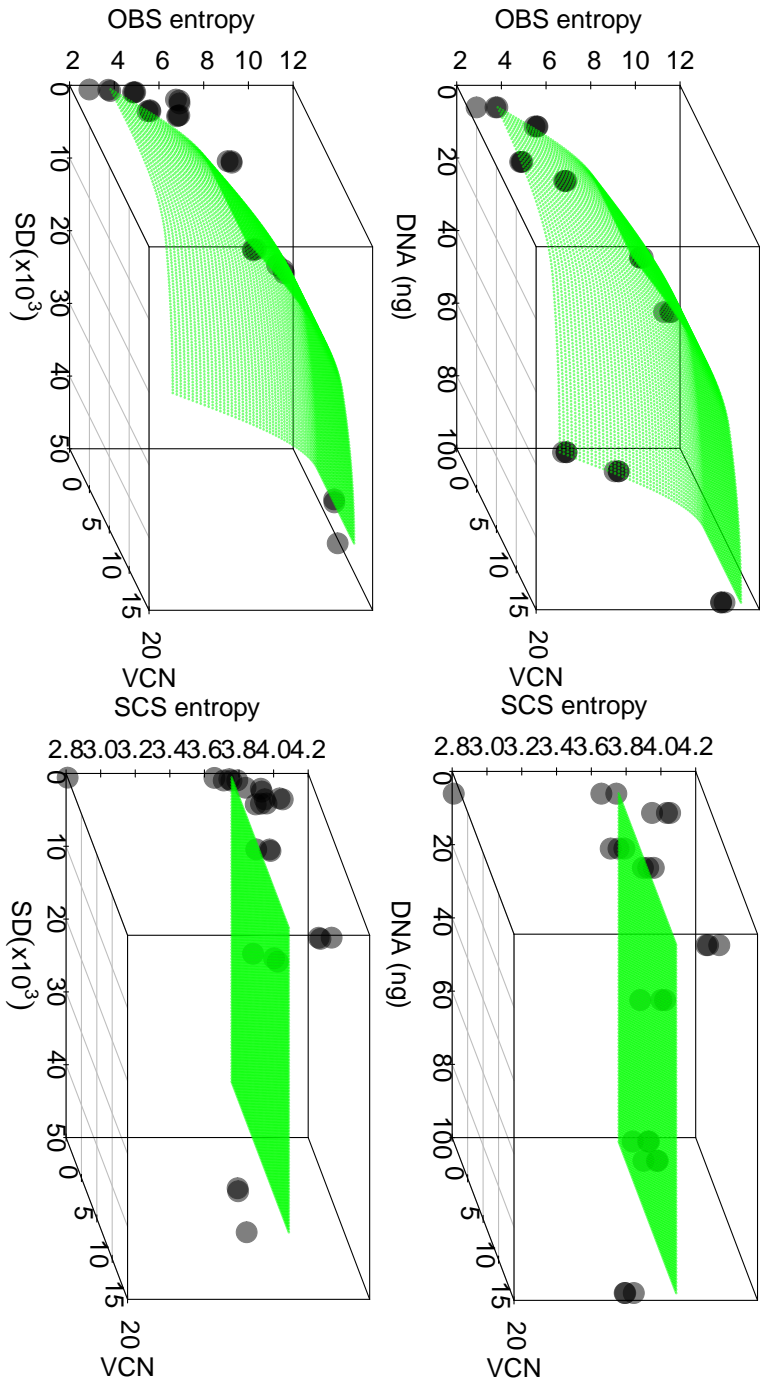
4

## 4.6. AVAILABILITY OF DATA AND MATERIALS

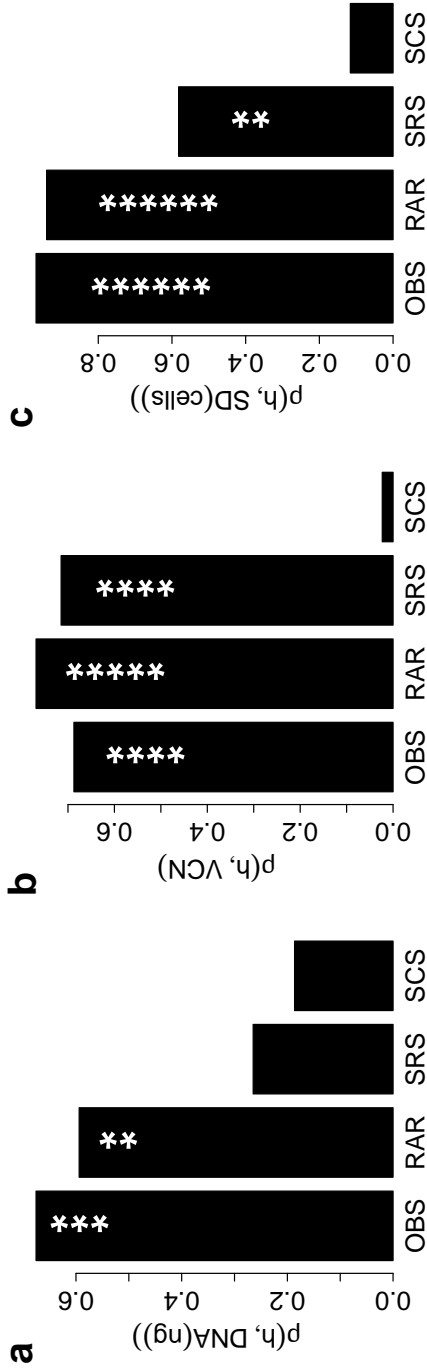
The code that supports the findings of this study is openly available at <https://github.com/delcore-luca/SCS>.



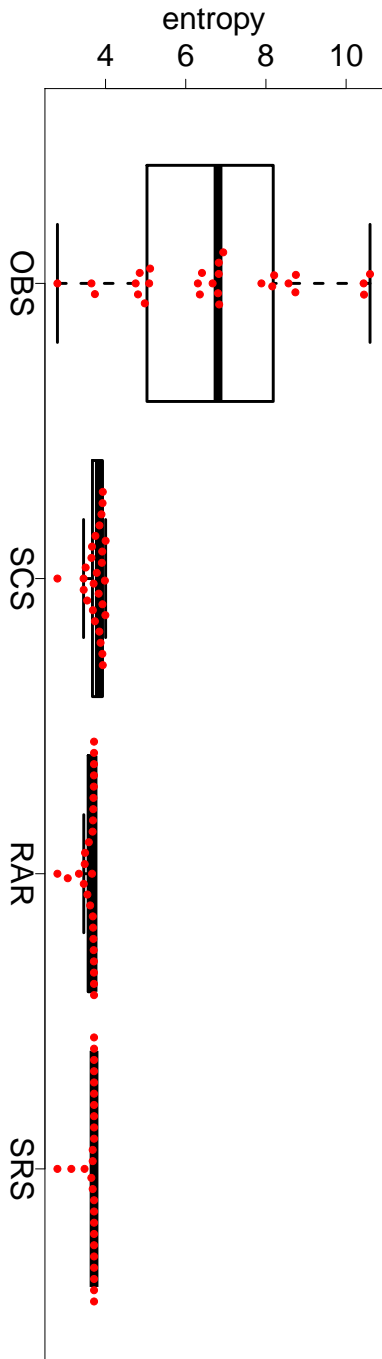
**Figure 4.2.1** | From left to right: Scatter plot of the Shannon entropy index against the DNA amount, the vector copy number (VCN) and the sequencing depth (SD) for all the samples included in the in-vitro assay. Only a single amount of DNA has been taken for every sample. The total amount of integrations found in a sample, namely the total number of sample's sequencing reads, has been used as proxy for the sample's sequencing depth (SD).



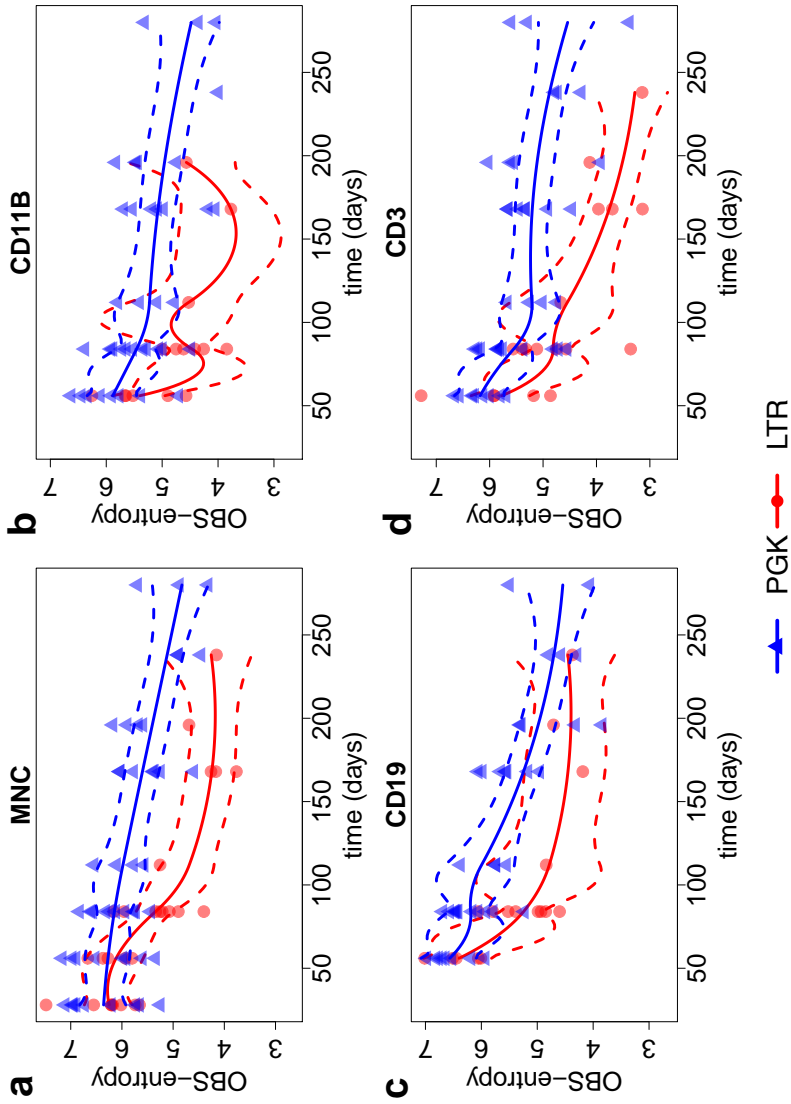
**Figure 4.4.1** | Observed (left) and SCS-rescaled (right) entropies (dot symbols) as a function of the confounders, together with the corresponding shape constrained (bivariate) splines (green surface). Top panels show the slices for the DNA and the VCN. Bottom panels show the slices for the SD and the VCN.



**Figure 4.4.2** | Each panel shows the absolute value of the Spearman's rank  $\rho$  correlation coefficient  $\rho(h, \text{confounder})$  ( $y$ -axis) between a confounder and the observed or rescaled Shannon entropies (different bars). For every correlation coefficient, we performed the two-sided Spearman's rank correlation test of Eq. (4.4.3) for checking the hypotheses  $H_0 : \rho(h, \text{confounder}) = 0$  VS  $H_1 : \rho(h, \text{confounder}) \neq 0$ . The number of leading zeros after the decimal point of the  $p$ -values are reported on top of each bar as white stars.

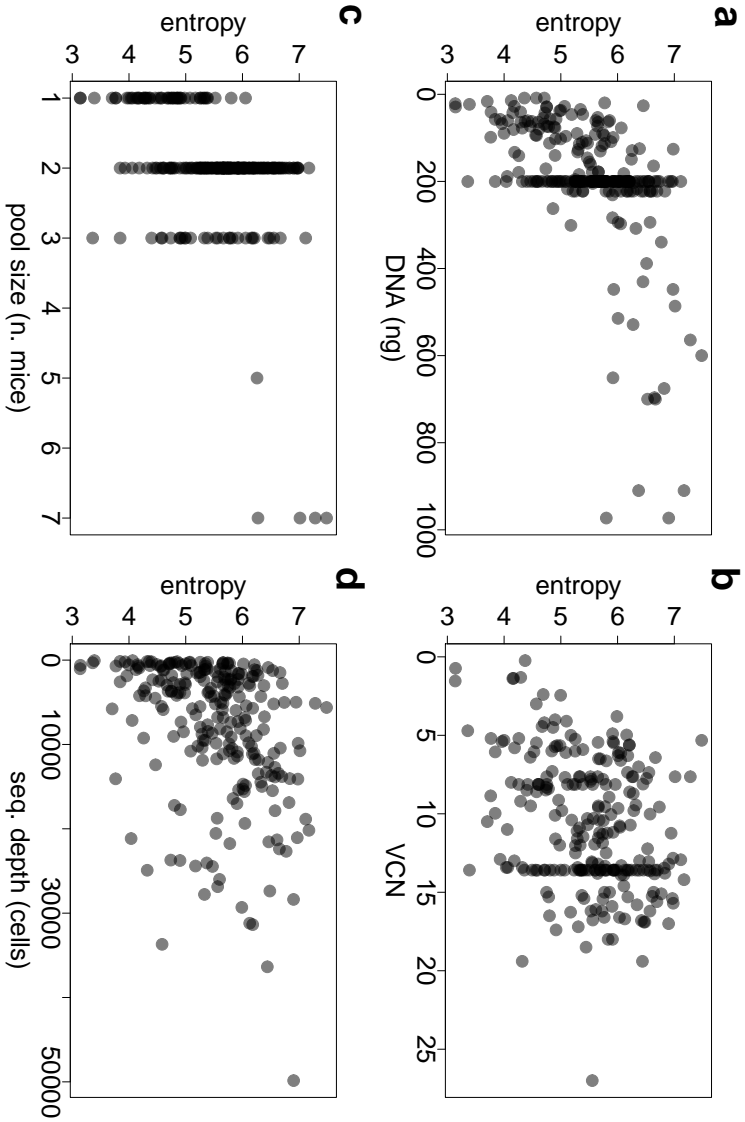


**Figure 4.4.3** | Box-plot (minimum, maximum, median, first quartile and third quartile) of the observed (OBS) and rescaled Shannon entropies using the SCS, RAR and SRS approaches.

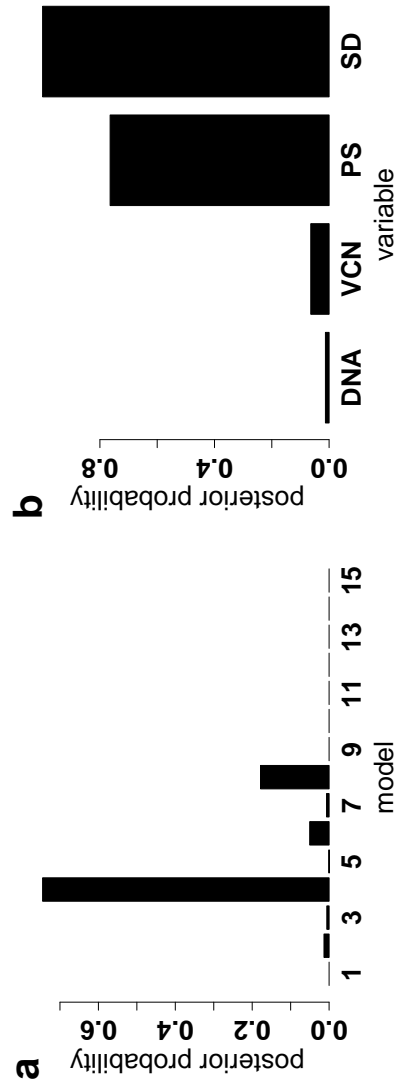


**Figure 4.4.4** | Observed Shannon entropies (dot symbols) over time (x-axis) in each treatment (different colors), along with a simple spline without any shape constraints and confounder adjustments for every combination of cell marker and viral vector. Quadratic splines are fitted using the standard  $1m()$  and  $bs()$  R functions.

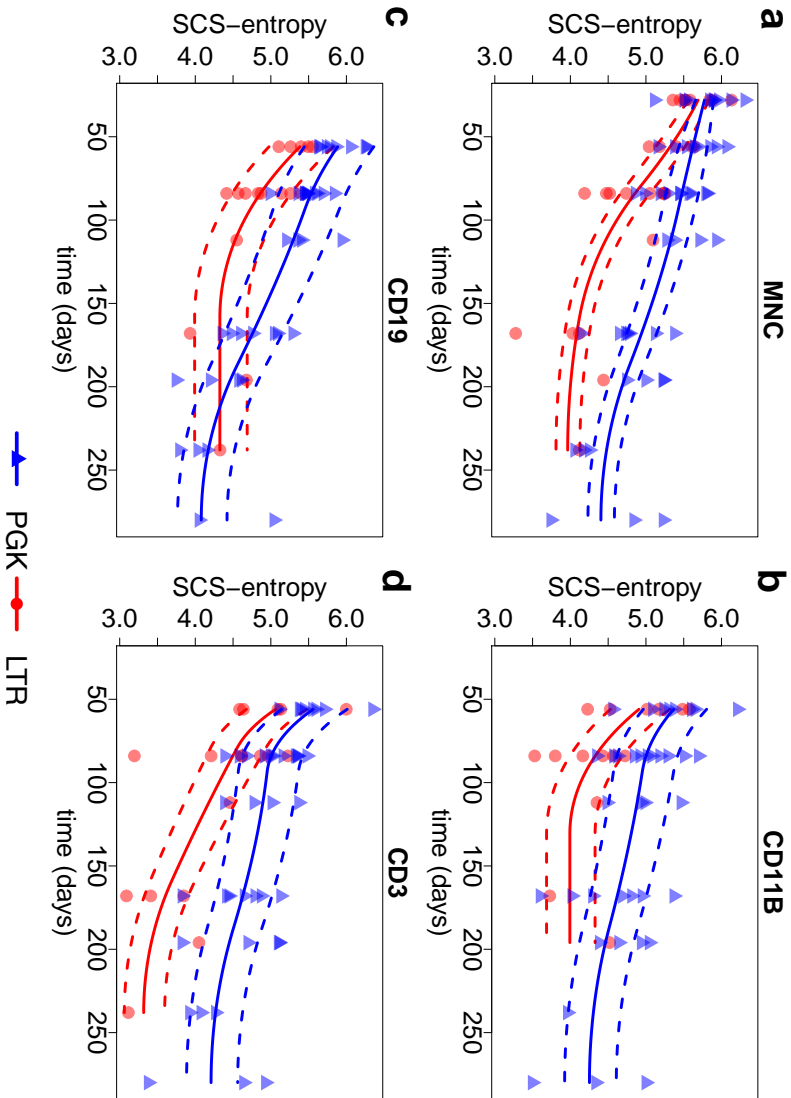




**Figure 4.4.5** | From top left to bottom right: Scatter plot of the raw (unscaled) Shannon entropy index computed from the entire dataset ( $n = 242$  IS samples) versus the DNA amount, the vector copy number (VCN), the pool size (PS) and the sequencing depth (SD).



**Figure 4.4.6** | Approximated posterior distribution (left) of the 15 candidate models according to the frequentist model averaging method of Eqs. (4.3.30) - (4.3.31), and the marginal posterior probabilities of the 4 potential confounders (right).



**Figure 4.4.7** | Rescaled Shannon entropies (y-axis) over time (x-axis) for every combination of cell marker (panels) and viral vector treatment (colors) together with the corresponding decay splines.

# APPENDIX

## 4.A. IN-VITRO ASSAY

In this experimental setup the genomic DNA of a bulk-transduced cell line harboring lentiviral insertions randomly distributed in the genome. This is a controlled environment designed to quantify the impact of several confounding factors to the clonal entropy. More specifically, the genomic DNA is obtained from a polyclonal lentiviral-vector (LV) marked cell line (JY). The DNA material was sheared by sonication, end-repaired and adenylated, split in technical triplicates, ligated of the barcoded linker cassettes and subjected to PCR amplification protocols for IS retrieval. Clonal quantification is obtained using the R package `SonicLength`[59]. The total amount of integrations found in a sample, namely the total number of sample's sequencing reads, has been used as proxy for the sample's sequencing depth (SD). Few summary statistics of the DNA amount, multiplicity of infection (MOI), VCN, number of distinct IS ( $n_{IS}$ ) and SD are provided in Table 4.A.1. In Table 4.A.2 we provide the VCN, the total number of distinct ISs and the total SD in each condition, that is for every combination of DNA amount and MOI. The sample-specific information is provided in Table 4.A.3.

## 4.B. SUPPLEMENTARY FIGURES

### 4.B.1. FITTING QUADRATIC AND CUBIC SPLINES

We fitted and compared quadratic and cubic splines on the VA data of Section 4.4.1. First, we used quadratic splines with one interior knot for the DNA amount and the VCN and two interior knots for the sequencing depth. Second, we used cubic splines with no interior knots for the DNA amount and the VCN and two interior knots for the sequencing depth. The number of knots were chosen, so as to get a full-rank design matrix,

DNA	MOI	VCN	$n_{IS}$	SD
Min. : 5.00	Min. : 0.1	Min. : 0.484	Min. : 22	Min. : 41
1st Qu.: 5.00	1st Qu.: 0.1	1st Qu.: 0.484	1st Qu.: 188	1st Qu.: 419
Median : 20.00	Median : 1.0	Median : 2.850	Median : 1004	Median : 1340
Mean : 41.67	Mean : 3.7	Mean : 7.478	Mean : 5544	Mean : 6032
3rd Qu.:100.00	3rd Qu.:10.0	3rd Qu.:19.100	3rd Qu.: 3811	3rd Qu.: 4398
Max. :100.00	Max. :10.0	Max. :19.100	Max. :40575	Max. :41787

**Table 4.A.1** | In-vitro assay (VA): Quartiles and ranges of the DNA amount, MOI, VCN,  $n_{IS}$  and SD for the  $n = 27$  samples.

	DNA	MOI	VCN	$n_{IS}$	SD
1	5	0.10	0.48	143.00	361.00
2	20	0.10	0.48	548.00	1288.00
3	100	0.10	0.48	3307.00	4936.00
4	5	1.00	2.85	578.00	1043.00
5	20	1.00	2.85	2008.00	3044.00
6	100	1.00	2.85	19239.00	22021.00
7	5	10.00	19.10	3166.00	4163.00
8	20	10.00	19.10	10506.00	12331.00
9	100	10.00	19.10	110182.00	113689.00

**Table 4.A.2** | In-vitro assay (VA): VCN, total number of distinct ISs and total SD in each of the nine conditions (combination of DNA amount and MOI).

and in turn, a positive-definite quadratic form, which is needed during optimization. Results are shown below in figure 4.B.1. The results in figure 4.B.1 show that there is not much difference between the two fitted surfaces (visual inspection). To be more objective, we also computed the two corrected Akaike Information criteria (cAIC) and found that the quadratic spline yields a lower cAIC value than the cubic spline.

## 4.B.2. COMPARISONS OF RESCALING METHODS

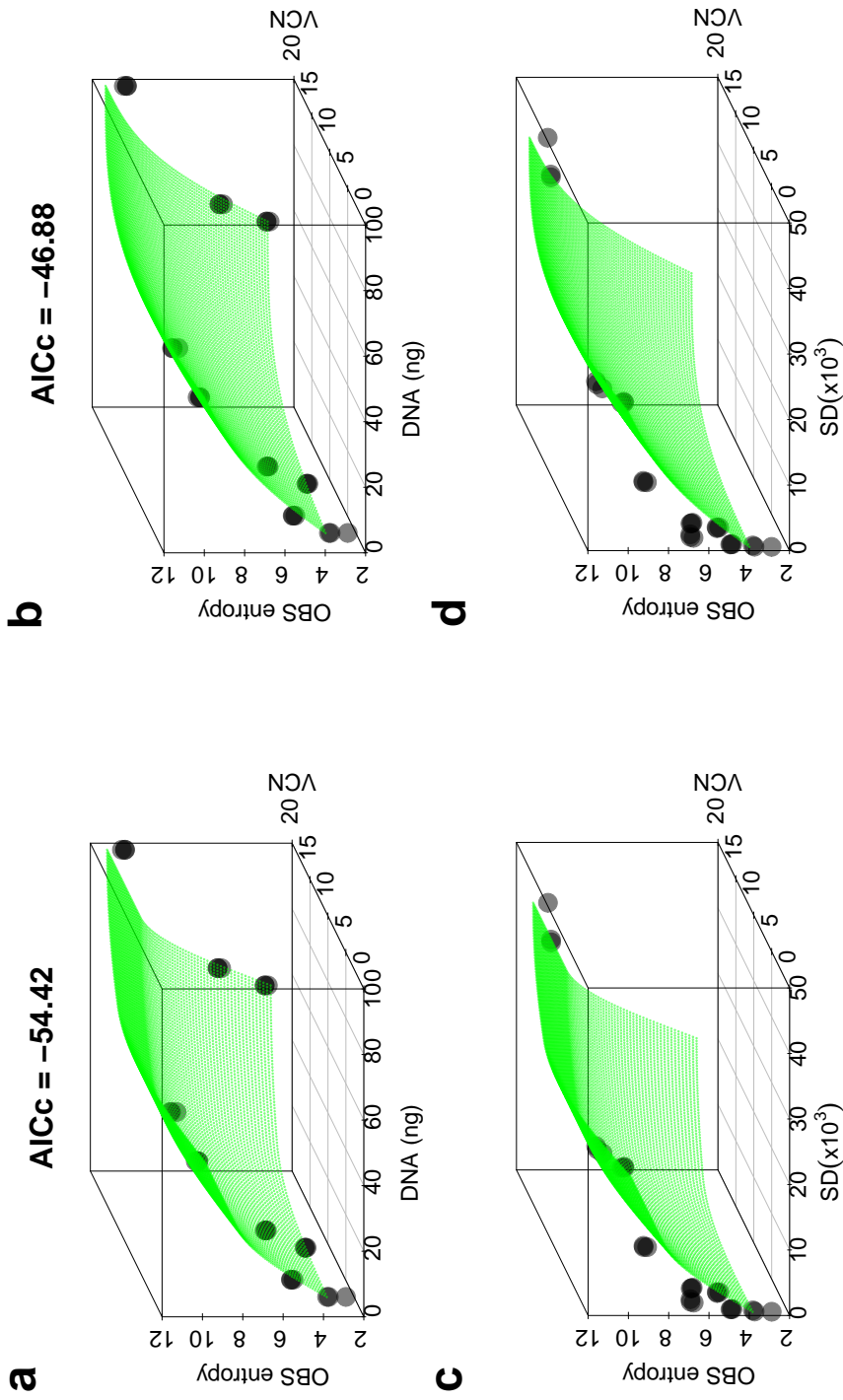
To make clearer what we mean by “RAR and SCS leave significant rank correlations (=dependencies)” in Section 4.4.1, we report here two additional figures. Figure 4.B.2 shows an additional graphical representation of the Spearman’s rank correlation tests performed in Section 4.4.1. Figure

4.B.2 clearly shows that our SCS-rescaled entropies outperform both RAR- and SRS- rescaled entropies in terms of dependence on the candidate confounders. Indeed, the red-highlighted square indicate that both RAR- and SRS- rescaled entropies still suffers from significant dependence on the confounders. Whereas, the SCS-rescaled entropies do not depend on any of the confounders anymore (all three correlations are insignificant). These are the results that we report in figure 3. Figures 4.B.3-4.B.3 show the observed entropies and the SCS-, RAR-, SRS- rescaled Shannon entropies against the candidate confounders, along with the fitted surfaces. These figures show that only the SCS-rescaled entropies do not depend on the confounders anymore, as the fitted surface is an (almost) horizontal plane. The surfaces fitted for RAR- and SRS- rescaled entropies are non-horizontal planes, as the entropies still depend on the confounders.

These analyses better clarify why SCS performs significantly better than the competitor methods. For the mice study data we do not know the ‘ground truth’, so that we cannot objectively cross-compare the performances of the different methods. However, we have applied all three methods to monitor the rescaled entropies in the mice study. The results are reported in Figures 4.B.5 - 4.B.7, showing that the proposed SCS-rescaling method yields the tightest confidence intervals and shows the clearest (smoothest) trends.

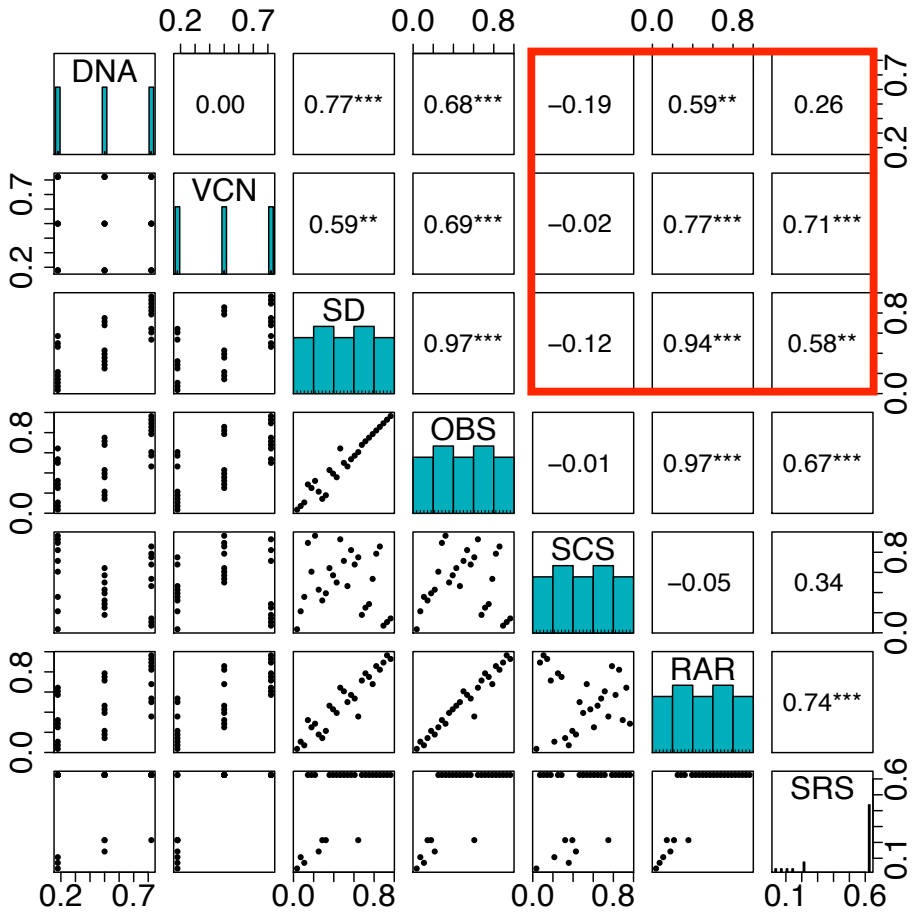
DNA	MOI	VCN	$n_{IS}$	SD
100	0.10	0.48	1300	1954.00
100	1.00	2.85	6082	7264.00
100	10.00	19.10	40575	41787.00
20	0.10	0.48	183	429.00
20	1.00	2.85	649	1087.00
20	10.00	19.10	3933	4631.00
5	0.10	0.48	78	250.00
5	1.00	2.85	204	395.00
5	10.00	19.10	1051	1492.00
100	0.10	0.48	1074	1557.00
100	1.00	2.85	6632	7479.00
100	10.00	19.10	34629	35743.00
20	0.10	0.48	177	450.00
20	1.00	2.85	692	966.00
20	10.00	19.10	2884	3534.00
5	0.10	0.48	43	70.00
5	1.00	2.85	188	387.00
5	10.00	19.10	1004	1340.00
100	0.10	0.48	933	1425.00
100	1.00	2.85	6525	7278.00
100	10.00	19.10	34978	36159.00
20	0.10	0.48	188	409.00
20	1.00	2.85	667	991.00
20	10.00	19.10	3689	4166.00
5	0.10	0.48	22	41.00
5	1.00	2.85	186	261.00
5	10.00	19.10	1111	1331.00

**Table 4.A.3** | In-vitro assay (VA): VCN, number of distinct ISs and SD in each of the  $n = 27$  samples.

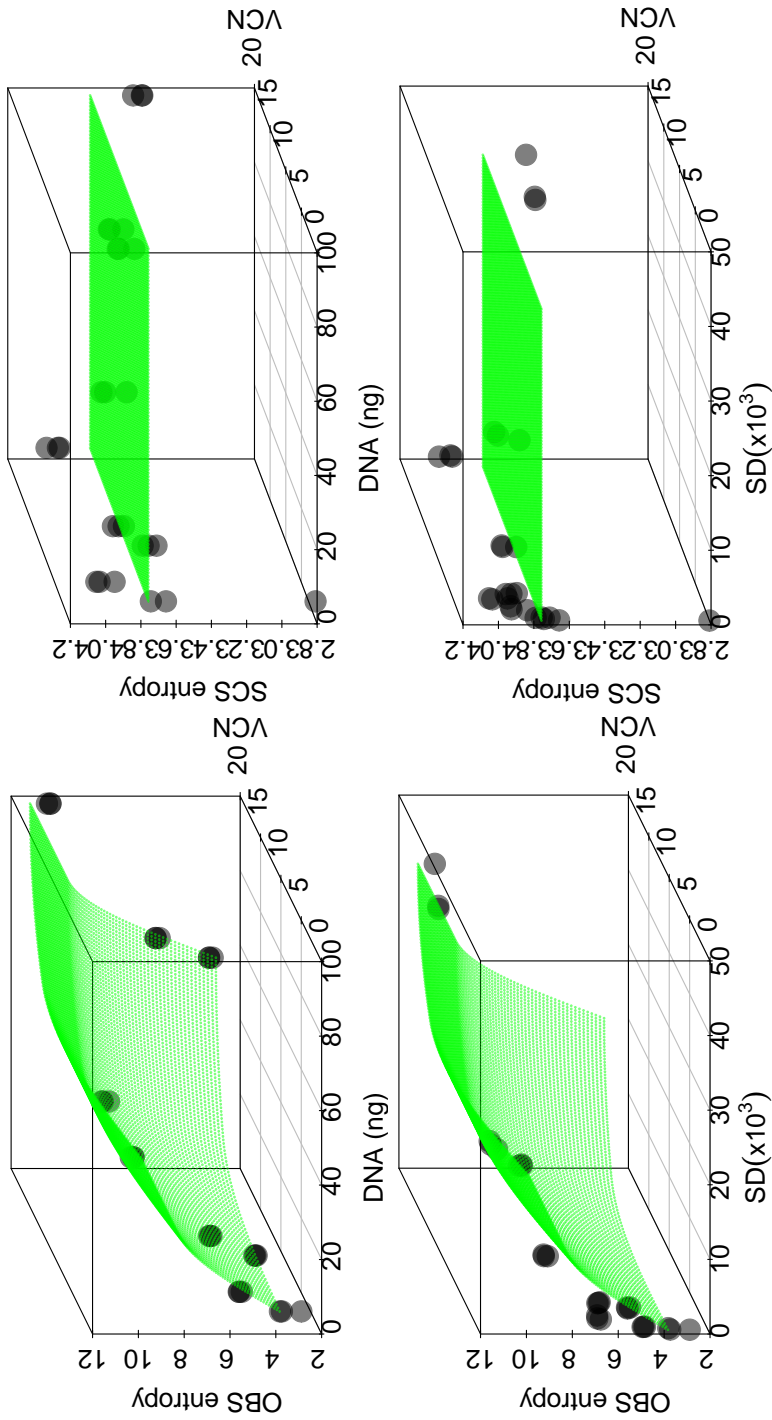


**Figure 4.B.1** | In-vitro Assay (VA): Quadratic (left) and cubic (right) spline fitting of the Shannon entropies ( $z$ -axis) against the candidate confounders. The  $x$ -axis refers to the DNA amount (top panels) or the SD (bottom panels). The  $y$ -axis refers to the VCN. The corresponding AIC is reported at the top.

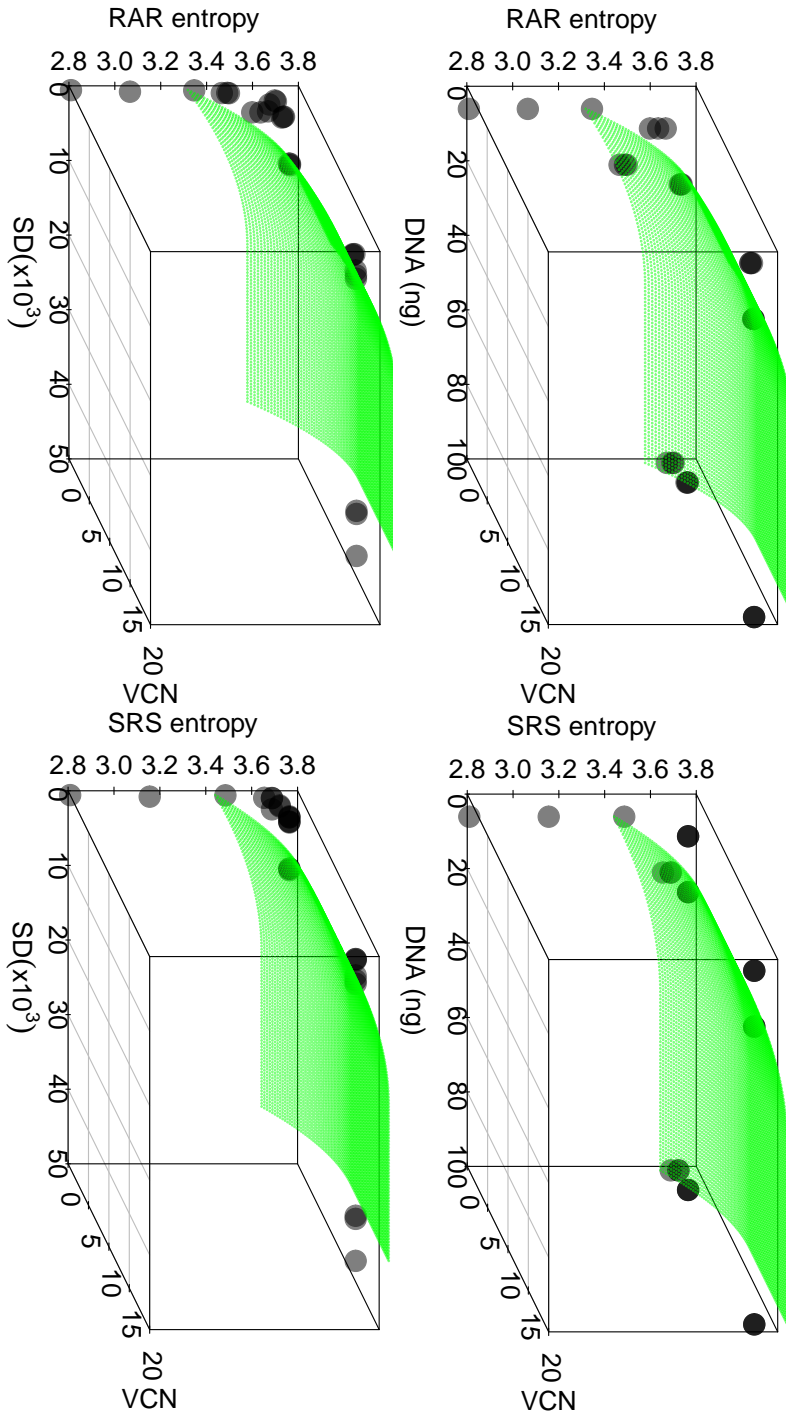




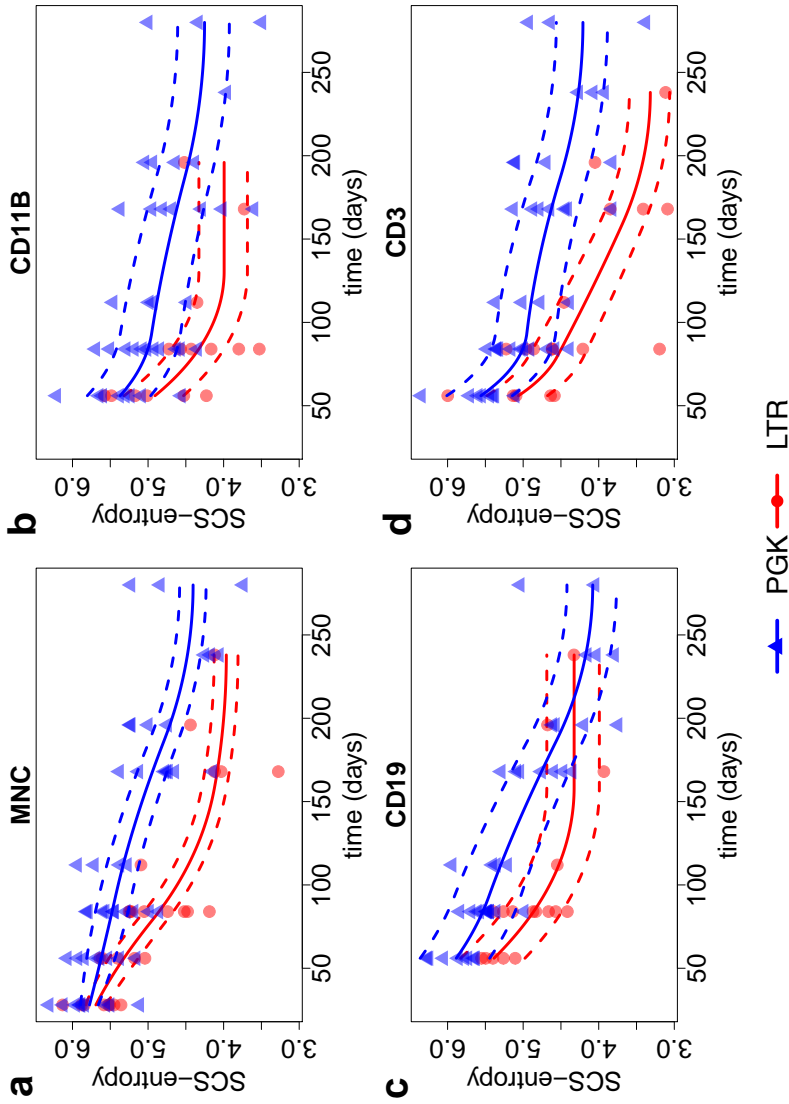
**Figure 4.B.2** | Pairwise correlation scatter plots between ranks of the candidate confounders (DNA, VCN, SD), the ranks of the observed entropies (OBS), and the ranks of the rescaled entropies obtained with the three methods SCS, RAR and SRS. The ranks have been rescaled by the sample size. Significant correlations are labelled with star symbols “\*”, where the significance level increases (p-values decreases) in the number of stars. We have put a red-highlighted square around the most relevant correlations.



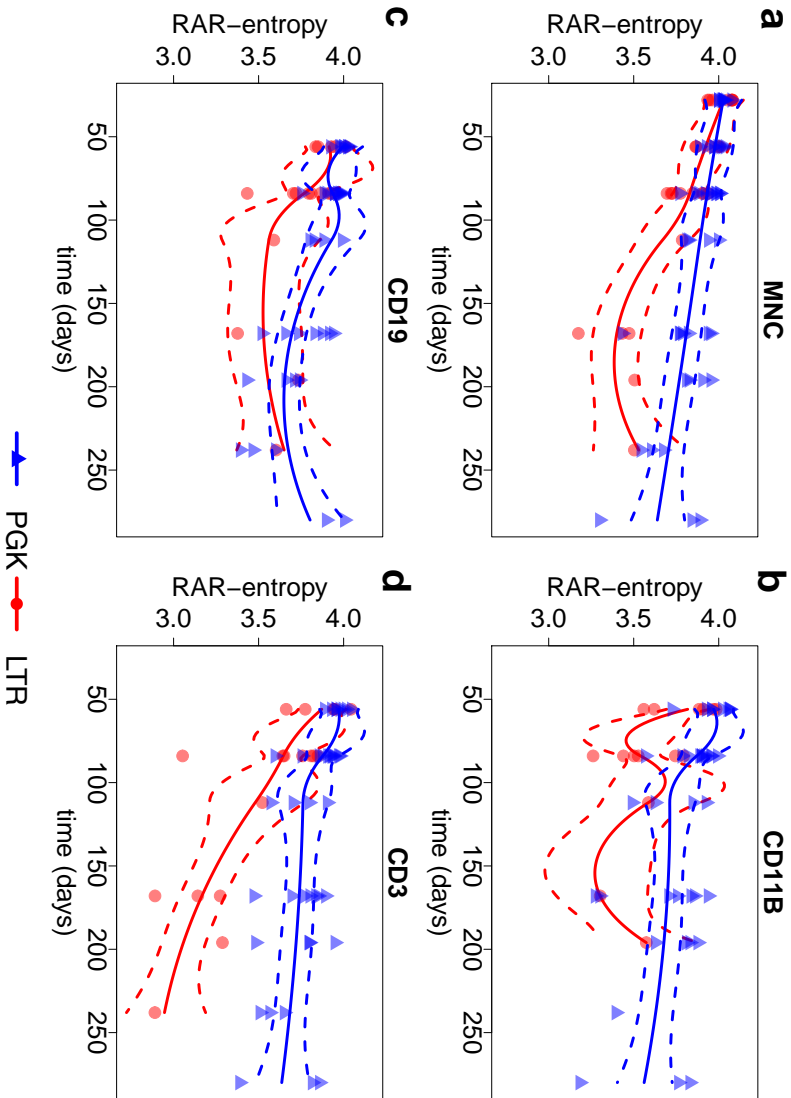
**Figure 4.B.3** | Each panel shows a three-dimensional scatterplot. The x-axes refer to the DNA amount or the SD. The y-axes refer to the VCN, and the z-axes refer to the entropies. The four panels refer to (left): the observed entropies, (right): the SCS-rescaled entropies. In each panel the fitted surfaces are shown in green.



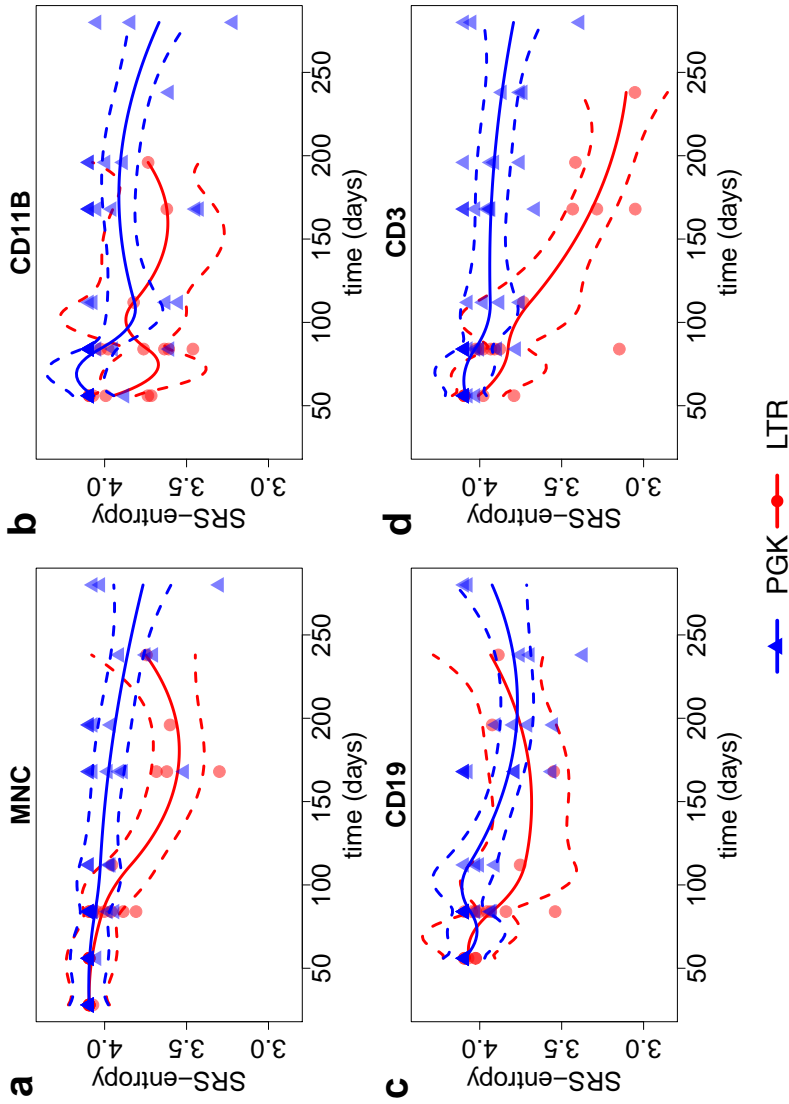
**Figure 4.B.4** | Each panel shows a three-dimensional scatterplot. The x-axes refer to the DNA amount or the SD. The y-axes refer to the VCN, and the z-axes refer to the entropies. The four panels refer to (left): the RAR-rescaled entropies, and (right): the SRS-rescaled entropies. In each panel the fitted surfaces are shown in green.



**Figure 4.B.5** | SCS-rescaled entropies: For each of the four lineages there is a panel with two fitted curves for the two viral vector conditions (different colors). The dotted lines refer to confidence intervals with 0.95 coverage.



**Figure 4.B.6** | RAR-rescaled entropies: For each of the four lineages there is a panel with two fitted curves for the two viral vector conditions (different colors). The dotted lines refer to confidence intervals with 0.95 coverage.



**Figure 4.B.7** | SRS-rescaled entropies: For each of the four lineages there is a panel with two fitted curves for the two viral vector conditions (different colors). The dotted lines refer to confidence intervals with 0.95 coverage.

## REFERENCES

- [1] L. Del Core, D. Cesana, P. Gallina, Y. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. Grzegorzczuk, *Normalization of clonal diversity in gene therapy studies using shape constrained splines*, *Scientific Reports* **12**, 1 (2022).
- [2] C. E. Dunbar, K. A. High, J. K. Joung, D. B. Kohn, K. Ozawa, and M. Sadelain, *Gene therapy comes of age*, *Science* **359**, eaan4672 (2018).
- [3] A. Aiuti, L. Biasco, S. Scaramuzza, F. Ferrua, M. P. Cicalese, C. Baricordi, F. Dionisio, A. Calabria, S. Giannelli, M. C. Castiello, M. Bosticardo, C. Evangelio, A. Assanelli, M. Casiraghi, S. Di Nunzio, L. Callegaro, C. Benati, P. Rizzardi, D. Pellin, C. Di Serio, M. Schmidt, C. Von Kalle, J. Gardner, N. Mehta, V. Neduva, D. J. Dow, A. Galy, R. Miniero, A. Finocchi, A. Metin, P. P. Banerjee, J. S. Orange, S. Galimberti, M. G. Valsecchi, A. Biffi, E. Montini, A. Villa, F. Ciceri, M. G. Roncarolo, and L. Naldini, *Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich Syndrome*, *Science* **341** (2013), 10.1126/science.1233151, <https://science.sciencemag.org/content/341/6148/1233151.full.pdf>.
- [4] A. Biffi, E. Montini, L. Lorioli, M. Cesani, F. Fumagalli, T. Plati, C. Baldoli, S. Martino, A. Calabria, S. Canale, F. Benedicenti, G. Vallanti, L. Biasco, S. Leo, N. Kabbara, G. Zanetti, W. B. Rizzo, N. A. L. Mehta, M. P. Cicalese, M. Casiraghi, J. J. Boelens, U. Del Carro, D. J. Dow, M. Schmidt, A. Assanelli, V. Neduva, C. Di Serio, E. Stupka, J. Gardner, C. von Kalle, C. Bordignon, F. Ciceri, A. Rovelli, M. G. Roncarolo, A. Aiuti, M. Sessa, and L. Naldini, *Lentiviral hematopoietic stem cell gene therapy benefits Metachromatic Leukodystrophy*, *Science* **341** (2013), 10.1126/science.1233158, <https://science.sciencemag.org/content/341/6148/1233158.full.pdf>.
- [5] D. Cesana, A. Calabria, L. Rudilosso, P. Gallina, F. Benedicenti, G. Spinozzi, G. Schirolli, A. Magnani, S. Acquati, F. Fumagalli, V. Calbi, M. Witzel, F. D. Bushman, A. Cantore, P. Genovese, C. Klein, A. Fischer, M. Cavazzana, E. Six, A. Aiuti, L. Naldini, and E. Montini, *Retrieval of vector integration sites from cell-free DNA*, *Nature Medicine* (2021).
- [6] D. B. Kohn, C. Booth, E. M. Kang, S.-Y. Pai, K. L. Shaw, G. Santilli, M. Armant, K. F. Buckland, U. Choi, S. S. De Ravin, M. J. Dorsey, C. Y. Kuo, D. Leon-Rico, C. Rivat, N. Izotova, K. Gilmour, K. Snell, J. X.-B. Dip, J. Darwish, E. C. Morris, D. Terrazas, L. D. Wang, C. A. Bauser, T. Paprotka, D. B. Kuhns, J. Gregg, H. E. Raymond, J. K. Everett, G. Honnet, L. Biasco, P. E. Newburger, F. D. Bushman, M. Grez, H. B. Gaspar, D. A. Williams, H. L. Malech, A. Galy, A. J. Thrasher, K. F. Buckland, C. A. Bauser, H. B. Gaspar, A. J. Thrasher, and the Net4CGD consortium, *Lentiviral gene therapy for X-linked chronic granulomatous disease*, *Nature Medicine* **26**, 200 (2020).
- [7] C. F. Magnani, G. Gaipa, F. Lussana, D. Belotti, G. Gritti, S. Napolitano, G. Mat-era, B. Cabiati, C. Buracchi, G. Borleri, *et al.*, *Sleeping Beauty-engineered CAR T*

- cells achieve antileukemic activity without severe toxicities*, The Journal of Clinical Investigation **130**, 6021 (2020).
- [8] S. Markt, S. Scaramuzza, M. P. Cicalese, F. Giglio, S. Galimberti, M. R. Lidonnici, V. Calbi, A. Assanelli, M. E. Bernardo, C. Rossi, *et al.*, *Intrabone hematopoietic stem cell gene therapy for adult and pediatric patients affected by transfusion-dependent  $\beta$ -thalassemia*, Nature Medicine **25**, 234 (2019).
- [9] S. Scala, L. Basso-Ricci, F. Dionisio, D. Pellin, S. Giannelli, F. A. Salerio, L. Leonardelli, M. P. Cicalese, F. Ferrua, A. Aiuti, *et al.*, *Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans*, Nature Medicine **24**, 1683 (2018).
- [10] T. S.-T. Yuo and T. A. Tseng, *The environmental product variety and retail rents on central urban shopping areas: A multi-stage spatial data mining method*, Environment and Planning B: Urban Analytics and City Science **0**, 2399808320966607 (0), <https://doi.org/10.1177/2399808320966607> .
- [11] Y. Fu, S. Wu, Y. Hu, T. Chen, Y. Zeng, C. Liu, and Q. Ou, *Mutational characterization of hbv reverse transcriptase gene and the genotype-phenotype correlation of antiviral resistance among chinese chronic hepatitis b patients*, Emerging Microbes & Infections **9**, 2381 (2020), PMID: 33124952, <https://doi.org/10.1080/22221751.2020.1835446> .
- [12] 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 **313**, 1550 (2015).
- [13] S. N. McNulty, P. R. Mann, J. A. Robinson, E. J. Duncavage, and J. D. Pfeifer, *Impact of reducing DNA input on Next-Generation Sequencing Library Complexity and Variant Detection*, The Journal of Molecular Diagnostics **22**, 720 (2020).
- [14] H. L. Sanders, *Marine benthic diversity: A comparative study*, The American Naturalist **102**, 243 (1968), <https://doi.org/10.1086/282541> .
- [15] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight, *Normalization and microbial differential abundance strategies depend upon data characteristics*, Microbiome **5**, 27 (2017).
- [16] P. J. McMurdie and S. Holmes, *Waste not, want not: Why rarefying microbiome data is inadmissible*, PLOS Computational Biology **10**, 1 (2014).
- [17] A. D. Willis, *Rarefaction, alpha diversity, and statistics*, Frontiers in Microbiology **10**, 2407 (2019).



- [18] R. H. Whittaker, *Evolution and measurement of species diversity*, TAXON **21**, 213 (1972), <https://onlinelibrary.wiley.com/doi/pdf/10.2307/1218190>.
- [19] K. P. Beule L, *Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities*. PeerJ (2020), 10.7717/peerj.9593.
- [20] E. Montini, D. Cesana, M. Schmidt, F. Sanvito, M. Ponzoni, C. Bartholomae, L. S. Sergi, F. Benedicenti, A. Ambrosi, C. Di Serio, C. Doglioni, C. von Kalle, and L. Naldini, *Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration*, Nature Biotechnology **24**, 687 (2006).
- [21] E. Montini, D. Cesana, M. Schmidt, F. Sanvito, C. C. Bartholomae, M. Ranzani, F. Benedicenti, L. S. Sergi, A. Ambrosi, M. Ponzoni, C. Doglioni, C. D. Serio, C. von Kalle, and L. Naldini, *The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy*, The Journal of Clinical Investigation **119**, 964 (2009).
- [22] R. Lu, N. Neff, S. Quake, and I. Weissman, *Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding*, Nature Biotechnology **29**, 928 (2011).
- [23] T. Nakamura and T. Omasa, *Optimization of cell line development in the GS-CHO expression system using a high-throughput, single cell-based clone selection system*, Journal of Bioscience and Bioengineering **120**, 323 (2015).
- [24] A. Gerrits, B. Dykstra, O. J. Kalmykova, K. Klauke, E. Verovskaya, M. J. C. Broekhuis, G. de Haan, and L. V. Bystrykh, *Cellular barcoding tool for clonal analysis in the hematopoietic system*, Blood **115**, 2610 (2010), <https://ashpublications.org/blood/article-pdf/115/13/2610/1324089/zh801310002610.pdf>.
- [25] M. A. Harkey, R. Kaul, M. A. Jacobs, P. Kurre, D. Bovee, R. Levy, and C. A. Blau, *Multi-arm high-throughput integration site detection: Limitations of LAM-PCR technology and optimization for clonal analysis*, Stem Cells and Development **16**, 381 (2007), pMID: 17610368, <https://doi.org/10.1089/scd.2007.0015>.
- [26] S. C. Schuster, *Next-generation sequencing transforms today's biology*, Nature Methods **5**, 16 (2008).
- [27] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P.

- McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, *Genome sequencing in microfabricated high-density picolitre reactors*, *Nature* **437**, 376 (2005).
- [28] U. Demkow and R. Ploski, *Clinical Applications for Next-Generation Sequencing* (Elsevier Science, 2015).
- [29] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, *Advanced sequencing technologies: methods and goals*, *Nature Reviews Genetics* **5**, 335 (2004).
- [30] C. Ledergerber and C. Dessimoz, *Base-calling for next-generation sequencing platforms*, *Briefings in Bioinformatics* **12**, 489 (2011).
- [31] F. Chang and M. M. Li, *Clinical application of amplicon-based next-generation sequencing in cancer*, *Cancer Genetics*, *Cancer Genetics* **206**, 413 (2013).
- [32] A. Kohlmann, V. Grossmann, and T. Haferlach, *Integration of Next-Generation Sequencing into clinical practice: Are we there yet?* *Seminars in Oncology* **39**, 26 (2012), molecular Pathogenesis of Hematologic Malignancies.
- [33] A. S. Gargis, L. Kalman, D. P. Bick, C. da Silva, D. P. Dimmock, B. H. Funke, S. Gowrisankar, M. R. Hegde, S. Kulkarni, C. E. Mason, R. Nagarajan, K. V. Voelkerding, E. A. Worthey, N. Aziz, J. Barnes, S. F. Bennett, H. Bisht, D. M. Church, Z. Dimitrova, S. R. Gargis, N. Hafez, T. Hambuch, F. C. L. Hyland, R. A. Luna, D. MacCannell, T. Mann, M. R. McCluskey, T. K. McDaniel, L. M. Ganova-Raeva, H. L. Rehm, J. Reid, D. S. Campo, R. B. Resnick, P. G. Ridge, M. L. Salit, P. Skums, L.-J. C. Wong, B. A. Zehnbaauer, J. M. Zook, and I. M. Lubin, *Good laboratory practice for clinical next-generation sequencing informatics pipelines*, *Nature Biotechnology* **33**, 689 (2015).
- [34] L. Biasco, D. Pellin, S. Scala, F. Dionisio, L. Basso-Ricci, L. Leonardelli, S. Scaramuzza, C. Baricordi, F. Ferrua, M. Cicalese, S. Giannelli, V. Neduva, D. Dow, M. Schmidt, C. Von Kalle, M. Roncarolo, F. Ciceri, P. Vicard, E. Wit, C. Di Serio, L. Naldini, and A. Aiuti, *In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases*, *Cell Stem Cell* **19**, 107 (2016).
- [35] C. Wu, B. Li, R. Lu, S. Koelle, Y. Yang, A. Jares, A. Krouse, M. Metzger, F. Liang, K. LorÃ©, C. Wu, R. Donahue, I. Chen, I. Weissman, and C. Dunbar, *Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells*, *Cell Stem Cell* **14**, 486 (2014).

- [36] F. Mazurier, O. I. Gan, J. L. McKenzie, M. Doedens, and J. E. Dick, *Lentivector-mediated clonal tracking reveals intrinsic heterogeneity in the human hematopoietic stem cell compartment and culture-induced stem cell impairment*, *Blood* **103**, 545 (2004), <https://ashpublications.org/blood/article-pdf/103/2/545/1694234/zh800204000545.pdf>.
- [37] L. Biasco, M. Rothe, J. W. Schott, and A. Schambach, *Integrating vectors for gene therapy and clonal tracking of engineered hematopoiesis*, *Hematology/Oncology Clinics*, *Hematology/Oncology Clinics* **31**, 737 (2017).
- [38] C. E. Shannon, *A mathematical theory of communication*, *The Bell System Technical Journal* **27**, 623 (1948).
- [39] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, *The Annals of Mathematical Statistics* **22**, 79 (1951).
- [40] I. Carboni, P. Fattorini, C. PreviderÅš, S. S. Ciglieri, S. Iozzi, A. Nutini, E. Contini, C. Pescucci, F. Torricelli, and U. Ricci, *Evaluation of the reliability of the data generated by next generation sequencing from artificially degraded DNA samples*, *Forensic Science International: Genetics Supplement Series* **5**, e83 (2015).
- [41] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth*, *Bioinformatics* **28**, 1420 (2012), <https://academic.oup.com/bioinformatics/article-pdf/28/11/1420/742285/bts174.pdf>.
- [42] J. Pereira-Marques, A. Hout, R. M. Ferreira, M. Weber, I. Pinto-Ribeiro, L.-J. van Doorn, C. W. Knetsch, and C. Figueiredo, *Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis*, *Frontiers in Microbiology* **10**, 1277 (2019).
- [43] A. Hahn, A. Sanyal, G. F. Perez, A. M. Colberg-Poley, J. Campos, M. C. Rose, and M. PÃ©rez-Losada, *Different next generation sequencing platforms produce different microbial profiles and diversity in cystic fibrosis sputum*, *Journal of Microbiological Methods* **130**, 95 (2016).
- [44] J. Sabina and J. H. Leamon, *Bias in whole genome amplification: Causes and considerations*, in *Whole Genome Amplification: Methods and Protocols*, edited by T. Kroneis (Springer New York, New York, NY, 2015) pp. 15–41.
- [45] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, *Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq illumina sequencing platform*, *Applied and Environmental Microbiology* **79**, 5112 (2013), <https://aem.asm.org/content/79/17/5112.full.pdf>.

- [46] Y. Nakayama, H. Yamaguchi, N. Einaga, and M. Esumi, *Pitfalls of DNA quantification using DNA-binding fluorescent dyes and suggested solutions*, PLOS ONE **11**, 1 (2016).
- [47] J. D. Robin, A. T. Ludlow, R. LaRanger, W. E. Wright, and J. W. Shay, *Comparison of DNA quantification methods for next generation sequencing*, Scientific Reports **6**, 24067 (2016).
- [48] N. Pya and S. N. Wood, *Shape constrained additive models*, Statistics and Computing **25**, 543 (2015).
- [49] K. Bollaerts, P. H. Eilers, and I. Van Mechelen, *Simple and multiple P-splines regression with shape constraints*, British Journal of Mathematical and Statistical Psychology **59**, 451 (2006).
- [50] A. Brezger and W. J. Steiner, *Monotonic regression based on bayesian p-splines: An application to estimating price response functions from store-level scanner data*, Journal of Business & Economic Statistics **26**, 90 (2008).
- [51] F. N. Fritsch and R. E. Carlson, *Monotone piecewise cubic interpolation*, SIAM Journal on Numerical Analysis **17**, 238 (1980).
- [52] M. C. Meyer, *Inference using shape-restricted regression splines*, The Annals of Applied Statistics **2**, 1013 (2008).
- [53] M. C. Meyer, *A framework for estimation and inference in generalized additive models with shape and order restrictions*, Statistical Science **33**, 595 (2018).
- [54] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor, *A practical guide to splines*, Vol. 27 (Springer-Verlag New York, 1978).
- [55] X. Liao and M. C. Meyer, *coneproj: An R package for the Primal or Dual Cone Projections with Routines for Constrained Regression*, Journal of Statistical Software **61**, 1 (2014).
- [56] K. P. Burnham, D. R. Anderson, and K. P. Huyvaert, *AIC model selection and multi-model inference in behavioral ecology: some background, observations, and comparisons*, Behavioral Ecology and Sociobiology **65**, 23 (2011).
- [57] F. Benedicenti, A. Calabria, D. Cesana, A. Albertini, E. Tenderini, G. Spinozzi, V. Neduva, A. Richard, M. Brugman, D. Dow, *et al.*, *Sonication linker mediated-pcr (slim-pcr), an efficient method for quantitative retrieval of vector integration sites*, in *HUMAN GENE THERAPY*, Vol. 30 (MARY ANN LIEBERT, INC 140 HUGUENOT STREET, 3RD FL, NEW ROCHELLE, NY 10801 USA, 2019) pp. A214–A215.
- [58] G. Spinozzi, A. Calabria, S. Brasca, S. Beretta, I. Merelli, L. Milanese, and E. Montini, *VISPA2: a scalable pipeline for high-throughput identification and annotation of vector integration sites*, BMC Bioinformatics **18**, 520 (2017).

- [59] C. C. Berry, N. A. Gillet, A. Melamed, N. Gormley, C. R. M. Bangham, and F. D. Bushman, *Estimating abundances of retroviral insertion sites from DNA fragment length data*, *Bioinformatics* **28**, 755 (2012).
- [60] F. Benedicenti, A. Calabria, D. Cesana, A. Albertini, E. Tenderini, G. Spinozzi, V. Neduva, A. Richard, M. Brugman, D. Dow, *et al.*, *Sonication linker mediated-PCR (SLiM-PCR), an efficient method for quantitative retrieval of vector integration sites*, in *Human Gene Therapy*, Vol. 30 (2019) pp. A214–A215.

# SUMMARY

## AIM OF THE WORK

Mathematical models of haematopoiesis provide relevant insights to support the design of novel gene therapy strategies [1, 2]. Clonal tracking is a recent high-throughput technology that allows to calibrate such mathematical models by quantitatively tracing the evolution of haematopoietic stem cells [3]. In particular, cell differentiation networks and clonality are key surveillance studies in gene therapy [4, 5]. While cell differentiation networks describe the hierarchical relationships underlying haematopoiesis, clonality is aimed at quantifying the clonal population complexity (clonal diversity) and to detect possible therapy side effects, such as events of clonal dominance [6].

In this work we investigate cell differentiation using stochastic quasi-reaction networks combined with extended Kalman filtering, leading to our proposed Kalman Reaction Networks framework Karen [7]. This framework takes into account that typical clonal tracking data suffers from missing cell types and false negative errors. It consists in a continuous-discrete state space model with a stochastic reaction network describing the dynamics, coupled with a linear Gaussian measurement model that links the noisy observations to the underlying states. In particular, (i) we develop an expectation-maximization inference algorithm; (ii) we extensively test our method on several simulation studies, including a direct comparison with the state-of-the-art methods, and finally (iii) we apply our framework to five in-vivo clonal tracking datasets to compare different biologically plausible models of cell differentiation.

Subsequently, in order to detect possible adverse events of clonal dominance we combine stochastic quasi-reaction networks with random-effects (RestoreNet) [8]. Our proposed method consists in a set of biochemical reactions translated into a generalized mixed-effects model including random effects on the clones for the dynamic parameters. The unknown

parameters are estimated with a tailor-made expectation maximization algorithm. In particular, (i) we validate our inference method in several synthetic studies, (ii) we compare its performance with the state of the art approaches, and finally (iii) we apply it in two in-vivo clonal tracking studies. Finally, to objectively measure clonal complexity, we propose a method that combines shape-constrained splines (SCS) with the Shannon entropy index [9]. Our proposed method leverages the effect of technical artefacts from the Shannon entropy index, thus providing an unbiased measure of clonal diversity. Our SCS-rescaling method was first validated in a specifically designed in-vitro assay, and then used to objectively evaluate the impact of vector genotoxicity on the entropy decays of tumor prone mice.

## CHAPTERS CONTRIBUTION

The content of the thesis can be summarised as follows:

- **Chapter 1** first introduces the research questions that we address, and provides relevant biology background. Subsequently, it follows a brief discussion on the state-of-the-art methods and our proposed approaches.
- **Chapter 2** focuses on our Kalman reaction network framework *Karen* aimed at inferring cell differentiation networks from typical clonal tracking data. We first introduce and validate our method, then we apply it on five in-vivo clonal tracking studies.
- **Chapter 3** introduces our proposed random-effects stochastic reaction networks *RestoreNet* to detect possible adverse events of clonal dominance in gene therapy clonal tracking studies. After validating our framework with several synthetic studies, we analyse two in-vivo models of haematopoiesis.
- **Chapter 4** focuses on our proposed shape-constraint rescaled Shannon entropy index to provide an artefacts-free measure of clonal diversity. We first validate our SCS method on a specifically designed in-vitro clonal tracking study, then we apply it on a genotoxicity pre-clinical study.

## FUTURE DIRECTIONS

The main approximation in both the basal LLA and random-effects RestoreNet formulations is the piece-wise linearity of the process. That is, in both cases we consider first a local linear approximation of the Ito equation, which then we use to infer the process parameters either with or without random-effects. Although the linearity assumption makes all the computations easier, this approximation becomes poor as the time lag increments (the  $\Delta ts$ ) of the collected data increase. This can be addressed by introducing in the likelihood higher-order approximation terms than the ones considered by the Euler-Maruyama method. The Milstein approximation is a possible choice. Another, completely different, approach is to employ extended Kalman filtering (EKF) which is suitable for non-linear state space formulations, as we did for Karen. Furthermore, RestoreNet cannot consider false-negative errors or missing values of clonal tracking data, as Karen does. Also for this limitation, an EKF formulation could be a possible extension of RestoreNet.

Although the Gaussian assumption of Karen makes the analytical formulations of the likelihoods explicitly available, this approximation may become poor when the data contains outliers or shows non-Gaussian behaviors. A distribution-free approach, such as the Kernel Kalman Rule, could be a possible extension [10, 11]. Besides, both frameworks RestoreNet and Karen consider reaction rates constant for the whole study period. Extensions that allow for modeling reaction rates as spline functions of clinically relevant variables are within reach and will be the goal of future research.

Furthermore, since our SCS approach for rescaling Shannon entropy is regression-based, its main limitation is related to the available sample size, a potential issue when defining a spline basis. Thus, the number of knots of the splines should be chosen carefully. Our model averaging approach allows to rank the impact of the confounders according to their approximated inclusion probabilities by means of the Bayesian Information Criterion. A more precise method to estimate the marginal likelihood, such as Laplace Integration or Bayesian thermodynamic integration, can be a possible improvement.



## REFERENCES

- [1] H. Kawamoto, H. Wada, and Y. Katsura, *A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model*, *International Immunology* **22**, 65 (2010).
- [2] T. Friedmann and R. Roblin, *Gene therapy for human genetic disease?* *Science* **175**, 949 (1972), <https://science.sciencemag.org/content/175/4025/949.full.pdf>.
- [3] E. Six, A. Guilloux, A. Denis, A. Lecoules, A. Magnani, R. Vilette, F. Male, N. Cagnard, M. Delville, E. Magrin, *et al.*, *Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs*, *Blood* **135**, 1219 (2020).
- [4] O. S. Kustikova, A. Wahlers, K. Kühlcke, B. Stähle, A. R. Zander, C. Baum, and B. Fehse, *Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population*, *Blood* **102**, 3934 (2003).
- [5] B. Fehse, O. Kustikova, M. Bubenheim, and C. Baum, *Pois (s) on—it's a question of dose...*, *Gene Therapy* **11**, 879 (2004).
- [6] D. Bryder, D. J. Rossi, and I. L. Weissman, *Hematopoietic stem cells: the paradigmatic tissue-specific stem cell*, *The American journal of pathology* **169**, 338 (2006).
- [7] L. Del Core, D. Pellin, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data*, *bioRxiv* (2022).
- [8] L. Del Core, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic inference of clonal dominance in gene therapy studies*, *bioRxiv* (2022).
- [9] L. Del Core, D. Cesana, P. Gallina, Y. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. Grzegorzcyk, *Normalization of clonal diversity in gene therapy studies using shape constrained splines*, *Scientific Reports* **12**, 1 (2022).
- [10] 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).
- [11] G. H. W. Gebhardt, A. Kupcsik, and G. Neumann, *The kernel kalman rule*, *Machine Learning* **108**, 2113 (2019).

# SAMENVATTING

## DOEL VAN HET WERK

Wiskundige modellen van hematopoëse bieden relevante inzichten ter ondersteuning van het ontwerp van nieuwe genterapiestrategieën [1, 2]. Clonal tracking is een recente high-throughput technologie die het mogelijk maakt om dergelijke wiskundige modellen te kalibreren door de evolutie van hematopoëtische stamcellen kwantitatief te volgen [3]. Met name celdifferentiatienetwerken en klonaliteit zijn belangrijke elementen in surveillancestudies in genterapie [4, 5]. Terwijl celdifferentiatienetwerken de hiërarchische relaties beschrijven die ten grondslag liggen aan hematopoëse, is klonaliteit gericht op het kwantificeren van de complexiteit van de klonale populatie (klonale diversiteit) en het detecteren van mogelijke therapie-bijwerkingen, zoals gebeurtenissen van klonale dominantie [6].

In dit werk onderzoeken we celdifferentiatie met behulp van stochastische quasi-reactienetwerken gecombineerd met uitgebreide Kalman-filtering, wat leidt tot het door ons voorgestelde Kalman Reaction Networks raamwerk Karen [7]. Dit raamwerk houdt er rekening mee dat typische klonale trackinggegevens te kampen hebben met ontbrekende celtypen en fout-negatieven. Het bestaat uit een continu-discreet toestandsruimtemodel met een stochastisch reactienetwerk dat de dynamiek beschrijft, gekoppeld aan een lineair Gaussiaans meetmodel dat de onzekere waarnemingen koppelt aan de onderliggende toestanden. In het bijzonder (i) ontwikkelen we een verwachtingsmaximaliserend inferentie-algoritme; (ii) we testen onze methode uitgebreid op verschillende simulatiestudies, waaronder een directe vergelijking met de state-of-the-art methoden, en tot slot (iii) passen we ons raamwerk toe op vijf in-vivo klonale tracking-datasets om verschillende biologisch plausibele datasets te vergelijken modellen van celdifferentiatie.

Om mogelijke nadelige gebeurtenissen van klonale dominantie te de-

tecteren, combineren we vervolgens stochastische quasi-reactienetwerken met hiërarchische effecten (RestoreNet) [8]. Onze voorgestelde methode bestaat uit een reeks biochemische reacties vertaald in een algemeen mixed-effects-model inclusief hiërarchische effecten op de klonen voor de dynamische parameters. De onbekende parameters worden geschat met een op maat gemaakt algoritme voor verwachtingsmaximalisatie. In het bijzonder (i) valideren we onze inferentiemethode in verschillende synthetische studies, (ii) vergelijken we de prestaties ervan met de nieuwste benaderingen, en tot slot (iii) passen we deze toe in twee in-vivo klonale tracking-studies. Ten slotte stellen we, om de klonale complexiteit objectief te meten, een methode voor die shape-constrained splines (SCS) combineert met de Shannon-entropie-index [9]. Onze voorgestelde methode maakt gebruik van het effect van technische artefacten van de Shannon-entropie-index, en biedt zo een onpartijdige maatstaf voor klonale diversiteit. Onze SCS-herschalingmethode is allereerst gevalideerd in een specifiek ontworpen in-vitro-assay en vervolgens gebruikt om de impact van vectorgenotoxiciteit op het entropieverval van tumorgevoelige muizen objectief te evalueren.

## HOOFDSTUKBIJDRAGE

De inhoud van het proefschrift kan als volgt worden samengevat:

- **Hoofdstuk 1** introduceert eerst de onderzoeksvragen die we behandelen, en geeft relevante biologische achtergrondinformatie. Vervolgens volgt een korte discussie over de state-of-the-art methoden en onze voorgestelde benaderingen.
- **Hoofdstuk 2** richt zich op ons Kalman-reactienetwerkraamwerk. Het is gericht op het afleiden van celdifferentiatienetwerken uit typische klonale volgggegevens. We introduceren en valideren eerst onze methode, daarna passen we deze toe op vijf in-vivo klonale trackingstudies.
- **Hoofdstuk 3** introduceert onze voorgestelde random-effects stochastische reactienetwerken RestoreNet om mogelijke bijwerkingen van klonale dominantie te detecteren in onderzoeken naar klonale

tracking van gentherapie. Na validatie van ons raamwerk met verschillende synthetische studies, analyseren we twee in-vivo modellen van hematopoëse.

- **Hoofdstuk 4** richt zich op onze voorgestelde vormbeperking herschaalde Shannon entropie-index om een artefactvrije maatstaf van klonale diversiteit te bieden. We valideren eerst onze SCS-methode op een specifiek ontworpen in-vitro klonale tracking-studie, daarna passen we deze toe op een preklinische genotoxiciteitsstudie.

## TOEKOMSTIGE RICHTINGEN

4

De belangrijkste benadering in zowel de basale LLA- als de random-effects RestoreNet-formuleringen is de stuksgewijze lineariteit van het proces. Dat wil zeggen, in beide gevallen beschouwen we eerst een lokale lineaire benadering van de Ito-vergelijking, die we vervolgens gebruiken om de procesparameters met of zonder hierarchische effecten af te leiden. Hoewel de aanname van lineariteit alle berekeningen eenvoudiger maakt, wordt deze benadering slecht naarmate de tijdsvertraging toeneemt (de  $\Delta ts$ ) van de verzamelde gegevens. Dit kan worden verholpen door in de waarschijnlijkheid benaderingstermen van hogere orde in te voeren dan degene die worden overwogen door de Euler-Maruyama-methode. De Milstein-benadering is een mogelijke keuze. Een andere, geheel andere benadering is het gebruik van uitgebreide Kalman-filtering (EKF) die geschikt is voor niet-lineaire formuleringen van toestandsruimten, zoals we deden voor Karen. Bovendien kan RestoreNet geen fout-negatieven of ontbrekende waarden van klonale trackinggegevens in overweging nemen, zoals Karen doet. Ook voor deze beperking zou een EKF-formulering een mogelijke uitbreiding van RestoreNet kunnen zijn.

Hoewel de Gaussiaanse aanname van Karen de analytische formuleringen van de waarschijnlijkheden expliciet beschikbaar maakt, kan deze benadering slecht worden wanneer de gegevens uitschieters bevatten of niet-Gaussiaans gedrag vertonen. Een distributievrije aanpak, zoals de Kernel Kalman Rule, zou een mogelijke uitbreiding [10, 11] kunnen zijn. Bovendien beschouwen beide frameworks RestoreNet en Karen de reactiesnelheden als constant voor de hele studieperiode. Uitbreidingen die het

mogelijk maken om reactiesnelheden te modelleren als spline-functies van klinisch relevante variabelen liggen binnen handbereik en zullen het doel zijn van toekomstig onderzoek.

Bovendien, aangezien onze SCS-benadering voor het herschalen van Shannon-entropie op regressie is gebaseerd, heeft de belangrijkste beperking te maken met de beschikbare steekproefomvang, een mogelijk probleem bij het definiëren van een spline-basis. Het aantal knopen van de splines moet dus zorgvuldig worden gekozen. Onze modelmiddelingsbenadering maakt het mogelijk om de impact van de confounders te rangschikken op basis van hun geschatte inclusiekansen door middel van het Bayesiaanse informatiecriterium. Een meer precieze methode om de marginale waarschijnlijkheid in te schatten, zoals Laplace-integratie of Bayesiaanse thermodynamische integratie, kan een mogelijke verbetering zijn.

**BIBLIOGRAFIE**

- [1] H. Kawamoto, H. Wada, and Y. Katsura, *A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model*, *International Immunology* **22**, 65 (2010).
- [2] T. Friedmann and R. Roblin, *Gene therapy for human genetic disease?* *Science* **175**, 949 (1972), <https://science.sciencemag.org/content/175/4025/949.full.pdf>.
- [3] E. Six, A. Guilloux, A. Denis, A. Lecoules, A. Magnani, R. Vilette, F. Male, N. Cagnard, M. Delville, E. Magrin, *et al.*, *Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs*, *Blood* **135**, 1219 (2020).
- [4] O. S. Kustikova, A. Wahlers, K. Kühlcke, B. Stähle, A. R. Zander, C. Baum, and B. Fehse, *Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population*, *Blood* **102**, 3934 (2003).
- [5] B. Fehse, O. Kustikova, M. Bubenheim, and C. Baum, *Pois (s) on—it's a question of dose...*, *Gene Therapy* **11**, 879 (2004).
- [6] D. Bryder, D. J. Rossi, and I. L. Weissman, *Hematopoietic stem cells: the paradigmatic tissue-specific stem cell*, *The American journal of pathology* **169**, 338 (2006).
- [7] L. Del Core, D. Pellin, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data*, *bioRxiv* (2022).
- [8] L. Del Core, M. A. Grzegorzcyk, and E. C. Wit, *Stochastic inference of clonal dominance in gene therapy studies*, *bioRxiv* (2022).
- [9] L. Del Core, D. Cesana, P. Gallina, Y. Secanechia, L. Rudilosso, E. Montini, E. C. Wit, A. Calabria, and M. Grzegorzcyk, *Normalization of clonal diversity in gene therapy studies using shape constrained splines*, *Scientific Reports* **12**, 1 (2022).
- [10] 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).
- [11] G. H. W. Gebhardt, A. Kupcsik, and G. Neumann, *The kernel kalman rule*, *Machine Learning* **108**, 2113 (2019).



# ACKNOWLEDGEMENTS

This work would not have been possible without the exceptional mentorship of my supervisors Prof. Dr. Marco Grzegorzczak and Prof. Dr. Ernst Wit. Thank you Marco and Ernst for your invaluable support, guidance, encouragement, and patience. I had great pleasure working with you, in a very nice and constructive environment. You were both of high inspiration and I learnt a lot from you. Thank you also for giving me the opportunity to attend a lot of conferences and scientific missions abroad.

I am really grateful to the members of the assessment committee, Prof. Dr. Dirk Husmeier, Prof. Dr. Casper Albers, and Prof. Dr. Clelia Di Serio, for their encouraging statements and constructive feedback. I extend my sincere thanks to Prof. Dr. Danilo Pellin for his contribution on chapters 2 and 3. I am really grateful to the research group members, the University staff, my office mates, the housemates, and all the friends for their warm welcome during my stay in Groningen and the visits in Lugano. Many thanks to Prof. Dr. Cristóbal Bertoglio and Dr. Victor Arturo Bernal for being my paranymphs during my PhD defence. I would like to acknowledge the generous support of the funding institutes: the European Cooperation for Statistics of Network data science (COSTNET), Fondazione Telethon, and the Swiss National Science Foundation (SNSF).

I really appreciate the love and support of my parents, my brothers, all my family members and friends. Many thanks to all of you my dears, your belief in me contributed to keeping my spirits, motivation and research curiosity high. Special thanks go to my partner: Claudia, my darling, thank you so much for your love, support, encouragement, patience, and for always being there for me. I am also very grateful to your family for being kind, supportive and welcoming. I would like to thank Prof. Dr. Gianfausto Salvadori, who introduced me to the research world, and follows my work with great enthusiasm. Finally, I express my gratitude to everyone who attended my PhD defence, one of the most important days of my life that marks the completion of this incredible and amazing journey.





# BIOGRAPHY

Luca Del Core was born on 8 September 1989 in Taranto, Italy. After obtaining his Bachelor (B.Sc.) and Master (M.Sc.) degrees in Mathematics from the University of Salento (Lecce, Italy) he was a software developer from 2016 to 2017 at a consulting company based in Milan, Italy. From 2017 to 2021 he was a researcher at the San Raffaele Telethon Institute for Gene Therapy, Milan, Italy. Luca met his prospective PhD supervisors, Prof. Ernst Wit and Prof. Marco Grzegorzczuk, at a COSTNET workshop on statistical network science that took place in Milan in early 2018, and he started his PhD project on September 2018 at the Bernoulli Institute, University of Groningen, The Netherlands. Luca's research interests include forward and inverse modelling of stochastic dynamical systems, state-space modelling and computational statistics, with a particular focus on systems biology. Recently, Luca started his postdoctoral research programme at the School of Mathematical Sciences, University of Nottingham, United Kingdom, where his research focuses on stochastic modelling with applications in cardiac electrophysiology.