


RESEARCH

Open Access



Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking

Danilo Pellin^{1,3}, Luca Biasco^{7,8}, Alessandro Aiuti^{2,6}, Maria Clelia Di Serio³ and Ernst C. Wit^{4,5*} 

*Correspondence: wite@usi.ch

⁴Institute of Computational Science, Università della Svizzera italiana, Via G. Buffi 13, 6900 Lugano, Switzerland

⁵Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands

Full list of author information is available at the end of the article

Abstract

Background: During their lifespan, stem- or progenitor cells have the ability to differentiate into more committed cell lineages. Understanding this process can be key in treating certain diseases. However, up until now only limited information about the cell differentiation process is known.

Aim: The goal of this paper is to present a statistical framework able to describe the cell differentiation process at the single clone level and to provide a corresponding inferential procedure for parameters estimation and structure reconstruction of the differentiation network.

Approach: We propose a multidimensional, continuous-time Markov model with density-dependent transition probabilities linear in sub-population sizes and rates. The inferential procedure is based on an iterative calculation of approximated solutions for two systems of ordinary differential equations, describing process moments evolution over time, that are analytically derived from the process' master equation. Network sparsity is induced by adding a SCAD-based penalization term in the generalized least squares objective function.

Results: The methods proposed here have been tested by means of a simulation study and then applied to a data set derived from a gene therapy clinical trial, in order to investigate hematopoiesis in humans, *in-vivo*. The hematopoietic structure estimated contradicts the classical dichotomy theory of cell differentiation and supports a novel myeloid-based model recently proposed in the literature.

Keywords: Cell differentiation tree, Penalized inference, Method-of-moments, Euler's method

Introduction

Over the past decade, gene therapy (GT) has proved its potential as a next-generation therapy for many diseases that were untreatable by conventional therapies (Naldini 2011). GT can be used to treat cellular defects due to a mutated gene by providing a fully functional copy of it or by equipping target cells with a new cellular function through genetic engineering. Most clinical approaches are based on the delivery of exogenous DNA molecules by viral vectors using retrovirus- or lentivirus-derived systems. The advent of next-generation sequencing (NGS) platforms (i.e. Roche/454 pyrosequencing and Illumina sequencing technology) substantially improved the accuracy and resolution

of viral integration site (IS) analyses (Biasco et al. 2011). This technological progress has resulted in the availability of IS data from single transduction experiments that are two orders of magnitude denser than previously possible. As a result, IS research has diversified from characterizing the mechanisms driving the virus integration process and its interactions with the host cell genome to investigating other biological questions. For example, retrovirus IS distribution over the genome has been used as an indicator of active gene enhancers and regulatory regions, involved in hematopoietic stem cell commitment (Romano et al. 2016) or as a tool to follow individual cell fate *in-vivo* (Biasco et al. 2016; Scala et al. 2018).

In clinical settings, GT has been successfully used, for example, to treat hematological diseases such as Wiskott-Aldrich Syndrome (WAS), an inherited immunodeficiency caused by mutations in the gene encoding WASP protein (Aiuti et al. 2013). In this context, to ensure life-long curative potential and limit possible treatment side effect, hematopoietic stem/progenitor cells (HSC) are harvested from patients' bone marrow (BM), corrected by means of virus-based manipulation and then re-infused to patients. Treated cells acquire a unique label, represented by IS genomic coordinates, and this label will be inherited by all cellular offspring generated by both duplication or differentiation events in more committed cell types. In other words, IS can be used as a molecular marker to track individual HSC and evolution.

The set of cells, among all lineages under investigation, sharing a specific genomic marker, and therefore deriving from a common HSC ancestor, is defined as a *clone*. The analysis of the *in-vivo* clone evolution by means of periodic IS analysis performed on patients' BM or peripheral blood (PB) sample, is called *clonal tracking*. In the experimental data analysed in this paper, 15 different cell sub-populations (named also cell types or lineages), distributed along the hematopoietic hierarchy, have been collected from three patients affected by WAS during their first three years after GT treatment. Lineage-specific population sizes have been measured by means of reads count values (Biasco et al. 2016), returned by NGS platforms. Given the amount of lineages, samples and patients, this study provides a unique opportunity to reveal novel insight into human hematopoiesis.

In the literature, various mathematical approaches for the quantitative analysis of hematopoiesis have been proposed. For example, stochastic models for simplified hierarchical structures reduced to two categories, stem cells reserve and contributing clones, have been developed in (Abkowitz et al. 1990; Catlin et al. 2001) and in (Becker et al. 1963) for cat and mouse models, respectively. In (Marciniak-Czochra and Stiehl 2013) authors proposed a more complex multi-compartment model described by a set of deterministic functions, aimed at evaluating the mechanisms of regulation governing reconstitution after HSCs transplantation in humans. However, in all these approaches, authors compared how known and alternative hierarchies support various types of experimental data, rather than making hierarchy estimation a goal of the inferential procedure itself. In this respect, an interesting discovery-oriented approach applied to GT for WAS data has been recently proposed in (Scala et al. 2018). By means of additive Bayesian network modeling of IS detection, simultaneous structural learning and associations estimation have been performed, in order to investigate differences in lineages dependence at early and late phase after treatment. Dichotomizing IS measurements is motivated by the necessity to alleviate the noisy nature of IS analysis, due to technical factors such as DNA

amplification and sequencing, but it also discards valuable information about clone size dynamics. Standard Bayesian network algorithms are in principle capable of both parameter inference and model structure learning. These methods are based on modeling conditionally independence among nodes, usually using a contingency table parametrization in the case of binary variables. In the context of clonal tracking data, the results derived from such a modeling approaches are difficult to interpret from a biological perspective. In general, the metrics adopted by learning methods like mutual information make no distinction between positive and negative association. However, they have opposite biological interpretations, where a positive dependence suggests a differentiation path connecting two nodes, whereas a negative one supports the hypothesis that the nodes belong to alternative differentiation branches. In addition, coefficients measuring dependence strength are difficult to compare among each other and are particularly sensitive to a significant amount of stochastic variation of the process itself and other effects, such as saturation.

The goals of this paper are to propose a statistical framework able to model the cell differentiation process (CDP) measured at single clone resolution, to provide an inferential procedure able to perform both process parameters estimation and model reconstruction and finally, to investigate the hematopoietic process in humans. Our proposal derives from the definition of a novel generative stochastic process for clonal tracking data, defined over a network of lineages (nodes). The model is able to properly address the stochastic nature of the cell differentiation process and given a specific setting for the network parameters, generate complete evolution of clone size dynamics among all lineages. Although the application in this paper is clearly geared towards the cell differentiation process, the methodology underlying the analysis is completely general and can be applied to any stochastic process that involves differentiation, replication and extinction, such as political systems, corporate organizational development or even insect colonies.

A description of the continuous time, density-dependent Markov model for CDP, along with the underlying assumptions, are detailed in “[Stochastic cell differentiation model](#)” section. In “[Approximate generalized method-of-moments estimation](#)” section an efficient generalized least square estimation procedure, relying on first order Euler’s method approximation for the evolution of first and second order process moments is derived. In “[Model selection](#)” section a sparsity-inducing penalty term is incorporated in the estimation procedure in order to reconstruct the differentiation structure of the systems under investigation. A overview of the inference algorithm is given in “[Schematic overview of the inferential procedure](#)” section. In “[Simulation study](#)” section the performance of our proposal is verified by means of a simulation study. In “[Investigating human hematopoiesis in vivo](#)” section the experimental data previously mentioned are described in more details and then analyzed. Finally, our findings are discussed in “[Discussion](#)” section. “[Conclusion](#)” section is dedicated to final considerations, possible extensions and future directions to improve the methodology.

Stochastic cell differentiation model

For each clone or integration site l a CDP can be defined as a N -dimensional Markov process, $\mathbf{X}^l(t)$, such that each element $X_i^l(t)$ of $\mathbf{X}^l(t) = (X_1^l(t), \dots, X_N^l(t))$, corresponds to the number of cells (counts) of type (or lineage) C_i , $i = 1, \dots, N$ present in clone l at time t . For notational convenience, we will drop the explicit dependence on each individual clone

l in this section, as clones can be thought of as independent copies of each other. Given an initial state vector, \mathbf{x}_0 , the process evolves according to a random sequence of events, divided in three categories: duplications, deaths and differentiations. Individual cells are assumed to be independent from each other and cells belonging to the same lineage are assumed to obey the same law. Single event rates are assumed to be non-negative and constant over time. A graphical representation of cellular events is available in Fig. 1a and a detailed description follows.

Cell duplication: $1C_i \xrightarrow{\alpha_i} 2C_i$

The *net effect*, or *process state change* induced by the duplication of a cell of type C_i is the increment, of one unit, of C_i population size. Duplication rates, $\alpha = (\alpha_i, i = 1, \dots, N)$, correspond approximately to the probability that a generic cell of type C_i undergoes duplication, in a time unit. The transition probability associated to a duplication event in lineage C_i occurring in time interval $[t, t + \Delta t)$ for process $X(t)$ being in state \mathbf{x}_t , is given by:

$$P(X_i(t + \Delta t) = x_{i,t} + 1 | X_i(t) = x_{i,t}) \approx x_{i,t} \alpha_i \Delta t.$$

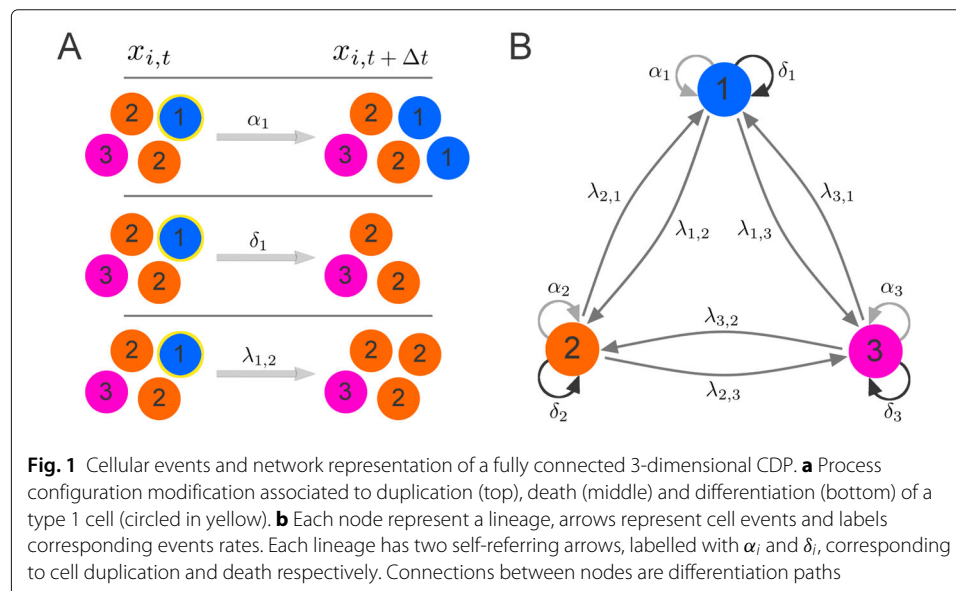
Cell death: $C_i \xrightarrow{\delta_i} \emptyset$

The *net effect* corresponding to a single death event is the decrease of one unit in the C_i population size. Death rates in vector $\delta = (\delta_i, i = 1, \dots, N)$, are the probabilities that a generic cell of type C_i dies, in a time unit. The transition probability associated to the generic death event in a time interval $[t, t + \Delta t)$ is:

$$P(X_i(t + \Delta t) = x_{i,t} - 1 | X_i(t) = x_{i,t}) \approx x_{i,t} \delta_i \Delta t.$$

Cell differentiation: $C_i \xrightarrow{\lambda_{i,j}} C_j$

According to the biological literature, it is possible to distinguish between two different models of differentiation: asymmetric cell division and signalling induced differentiation. In the first case, cell division gives rise to two daughter cells with distinct features



and fates. In the second case, differentiation is a process induced by a set of cell-to-cell signals, leading to conformational and receptor modifications and is not coupled with a duplication event. The kind of differentiation considered in this work consists in the transition of a single cell from lineage C_i to lineage C_j , and is equivalent to assume signalling induced differentiation. Similarly to duplication and death events, event rates $\lambda = (\lambda_{i,j}, i, j = 1, \dots, N, i \neq j)$ measure the probabilities that a generic cell of type C_i undergoes differentiation into C_j , in a time unit. The transition probability associated to a single differentiation event $C_i \rightarrow C_j$ in the time interval $[t, t + \Delta t)$, for a system being in state \mathbf{x}_t is given by:

$$P(X_i(t + \Delta t) = x_{i,t} - 1, X_j(t + \Delta t) = x_{i,t} + 1 | X_i(t) = x_{i,t}, X_j(t) = x_{i,t}) \approx x_{i,t} \lambda_{i,j} \Delta t.$$

In a CDP involving N lineages, the complexity of the system depends on the number of positive differentiation rates. This could vary from a minimum of 0 up to a maximum of $N(N - 1)$ in a fully interconnected system.

Readers familiar with literature concerning the modelling of systems of coupled biochemical reactions, could recognize similarities between those models and the presentation made so far for the CDP. Pursuing this parallelism, the set of cell events can be interpreted as first-order mass-action kinetics (*reactions*), whilst the transition probabilities per time unit, $X_i(t)\alpha_i$, $X_i(t)\delta_i$ and $X_i(t)\lambda_{i,j}$ correspond to the *propensity functions* (Wilkinson 2006; Purutcuoglu and Wit 2008). Defining $\theta = (\alpha, \delta, \lambda)$ as the r -dimensional column vector of parameters, the set of the propensity functions can be expressed in a compact matrix notation as a r -dimensional column vector, obtained from the product $D(\mathbf{X}(t))\theta$, where $D(\mathbf{X}(t))$ is as a $r \times r$ diagonal matrix with elements of vector $\mathbf{X}(t)$ replicated. Finally, the state changes associated to cell events can be recast into a *net effect* or *stoichiometric matrix*, V , defined as the $r \times N$ integer matrix.

In order to clarify the elements just introduced, we present a derivation for a fully connected CDP of size $N = 3$, that is graphically represented in Fig. 1b. In total, $r = 3 + 3 + (3 \times 2) = 12$ distinct cell events can be defined and accordingly, the parameter vector is given by: $\theta = (\alpha_1, \alpha_2, \alpha_3, \delta_1, \delta_2, \delta_3, \lambda_{2,1}, \lambda_{3,1}, \lambda_{1,2}, \lambda_{3,2}, \lambda_{1,3}, \lambda_{2,3})^T$. The net effect

matrix equals to: $V = \begin{bmatrix} V_{\text{dupl}} \\ V_{\text{death}} \\ V_{\text{diff}} \end{bmatrix}$ where

$$V_{\text{dupl}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; V_{\text{death}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}; V_{\text{diff}} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$$

and $D(\mathbf{X}) = \text{diag}(X_1, X_2, X_3, X_1, X_2, X_3, X_2, X_3, X_1, X_3, X_1, X_2)$. Because $\mathbf{X}(t)$ has been defined as a Markov process, given an initial condition vector, Kolmogorov forward equations enables to determine the time evolution of the process probability distribution $P(\mathbf{X}(t), t)$. An alternative and equivalent formulation of the Kolmogorov equation, widely used for modelling physical and chemical dynamic systems, is known as *master equation* (Kampen 1981; Risken 1984; Gardiner 1985).

For the CDP described in this paper, the master equation is given by:

$$\frac{dP(\mathbf{X}_t, t)}{dt} = \sum_{k=1}^r \{ [D(\mathbf{X}_t - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{X}_t - \mathbf{V}_{k,\cdot}, t) - [D(\mathbf{X}_t)\boldsymbol{\theta}]_k P(\mathbf{X}_t, t) \} \tag{1}$$

The solution of (1) involves the evaluation of the evolution of $P(\mathbf{X}(t), t)$ over the whole set of admissible configurations for process $\mathbf{X}(t)$. Clearly, for systems of realistic size and complexity this do not represent a feasible option. However, starting from (1), important information about the dynamics of characteristic statistical features of the system can be obtained. In particular, as shown in Appendix 1, two coupled sets of ordinary differential equations (ODEs) are derived, describing the time evolution of lineage population size averages, $E[X_i(t)]$, $i = 1, \dots, N$, and variances-covariances $\Sigma_{X_i, X_j}(t)$, $i, j = 1, \dots, N$:

$$\frac{dE[X_i(t)]}{dt} = \sum_{k=1}^r v_{k,i} [D(E[\mathbf{X}(t)])\boldsymbol{\theta}]_k$$

with initial condition:

$$E[X_i(t_0)] = x_{i,0} \tag{2}$$

and

$$\begin{aligned} \frac{d\Sigma_{X_i, X_j}(t)}{dt} &= \sum_{k=1}^r v_{k,j} [D(E[X_i(t)\mathbf{X}(t)])\boldsymbol{\theta}]_k + \sum_{k=1}^r v_{k,i} [D(E[X_j(t)\mathbf{X}(t)])\boldsymbol{\theta}]_k \\ &+ \sum_{k=1}^r v_{k,i} v_{k,j} [D(E[\mathbf{X}(t)])\boldsymbol{\theta}]_k - E[X_i(t)] \sum_{k=1}^r v_{k,j} [D(E[\mathbf{X}(t)])\boldsymbol{\theta}]_k \\ &- E[X_j(t)] \sum_{k=1}^r v_{k,i} [D(E[\mathbf{X}(t)])\boldsymbol{\theta}]_k; \end{aligned}$$

with initial conditions:

$$E[X_i(t_0)] = x_{i,0}; \quad E[X_i(t_0)X_j(t_0)] = x_{i,0}x_{j,0}. \tag{3}$$

In the next session, starting from (2) and (3), an inference procedure is presented.

Inference

In this section, a two step inference procedure for parameters estimation and model selection is described. The majority of studies aimed at answering biological questions through clonal tracking experiments provide information about the simultaneous evolution of several clones, observed at a limited set of timepoints. Assuming in a single experiment in total L clones have been tracked, with S total observations at times $t_1 < \dots < t_S$, each clone trajectory $\mathbf{x}^l = (\mathbf{x}_{t_1}^l, \dots, \mathbf{x}_{t_S}^l)$, with $l = 1, \dots, L$ corresponds to an independent realization of a unique, common CDP. Conditioning on the generic $\mathbf{x}_{t_S}^l$, the mean and variance-covariance values for lineage counts at time t_{s+1} , can be estimated by solving (2) and (3) with the following initial conditions:

$$E[X_i^l(t_s)] = x_{i,t_s}^l; \quad E[X_i^l(t_s)X_j^l(t_s)] = x_{i,t_s}^l x_{j,t_s}^l \tag{4}$$

A computationally efficient, albeit approximated, solution for $E[X_i^l(t_{s+1})]$ and $\Sigma_{X_i^l, X_j^l}(t_{s+1})$ can be calculated by Euler’s method. Accordingly, for the lineage specific mean population count, the following expression is derived:

$$E \left[X_i^l(t_{s+1}) \right] \simeq E \left[X_i^l(t_s) \right] + \frac{dE \left[X_i^l(t_s) \right]}{dt} [t_{s+1} - t_s]. \tag{5}$$

By substituting the second term on the right-hand-side (RHS) of (5) with (2) and including initial condition defined in (4), equation (5) becomes:

$$E \left[X_i^l(t_{s+1}) | X^l(t_s) = \mathbf{x}_{t_s}^l \right] \simeq x_{t_s}^{l,i} + \sum_{k=1}^r v_{k,i} \cdot \left[D \left(\mathbf{x}_{t_s}^l \right) \cdot \boldsymbol{\theta} \right]_k \cdot [t_{s+1} - t_s]$$

or, in an equivalent matrix notation:

$$E \left[X^l(t_{s+1}) \right] \simeq \mathbf{x}_{t_s}^l + V^T \cdot D \left(\mathbf{x}_{t_s}^l \right) \cdot \boldsymbol{\theta} \cdot [t_{s+1} - t_s]. \tag{6}$$

A similar reasoning can be applied to approximate the solutions for variance-covariance indexes, $\Sigma_{X_i^l, X_j^l}(t_{s+1})$, introduced in (3):

$$\Sigma_{X_i^l, X_j^l}(t_{s+1}) \simeq \Sigma_{X_i^l, X_j^l}(t_s) + \frac{d\Sigma_{X_i^l, X_j^l}(t_s)}{dt} [t_{s+1} - t_s]. \tag{7}$$

The first term in on the RHS of (7), by definition of variances-covariances as second central moments and initial conditions in (4), equals 0, while the second term simplifies to:

$$\frac{d\Sigma_{X_i^l, X_j^l}(t_s)}{dt} = \sum_{k=1}^r v_{k,i} \cdot v_{k,j} \cdot \left[D \left(\mathbf{x}_{t_s}^l \right) \cdot \boldsymbol{\theta} \right]_k$$

or, in matrix notation:

$$\Sigma_{X^l}(t_{s+1}) \simeq V^T \cdot D \left(\mathbf{x}_{t_s}^l \right) \cdot \text{Diag} \left(\boldsymbol{\theta} \right) \cdot V \cdot [t_{s+1} - t_s]. \tag{8}$$

Let now define $\Delta X^l(t_s)$ as:

$$\Delta X_{t_s}^l = X^l(t_{s+1}) - X^l(t_s) \tag{9}$$

a collection of N -dimensional r.v. modelling the state increment occurring in a time interval $[t_{s+1} - t_s]$. It is straightforward to show by linearity of expectation operator that:

$$E \left[\Delta X_{t_s}^l | X^l(t_s) = \mathbf{x}_{t_s}^l \right] \simeq V^T \cdot D \left(\mathbf{x}_{t_s}^l \right) \cdot \boldsymbol{\theta} \cdot [t_{s+1} - t_s] \tag{10}$$

and by invariance property with respect to shift in location parameters of variance-covariance index:

$$\Sigma_{\Delta X_{t_s}^l | X^l(t_s) = \mathbf{x}_{t_s}^l} \simeq V^T \cdot D \left(\mathbf{x}_{t_s}^l \right) \cdot \text{Diag} \left(\boldsymbol{\theta} \right) \cdot V \cdot [t_{s+1} - t_s]. \tag{11}$$

The piece-wise constant nature of the propensity functions over time, allows to make some considerations concerning the degree of the approximation provided by Euler’s method in (10) and (11). In the time elapsing between consecutive cellular events, the propensity functions are not subjected to variation, since they depend on the current process configuration and on parameters $\boldsymbol{\theta}$, assumed constant over time. Modification of their values can eventually occur only in coincidence with cellular events. It follows that if the set of time t_1, \dots, t_S corresponds to the sequence of events times, the probabilities $D \left(\mathbf{x}_{t_s}^l \right) \cdot \boldsymbol{\theta} \cdot [t_{s+1} - t_s]$ associated to the set of possible state variations are constant as well. It is thus possible to consider state increment as a regular discrete random variables and conclude that, under this specific sampling scheme, (10) and (11) are exact results rather than an approximation. When the time distance between consecutive observations increases, it becomes more likely that more than one event occur within them,

decreasing the quality of the approximation. The impact of sampling times intervals will be investigated by means of a simulation study in “[Simulation study](#)” section.

Approximate generalized method-of-moments estimation

Formulas given in (10) and (11) state that first order approximation for both increments conditional expectation and variance-covariances can be calculated as linear combination of the propensity functions at time t_s , in turn linear with respect to both process state and parameters vector. This result suggests an estimation of θ as a linear regression problem of type

$$dx \simeq M\theta + \varepsilon; \quad E[\varepsilon] = \mathbf{0}; \quad \Sigma_\varepsilon = W \tag{12}$$

where

$$dx = \begin{bmatrix} dx_{t_0}^1 \\ dx_{t_1}^1 \\ \vdots \\ dx_{t_{S-1}}^L \end{bmatrix} \quad M = \begin{bmatrix} M_{t_0}^1 \\ M_{t_1}^1 \\ \vdots \\ M_{t_{S-1}}^L \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} W_{t_0}^1 & 0 & \dots & 0 \\ 0 & W_{t_1}^1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_{t_{S-1}}^L \end{bmatrix}, \tag{13}$$

respectively, (i) dx is a $[L \cdot S \cdot N]$ -dimensional column vector in which the generic element $dx_{t_s}^l = \text{vec}(x_{t_{s+1}}^l - x_{t_s}^l)$ is a N -dimensional column vector corresponding to observed increments in cells counts, interpretable as realizations of the r.v. defined in (9); (ii) M is a $[L \cdot S \cdot N] \times r$ predictors matrix where the generic element $M_{t_s}^l = V^T \cdot D(x_{t_s}^l) \cdot [t_{s+1} - t_s]$ is a $N \times r$ matrix; (iii) The covariance matrix W is a blocks diagonal matrix describing the dependence between lineages counts increments belonging to the same time-point and independence among all the other. Each block $W_{t_s}^l$ is a $N \times N$ matrix and is approximated by (11) as $V^T \cdot D(x_{t_s}^l) \cdot [t_{s+1} - t_s] \cdot \text{diag} \cdot (\theta) \cdot V$.

Both duplication and death events involve only a single lineage and associated columns in M matrix result in pairs numerically equal but with opposite sign, causing rank deficiency. To overcome this issue, a new set of parameters, $\gamma = \alpha - \delta$ named *net duplication rates* is defined. Differently from others, elements of γ take values in \mathbb{R} . In particular, γ_i is positive if cells in lineage C_i undergo duplication with a higher rate than death ($\alpha_i \geq \delta_i$), and with a lower rate if the reverse is true ($\alpha_i \leq \delta_i$). Therefore, the following modifications have to be done. Parameters vector θ reduces to $\theta^* = (\gamma, \lambda)$, an r^* -dimensional vector, where r^* is equal to $r - N$, V^* is the *reduced* net effect matrix of dimension $r^* \times N$ and M^* is a $[L \cdot S \cdot N] \times r^*$ matrix equal to M but columns related to death events removed. Although it is not possible to infer simultaneously both duplication and death rates, the modifications introduced have limited impact on the precision of the estimators, since the product $M\theta$ is equal to $M^*\theta^*$.

It is now possible to define the constrained generalized least squares (CGLS) estimator as:

$$\hat{\theta}^* = \arg \min_{\theta^*} (dx - M^*\theta^*)^T W^{*-1} (dx - M^*\theta^*) \quad s.t. \quad \lambda \geq 0 \tag{14}$$

Due to the non-negativity constraint on differentiation rates λ and the dependence of W^* on the unknown parameters θ^* , a closed form solution for $\hat{\theta}^*$ is not available. A description of the iterative procedure for solving the CGLS minimization problem is presented in Appendix 2.

Model selection

Biologically meaningful and realistic cell differentiation structures are characterized by a limited number of connections between lineages. From a statistical modelling point of view, this type of result can be encouraged by means of sparsity promoting components in the differentiation parameters vector. Within the class of differentiation rates, it is usually possible to distinguish two types of relations between cell lineages: relations with differentiation rates known to be equal to zero due to the hierarchical cellular structure, and relations that do not violate hierarchical constraints and whose rates can be either zero or positive. The first category includes all rates related to the differentiation of mature lineages into stem/progenitors cells. The associated parameters are simply set to zero a priori. For the second category, a penalization procedure based on the smoothly clipped absolute deviation (SCAD) penalty function (Fan 1997) is applied. SCAD penalization has several advantages over another widely used technique, i.e., least absolute shrinkage and selection operator (LASSO). Although the LASSO has many excellent properties and very efficient implementations, it is a biased estimator. This bias affects in particular parameters that are truly non-zero, and does not disappear when the sample size increases. The SCAD penalty, instead, retains the penalization rate of the LASSO for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases, leading to asymptotically unbiased estimates (Fan and Li 2001). The SCAD penalty function is defined as:

$$p_\eta(\theta) = \begin{cases} \eta|\theta|, & \text{if } 0 \leq |\theta| \leq \eta \\ \frac{(\xi^2-1)\eta^2 - (|\theta|-\xi\eta)^2}{2(\xi-1)}, & \text{if } \eta \leq |\theta| \leq \xi\eta \\ \frac{(\xi+1)\eta^2}{2}, & \text{if } |\theta| \geq \xi\eta \end{cases} \quad (15)$$

where $\eta > 0$ and $\xi > 2$. As for many other penalization procedures, the role of the threshold parameters, η and ξ , is to tune the degree of sparseness in the final model and valid setting criteria for them are needed to ensure the accuracy of the estimator. Setting $\xi = 3.7$, the SCAD penalty has been demonstrated to give a satisfactory performance in a variety of variable selection problems (Fan and Li 2001) and it has therefore been adopted in this paper.

The optimal value of η has been chosen according to a generalized cross-validation (GCV) minimization criteria (Golub et al. 1979; Tibshirany 1996). The GCV is defined as:

$$GCV_\eta = \frac{1}{n} \frac{\|\mathbf{dx} - \widehat{\mathbf{dx}}\|^2}{(1 - e/n)^2} \quad (16)$$

where $\widehat{\mathbf{dx}} = \mathbf{M}^* \hat{\boldsymbol{\theta}}^*_{p,e}$, $e = \text{tr}[\mathbf{M}^*(\mathbf{M}^{*\top} \mathbf{W}^{*-1} \mathbf{M}^* + \mathbf{P}_\eta)^{-1} \mathbf{M}^{*\top} \mathbf{W}^{*-1}]$ corresponds to the number of effective parameters and \mathbf{P}_η is a $r^* \times r^*$ parameters penalization matrix described in detail in Appendix 3. Finally, given a particular value for η , the parameters estimates are calculated by minimizing the following objective function:

$$\hat{\boldsymbol{\theta}}^*_p = \arg \min_{\boldsymbol{\theta}^*} (\mathbf{dx} - \mathbf{M}^* \boldsymbol{\theta}^*)^\top \mathbf{W}^{*-1} (\mathbf{dx} - \mathbf{M}^* \boldsymbol{\theta}^*) + n \sum p_\eta(\lambda_{ij}) \text{ s. t. } \lambda \geq 0 \quad (17)$$

In the penalized CGLS (PCGLS) algorithm described in Appendix 3, the penalization function is included in an iterative procedure able to perform model selection and parameters estimation simultaneously.

Schematic overview of the inferential procedure

In this section, we outline whole inferential procedure in pseudo-code notation. The algorithm can be split into two major and consecutive parts that have been introduced in “[Approximate generalized method-of-moments estimation](#)” and “[Model selection](#)” sections: CGLS and PCGLS. Detailed description for each of them can be found in Appendices B and C, respectively. Algorithm 1 starts with the calculation of increments vector \mathbf{dx} and the regression matrix \mathbf{M}^* in (13). It receives as input a set η of candidate values for the SCAD tuning parameter. The initial values for the CGLS iterative procedure are calculated by solving a constrained ordinary least square problem (COLS), in which errors are assumed independent and homoscedastic. The COLS estimates $\hat{\theta}^{*0}$ are then used to calculate a first estimate for the covariance matrix $\hat{\mathbf{W}}^*$. By means of an iterative procedure, estimates for $\hat{\theta}^*$ and $\hat{\mathbf{W}}^*$ are then sequentially refined, until the convergence criteria are satisfied. Final $\hat{\theta}^*$ returned by the CGLS part is then used as parameter starting values in the PCGLS procedure, aimed at reconstructing the true, sparse model configuration by shrinking small coefficients to zero. In the PCGLS algorithm, parameter estimates and model structure identification are simultaneously updated. Once the convergence criterion is met, general cross-validation (GCV) statistics are calculated as shown in (16) to select the network configuration corresponding to minimum cross-validation error.

Simulation study

In this section we present a simulation study to evaluate the performance of the inference procedure. The settings used in the simulation study closely correspond to the gene therapy dataset analyzed in “[Investigating human hematopoiesis in vivo](#)” section, using the most recent model of hematopoiesis that has been suggested for non-human primates (Goyal et al. 2015). A simulated hierarchical differentiation process (SHDP) of size $N = 15$ has been designed with 3 hierarchical layers where differentiation paths are only allowed between adjacent levels and in a unidirectional way. The top layer, constituted by a single “stem cell” lineage (node: 1), is characterized by a positive net duplication rate. The middle layer is composed of 7 partially interconnected lineages (nodes: 2-8), all derived from differentiation events occurred in top lineage cells and able to generate bottom level cell types. The net duplication rates in this layer are heterogeneous and can be both positive or negative. Finally, the 7 lineages in the bottom layer (nodes: 9-15) have no differentiation potential and they all die faster than duplicate. A biological interpretation is that the top layer corresponds to stem cells, responsible for the generation of a set of myeloid and lymphoid branches specific progenitors in the BM (second layer). Each progenitor is then able to give rise to a small subset of committed cell, circulating in the PB (third layer) and characterized by limited lifespan. A graphical representation of the SHDP model is given in Fig. 2 along with a matrix representation of rates intensities, hierarchical constraints and the differentiation rates.

The aims of the simulation study are (i) to verify the performance of the proposed inferential procedure in case of a hierarchically structured systems and (ii) to measure the impact of different sampling time interval lengths on both estimation precision and model selection, i.e., the reconstruction of the true underlying differentiation process. Each experiment is composed of $N = 1000$ simulated clone evolutions, all generated starting from the same initial state vector consisting of a single hematopoietic stem cell. Continuous-time clones dynamics

Algorithm 1: Structure of the proposed inferential procedure composed by CGLS and PCGLS

Data: dx, M^*, η
Result: Parameters estimates $\hat{\theta}^*$ and model structure with minimum GCV
begin

CGLS (Appendix B): begin

Initialization: $tol = \epsilon$, $iter = 0$;
 $\hat{\theta}^{*0} = \arg \min_{\theta^*} (dx - M^* \theta^*)^\top (dx - M^* \theta^*)$ s.t. $\lambda_{i,j} \geq 0$;
while $\sum_{i=1}^{r^*} (|\hat{\theta}^{*iter} - \hat{\theta}^{*iter-1}|) \geq tol$ **do**
 Update \hat{W}^{*iter} with current $\hat{\theta}^{*iter}$;
 Update $\hat{\theta}^{*iter}$ with current \hat{W}^{*iter} ;
end while
Return $\hat{\theta}^*$

end

PCGLS (Appendix C): begin

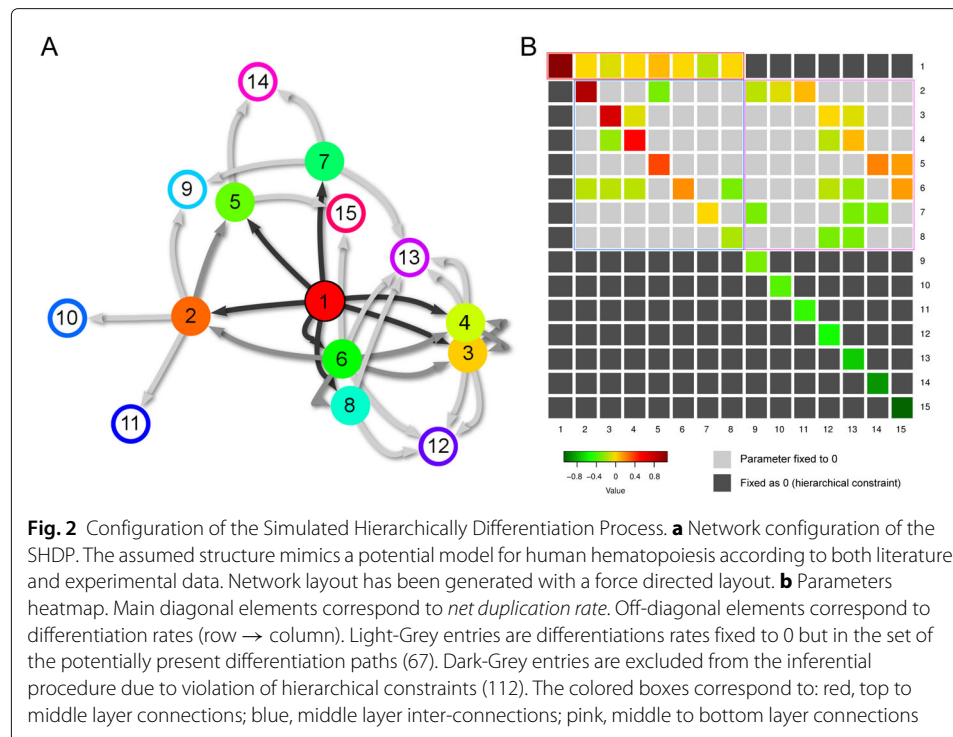
foreach η *in* η **do**
 $iter = 0$;
 $\hat{\theta}^*_p = \hat{\theta}^*$;
 while $\sum_{i=1}^{r^*} (|\hat{\theta}^{*iter}_p - \hat{\theta}^{*iter-1}_p|) \geq tol$ **do**
 Update \hat{W}^{*iter}_p with current $\hat{\theta}^{*iter}_p$;
 Update $P^{iter}_{\eta, \lambda}$ with current η , $\hat{\theta}^{*iter}_p$;
 Update $\hat{\theta}^{*iter}_p$ with current \hat{W}^{*iter}_p , $P^{iter}_{\eta, \lambda}$;
 end while
 Calculate $GCV(\eta)$;
end foreach

end
Return model associated to minimum $GCV(\eta)$;

end

are simulated by means of Gillespie algorithm (Gillespie 1977; Wilkinson 2006) according to the SHDP configuration in Fig. 2. Process states are then recorded at equispaced time intervals of lengths, $dt = (0.1, 0.2, 0.5, 0.7, 1)$, up to the end time-point fixed at $t_{end} = 4$. Based on these fixed time observations, state increments vector are calculated and given as input to the algorithm described in “Schematic overview of the inferential procedure” section. In the GCV procedure we consider a sequence of candidate values η from 0.001 to 0.1 with 0.001 step sizes.

In total, 100 independent experiments have been analyzed, each composed by 1000 simulated clones. Results are graphically represented in two figures. Figure 3 shows the impact of sampling interval length on the ability to correctly estimate process parameters. The performance regarding four specific rates are reported: a positive and a negative net duplication rate, one positive differentiation rate and one absent differentiation coefficient. The distribution of the estimates obtained from the 100 replicates are represented as a boxplot. In Fig. 4 the accuracy in terms of network reconstruction for increasing dt

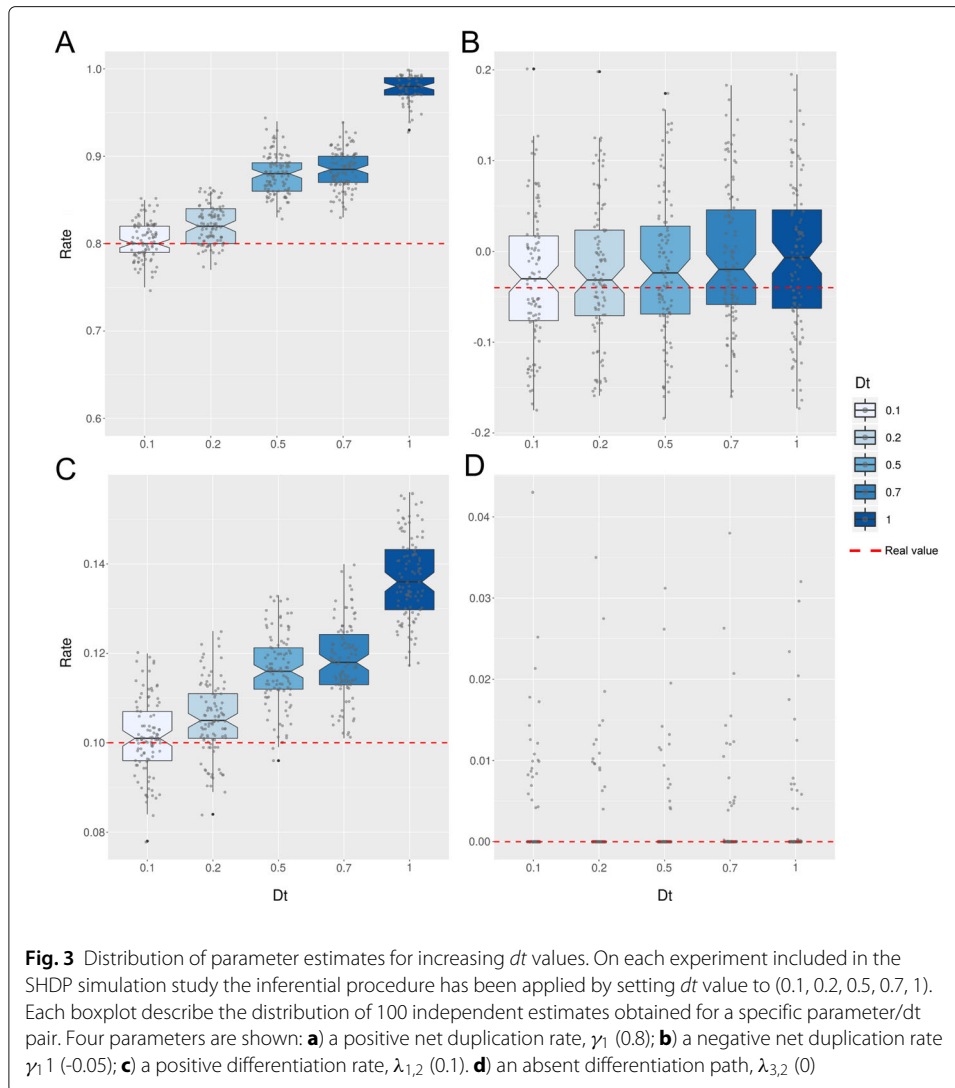


values is summarized. The distribution of two indices, recall and precision, are plotted by means of boxplots. In particular, we focus on verifying how reliable our proposal is in estimating absent connections among lineages. Precision measures the proportion of true differentiations among the identified differentiations, while recall (also known as sensitivity) is the proportion of identified differentiations among the true differentiations.

From a computational point of view, the Gillespie algorithm has been implemented in C++ (Stroustrup 1997) with the support of the **Eigen** library (Guennebaud et al. 2010). The inferential and penalization procedures are implemented in R (R Core Team 2015) by means of custom scripts requiring sparse Matrix packages for efficient dense and sparse matrix manipulations (Bates and Maechler 2015). QP problems (21) and (22) are solved by means of **IBM ILOG CPLEX Optimizer**, freely available under the IBM Academic Initiative program (IBM 2010). The simulated clone trajectories included in a single experiment (1000 clones) are generated in approximately 4 minutes. The inferential procedure takes from 5 to 12 minutes to complete on a single dataset, using candidate values for SCAD parameter (η) as mentioned above. In general, with a small dt value (0.1), the amount of data to be processed is about 10 times higher than with $dt = 1$, increasing the computational burden and time. This aspect is slightly counter-balanced by the fact that with higher dt values, the number of iterations required for convergence is higher (3.8 vs. 7.1 with dt equal to 0.1 and 1 respectively). For the setting tested and reported in this manuscript, no convergence issues have been observed.

Investigating human hematopoiesis in vivo

In this section, we return to the motivating Wiskott-Aldrich syndrome (WAS) gene therapy (GT) clinical study. The aim is to infer the network structure of the hematopoietic process in humans, along with lineage-specific duplication, death and differentiation



rates. Technical and experimental protocols used to collect the data have been described in Aiuti et al. (2013); Biasco et al. (2016); Scala et al. (2018) and are briefly summarized below.

At time 0, corrected HSCs harvested from BM are re-infused in 3 patients previously treated with bone marrow suppressive drugs enhancing immunosuppression in order to ensure a higher level of engraftment for corrected HSCs. In the patient’s body, marked HSCs start to duplicate, die and possibly differentiate into functionally more specialized cells, passing on the copy of the WASP gene to all the offspring generated, reconstituting a functional hematopoietic heritage. Viral IS selection is itself a quasi-random process and the probability that two integration events occur in the same genomic position in two distinct cells is negligible (Ambrosi et al. 2008; Biasco et al. 2011; Pellin and Di Serio 2016). Therefore, IS coordinates can be used as a molecular marker to monitor the *in-vivo* evolution of a single HSC and of its progeny. BM and PB samples have been taken for the 3 patients at 1, 2 and 3 years after treatment, enriched by means of magnetic cell sorting (MACS) technology according to a set of antibodies known to be lineage-specific.

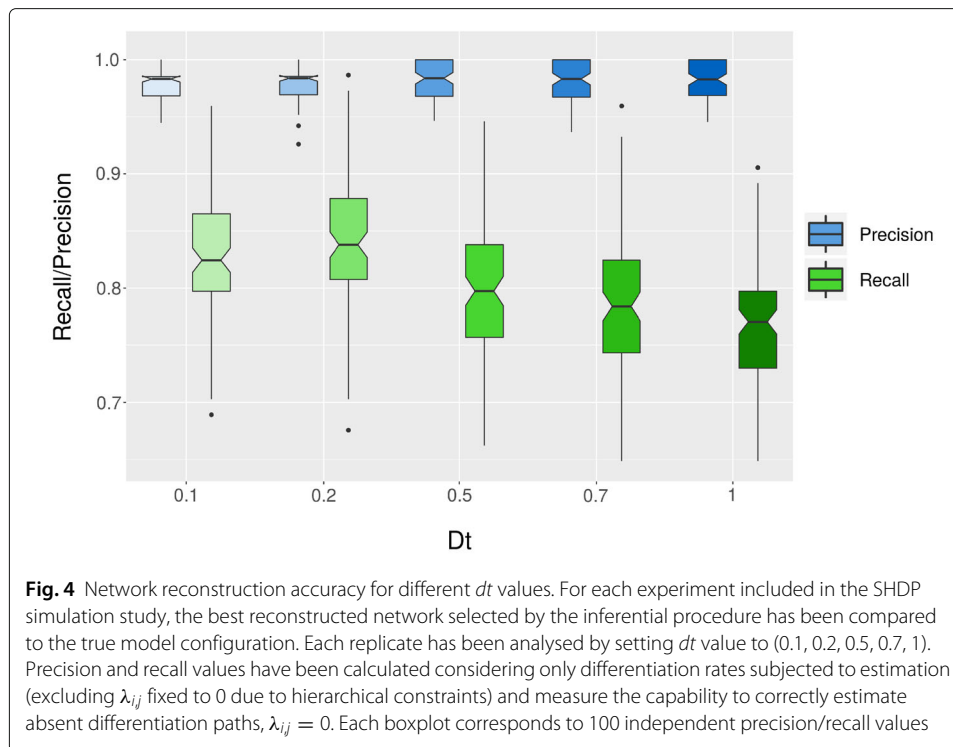


Fig. 4 Network reconstruction accuracy for different dt values. For each experiment included in the SHDP simulation study, the best reconstructed network selected by the inferential procedure has been compared to the true model configuration. Each replicate has been analysed by setting dt value to (0.1, 0.2, 0.5, 0.7, 1). Precision and recall values have been calculated considering only differentiation rates subjected to estimation (excluding λ_{ij} fixed to 0 due to hierarchical constraints) and measure the capability to correctly estimate absent differentiation paths, $\lambda_{ij} = 0$. Each boxplot corresponds to 100 independent precision/recall values

Finally, these samples were sequenced by means of the Illumina Miseq platform (Biasco et al. 2011). A bioinformatic pipeline starting from the sequencing output detects the IS coordinate (labels) and quantifies by means of reads count values the label distributions over lineages and time.

In total 37,637 distinct clones have been tracked covering 15 cell types divided in a three hierarchical levels: (i) the HSC level: CD34; (ii) the BM level: CD3, CD14, CD15, CD19, CD56, CD61, GLYCO and (iii) the PB level: CD3, CD4, CD8, CD14, CD15, CD19, CD56. In order to limit potential bias introduced by the low recapture probability of clones known to affect clonal tracking data, we kept only the 1083 clones with more than 15 observations across all lineages and time-points.

In accordance with the current state of the biological literature, the following assumptions have been made: (i) lineages in the HSC level can differentiate in any other cell type in BM level; (ii) lineages in the BM level can be connected to any cell type in BM and PB level; and (iii) lineages in the PB level cannot differentiate.

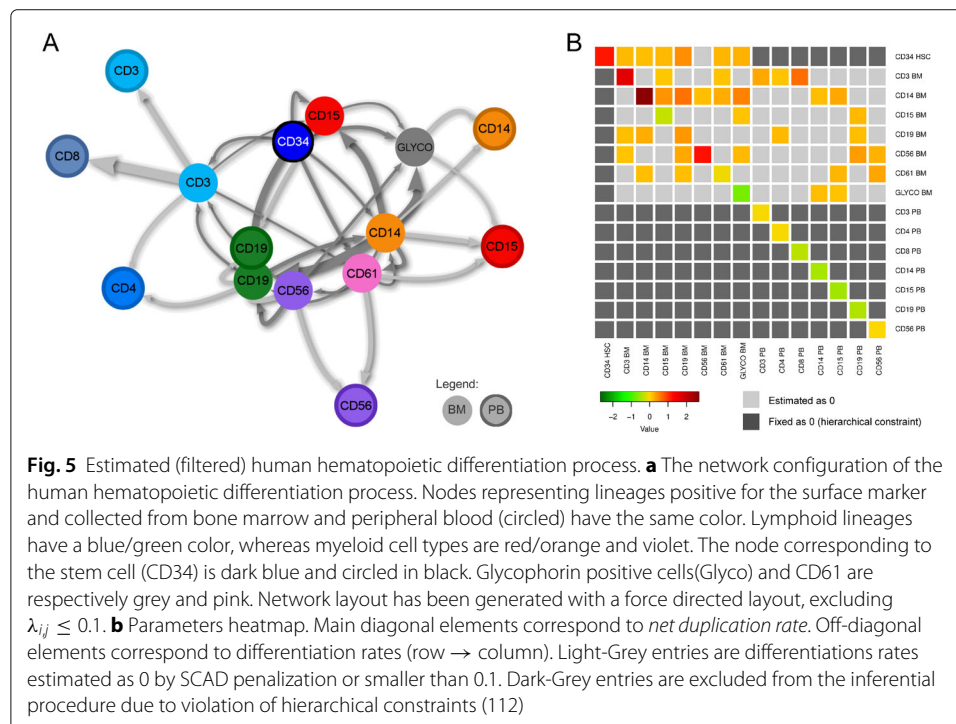
DNA library preparation for NGS based sequencing requires a linear amplification step known to be a source of noise potentially affecting cells counts measurement. To investigate the reliability of available information, part of the HSCs sample has been sequenced after a time interval in which is reasonable to assume that no or few cell events occurred, therefore are clones of size 1. Based on 3104 ISs, a *median absolute deviation* (MAD) statistics equal to 6.1 has been calculated, leading to a process noise estimate $\hat{\sigma}^2 = (1.48 \times MAD)^2 = 81.5$ (Rousseeuw and Croux 1993). This value has been incorporated in the inference and model selection procedures by modifying variance-covariance matrix as $\hat{W}^* = W^* + \hat{\sigma}^2 I$. In order to facilitate the biological interpretation of the results

obtained, a filtered version (only $\lambda_{i,j} \geq 0.1$) of estimated human hematopoiesis structure is given in Fig. 5. The full model can be found in Appendix 4.

Discussion

In the simulation study presented in “Simulation study” section, a candidate model of hematopoiesis of realistic complexity has been considered. The inferential procedures presented in “Approximate generalized method-of-moments estimation” section relies on iteratively updated approximate solutions for 2 coupled systems of ODEs with initial conditions. The accuracy of such approximations is known to be inversely proportional to the time distance between consecutive observations. This is particularly relevant given the sampling schema adopted in the experimental study. For these reasons, different values for dt have been considered and the clones’ evolution has been observed until $t_{end} = 4$.

As shown in Fig. 3, dt affects the estimation precision. This result was expected and can be attributed to the loss of approximation quality provided by Euler’s method for moments evolution with higher dt values. Bias consistently increases with interval length and for the settings considered in this paper only for $dt = 0.1$ parameter estimates are centered around the true values. The only exception is $\lambda_{3,2}$, shown in panel Fig. 3d, for which the performance is of good quality and similar across all dt settings, despite the fact that the overall amount of information available on clone dynamics decreases 10 times. This remarkable feature is the result of the penalization procedure that attenuate the dt effect by shrinking small coefficients to zero. In terms of model reconstruction accuracy, data in Fig. 4 show that precision values are close to 1 for all settings evaluated, meaning that we are very confident that $\hat{\lambda}_{i,j} = 0$ correspond to truly absent differentiation paths. This is consistent across dt . On the other hand, recall behaviour suggests that our proposal



is not able to identify all absent differentiations (i.e. $\lambda_{i,j} = 0$), but on average only 4% more are missed with $dt = 1$ (78%) compared to $dt = 0.1$ (82%).

In view of the above encouraging results, in particular regarding the network structure, it is possible to give the following interpretation of the final model for human hematopoiesis shown in Fig. 5. According to the hematological classification, the following branches can be defined: (i) the *lymphoid branch*, including CD3 and CD19 in BM and CD3, CD4, CD8 and CD19 in PB; (ii) the *myeloid branch* composed of CD14, CD15 and CD56 in both BM and PM; (iii) *Glycophorin positive cells*, corresponding to Glyco BM and (iv) the *CD61 positive lineage*. The flexibility of such a classification is currently debated and evidence for the presence of progenitors in BM straddling multiple branches emerged in multiple independent studies (Kawamoto et al. 2010; Kawamoto et al. 2010; Aiuti et al. 2013). Our results support this hypothesis. The complexity of the relationship among BM lineages is high and characterized by relevant cross-branches differentiation paths, mainly in myeloid to lymphoid direction, such as $CD14 \rightarrow CD19$, $CD14 \rightarrow GLYCO$ and $CD56 \rightarrow CD19$. A possible explanation can be found in a recent investigation aimed at dissecting the heterogeneous CD34 HSC population, suggesting the presence of intermediate stages, named Myeloid PluriPotent then followed by a Multi Lymphoid Progenitor (Biasco et al. 2016; Scala et al. 2018). All lineages in BM, with exception of CD15, have connections with their committed homologous subpopulation in PB compartment, as biologically expected.

The full model represented in Appendix 4 displays a much higher level of complexity with respect to Fig. 5. As showed in Appendix 5, a considerable amount of low differentiation rates are present. We consider them mostly related to the intrinsic sampling issues associated with clonal tracking experiments. In fact, it is difficult to obtain a consistent detection of all clones contributing to a given lineage across all time-points. The missing observation of clones in an intermediate cell population, say C_k , connecting lineages C_i and C_j for example, leads the inference algorithm at estimating weak differentiation rates between $C_i \rightarrow C_j$ directly, in addition to the true $C_i \rightarrow C_k \rightarrow C_j$ path. This problem affects in particular BM data, where bone marrow aspiration location, not always maintained unchanged over the follow-up period, can strongly affects clone capture probabilities. We set a threshold at 0.1, that we consider offering a good balance between model complexity and interpretability of the results.

Conclusion

In this paper, we presented a statistical model for cell differentiation process along with an inferential procedure able to provide parameters estimation and cell differentiation network reconstruction. The model has been defined as a continuous-time Markov chain, with density-dependent transition probabilities and considers three categories of cellular events: duplications, deaths, and differentiations. Starting from a special formulation of the Kolmogorov forward equation, two coupled set of ODEs have been derived, describing the time evolution of process first and second central moments over time. ODEs solutions have been approximated by Euler's method allowing for parameter estimation via a general linear regression setup.

In order to take into account the dependence among process components due to differentiation events and non-negativity constraints on a subset of parameters, estimation

is performed by means of an iterative generalized least square procedure, solved using an efficient quadratic programming approach. However, a biologically meaningful differentiation network is expected to be characterized by a limited amount of connections between lineages. To encourage a data-driven parsimonious solution and to provide an estimation of the *best candidate* differentiation pathway, a penalization step based on SCAD penalty function and a GCV criterion have been introduced.

Inferential procedures have been tested in a simulation study, mimicking a realistic candidate structure for human hematopoiesis. As expected, in case of frequently repeated observations for process state over time, Euler’s method provides a good approximation for moments evolution and, consequently, both parameters and model structure estimations are more accurate. When the time elapsing between consecutive observations increases, the quality of the estimations decreases, in particular for a subset of parameters. Despite this, model structure reconstruction is still reliable. The main limitation of the proposed model for cell differentiation, from a biological point of view, is represented by the linearity of the propensity functions, potentially allowing for unlimited clone expansion. However, it is worth noting that patients are immunodepressed at the time of treatment and that time necessary to reach a stable steady state equilibrium for lineages total populations is in the order of years.

Finally, human hematopoiesis structure and lineages specific parameters have been investigated by applying the developed method to a recent Wiskott-Aldrich syndrome gene therapy clinical study. The obtained result supports a recently proposed complex, interconnected myeloid/lymphoid branching model over previous simpler alternatives.

In future work we aim to extend the statistical framework presented in this paper to other, more flexible, dynamics formulation, such as Gompertz and logistic growth models. Additional attention will be paid to the inferential procedure, where the computationally efficient Euler’s method will be substituted with more complex alternatives, able to better approximate the solution of the ODE systems. From an application perspective, we consider of particular interests the potential comparison of the results obtained from the analysis of gene therapy data for WAS to those retrieved from different ongoing clinical trials, such as gene therapy for Adenosine Deaminase deficiency (ADA), Metachromatic Leukodystrophy (MLD) or sickle cell disease.

Appendix 1: Derivation of moment equations

By means of the summation operator, $\sum_{\mathbf{x} \in \tilde{\mathbf{x}}}$, spanning over the whole set of possible state for process $X(t)$, $\tilde{\mathbf{x}} = \mathbb{Z}^N$, it is possible to derived a functional connection between the evolution for the expected population size of each process component and the dynamics of the process probability distribution $P(X(t), t)$:

$$\begin{aligned} \frac{dE[X_i(t)]}{dt} &= \frac{d \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i P(X(t) = \mathbf{x}, t)}{dt} \\ &= \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i \frac{dP(X(t) = \mathbf{x}, t)}{dt} \end{aligned}$$

The evolution of $P(X(t), t)$ can be expressed by means of the master equation introduced in 1:

$$\frac{dE[X_i(t)]}{dt} = \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i \sum_{k=1}^r \{ [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{x} - \mathbf{V}_{k,\cdot}, t) - [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \}$$

Due to the fact that the summation operator $\sum_{\mathbf{x} \in \tilde{\mathbf{x}}}$ span over the all possible state configurations, the order of summation operators in the RHS can be inverted:

$$\begin{aligned} \frac{dE[X_i(t)]}{dt} &= \sum_{k=1}^r \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i \{ [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{x} - \mathbf{V}_{k,\cdot}, t) - [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \} \\ &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{x} - \mathbf{V}_{k,\cdot}, t) - \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \end{aligned}$$

Now, the summation variable in the first term of the RHS can be modified, without affecting the sum domain, since it cover all the possible state configuration:

$$\begin{aligned} \frac{dE[X_i(t)]}{dt} &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} (x_i + v_{k,i}) [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) - \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\ &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + v_{k,i} [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) - \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\ &= \sum_{k=1}^r \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,i} [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \end{aligned}$$

Given the known property for expected value of function $f(x)$ of a r.v. x with probability distribution $P(x)$, $E[f(x)] = \sum_x f(x)P(x)$:

$$\frac{dE[X_i(t)]}{dt} = \sum_{k=1}^r E[v_{k,i} [D(\mathbf{X}(t))\boldsymbol{\theta}]_k]$$

Finally, by linearity of expectation:

$$\frac{dE[X_i(t)]}{dt} = \sum_{k=1}^r v_{k,i} [D(E[\mathbf{X}(t)])\boldsymbol{\theta}]_k \tag{18}$$

A similar approach can be extended to define a system of ODEs for the time evolution for second order moments of $\mathbf{X}(t)$:

$$\begin{aligned}
 & \frac{dE[X_i(t)X_j(t)]}{dt} = \frac{d \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j P(\mathbf{X}(t) = \mathbf{x}, t)}{dt} \\
 &= \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j \frac{dP(\mathbf{X}(t) = \mathbf{x}, t)}{dt} \\
 &= \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j \sum_{k=1}^r \{ [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x} - \mathbf{V}_{k,\cdot}, t) - [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \} \\
 &= \sum_{k=1}^r \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x} - \mathbf{V}_{k,\cdot}, t) - [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \\
 &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x} - \mathbf{V}_{k,\cdot})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x} - \mathbf{V}_{k,\cdot}, t) - \right. \\
 & \quad \left. \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\
 &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} (x_i + v_{k,i})(x_j + v_{k,j}) [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) - \right. \\
 & \quad \left. \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\
 &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,j} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + \right. \\
 & \quad \left. \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,i} x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,i} v_{k,j} [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) - \right. \\
 & \quad \left. \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} x_i x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\
 &= \sum_{k=1}^r \left\{ \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,j} x_i [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,i} x_j [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) + \right. \\
 & \quad \left. \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} v_{k,i} v_{k,j} [D(\mathbf{x})\boldsymbol{\theta}]_k P(\mathbf{X}(t) = \mathbf{x}, t) \right\} \\
 &= \sum_{k=1}^r E[v_{k,j} X_i(t) [D(\mathbf{X}(t))\boldsymbol{\theta}]_k] + \sum_{k=1}^r E[v_{k,i} X_j(t) [D(\mathbf{X}(t))\boldsymbol{\theta}]_k] + \\
 & \quad \sum_{k=1}^r E[v_{k,i} v_{k,j} [D(\mathbf{X}(t))\boldsymbol{\theta}]_k] \\
 &= \sum_{k=1}^r v_{k,j} [D(E[X_i(t)X(t)])\boldsymbol{\theta}]_k + \sum_{k=1}^r v_{k,i} [D(E[X_j(t)X(t)])\boldsymbol{\theta}]_k + \\
 & \quad \sum_{k=1}^r v_{k,i} v_{k,j} [D(E[X(t)])\boldsymbol{\theta}]_k
 \end{aligned}$$

The generic element of the covariance matrix $\Sigma_{X_i, X_j}(t)$, describing the covariance between the $\mathbf{X}(t)$ components $X_i(t)$ and $X_j(t)$, can be expressed as a combination of first and second order moments:

$$\Sigma_{X_i, X_j}(t) = E[X_i(t)X_j(t)] - E[X_i(t)]E[X_j(t)] \tag{19}$$

Applying derivation rule to both sides, it is possible to derived a system of ODE for the evolution of covariance matrix elements as:

$$\frac{d\Sigma_{X_i, X_j}(t)}{dt} = \frac{dE[X_i(t)X_j(t)]}{dt} - \left(E[X_i(t)] \frac{dE[X_j(t)]}{dt} + E[X_j(t)] \frac{dE[X_i(t)]}{dt} \right)$$

Finally, substituting the RHS elements with the corresponding expression derived in (18), the following is obtained:

$$\begin{aligned} \frac{d\Sigma_{X_i, X_j}(t)}{dt} &= \sum_{k=1}^r v_{k,j} [D(E[X_i(t)X(t)])\theta]_k + \sum_{k=1}^r v_{k,i} [D(E[X_jX(t)])\theta]_k + \\ &\quad \sum_{k=1}^r v_{k,i}v_{k,j} [D(E[X(t)])\theta]_k - E[X_i(t)] \sum_{k=1}^r v_{k,j} [D(E[X(t)])\theta]_k - \\ &\quad E[X_j(t)] \sum_{k=1}^r v_{k,i} [D(E[X(t)])\theta]_k \end{aligned}$$

Appendix 2: Constrained generalized least square procedure

The algorithm described in pseudo-code notation in Algorithm 2 starts with the calculation of increments vector \mathbf{dx} and predictors matrix M^* according to (13).

Algorithm 2: Iterative procedure for CGLS based parameters estimation.

Data: \mathbf{dx}, M^*

Result: Get parameters estimates $\hat{\theta}^*$

begin

Initialization: $\text{tol} = \epsilon, \text{iter} = 0;$

$\hat{\theta}^{*0} = \arg \min_{\theta^*} (\mathbf{dx} - M^*\theta^*)^\top (\mathbf{dx} - M^*\theta^*) \text{ s.t. } \lambda_{i,j} \geq 0;$

while $\sum_{i=1}^r (|\hat{\theta}^{*iter} - \hat{\theta}^{*iter-1}|) \geq \text{tol}$ **do**

$$\hat{W}^*_{iter} = \begin{bmatrix} \hat{W}^*_{t_0} & 0 & \dots & 0 \\ 0 & \hat{W}^*_{t_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{W}^*_{t_{S-1}} \end{bmatrix} \text{ where}$$

$$\hat{W}^*_{t_s} = V^{*\top} D(\mathbf{x}_{t_s})(t_{s+1} - t_s) \text{Diag}(|\hat{\theta}^{*iter}|) V^*;$$

$\text{iter} = \text{iter} + 1;$

$$\hat{\theta}^{*iter} = \arg \min_{\theta^*} \left[\theta^{*\top} (M^{*\top} \hat{W}^*_{iter} M^*) \theta^* - 2(\mathbf{dx}^\top \hat{W}^*_{iter} M^*)^\top \theta^* \right]$$

$\text{s.t. } A\theta^* \leq \mathbf{b}$

end while

end

Parameters initial values for the proposed iterative procedure are calculated by solving a constrained ordinary least square problem (COLS), in which errors are assumed independent and homoscedastic. The COLS estimates $\hat{\theta}^{*0}$ are then used to calculate a first estimation for the covariance matrix \hat{W}^* . By means of an iterative procedure, estimates for $\hat{\theta}^*$ and \hat{W}^* are then sequentially refined, until convergence criteria on parameters vector is satisfied. From an optimization point of view, both COLS and CGLS estimations

can be interpreted as a quadratic programming (QP) problems, a special type of mathematical optimization problem in which a quadratic function has to be minimized (or maximized) taking into account for a set of linear constraints on variables. In general, a quadratic programming problem with n variables and m constraints can be formulated as follows.

Given a n -dimensional vector c , an $n \times n$ symmetric matrix Q , an $m \times n$ matrix A and an m -dimensional vector b , the goal is to find the n -dimensional vector x , such that:

$$\hat{x} = \arg \min_x \left(\frac{1}{2} x^T Q x + c^T x \right) \text{ s.t. } A x \leq b. \tag{20}$$

The CGLS problem defined in (14) can be converted in a QP problem of type (20) by setting $x = \theta^*$, $Q = 2(M^{*T} W^{*-1} M^*)$, $c = -2(dx^T W^{*-1} M^*)$, $b = 0_{r^*}$ and defining A as a $r^* \times r^*$ diagonal matrix with elements $A_{i,i} = 0$ if i_{th} element of θ^* refers to a net duplication rate (unconstrained) and $A_{i,j} = -1$ if it is a differentiation rate (non-negativity constrained). Finally, the QP problem becomes:

$$\hat{\theta}^* = \arg \min_{\theta^*} \left[\theta^{*T} (M^{*T} W^{*-1} M^*) \theta^* - 2(dx^T W^{*-1} M^*)^T \theta^* \right] \text{ s.t. } A \theta^* \leq 0_{r^*}. \tag{21}$$

For COLS is sufficient to remove W^{*-1} in Q and c formulas, other terms remain unchanged. It is worth noting that in case of large systems and/or when the amount of observations is high, it is possible to take advantage of the following property for block structured matrix, in order to calculate the inverse W^{*-1} :

$$\begin{bmatrix} \hat{W}^{*iter}_{t_0} & 0 & \dots & 0 \\ 0 & \hat{W}^{*iter}_{t_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{W}^{*iter}_{t_{S-1}} \end{bmatrix}^{-1} = \begin{bmatrix} \hat{W}^{*iter-1}_{t_0} & 0 & \dots & 0 \\ 0 & \hat{W}^{*iter-1}_{t_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{W}^{*iter-1}_{t_{S-1}} \end{bmatrix}$$

In real data analysis, this aspect allows to remarkably reduce both computational complexity and memory requirements of the estimation algorithm.

Appendix 3: Penalized constrained generalized least square procedure

The minimization problem in (17) can be formulated as a QP problem similarly to what presented in “Conclusion” section, by making the following modification to the definition of matrix Q :

$$Q_P = 2(M^{*T} W^{*-1} M^* + P_\eta)$$

where

$$P_\eta = \begin{bmatrix} P_\gamma & \mathbf{0} \\ \mathbf{0} & P_{\eta,\lambda} \end{bmatrix}$$

is $r^* \times r^*$ diagonal penalization matrix, with elements defined as:

$$P_\gamma = \mathbf{0}_{N,N}; \quad P_{\eta,\lambda} = n \begin{bmatrix} \frac{p'_\eta(\lambda_{1,2})}{\lambda_{1,2}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{p'_\eta(\lambda_{N-1,N})}{\lambda_{N-1,N}} \end{bmatrix}$$

and the derivative of SCAD penalty function (15) is given by:

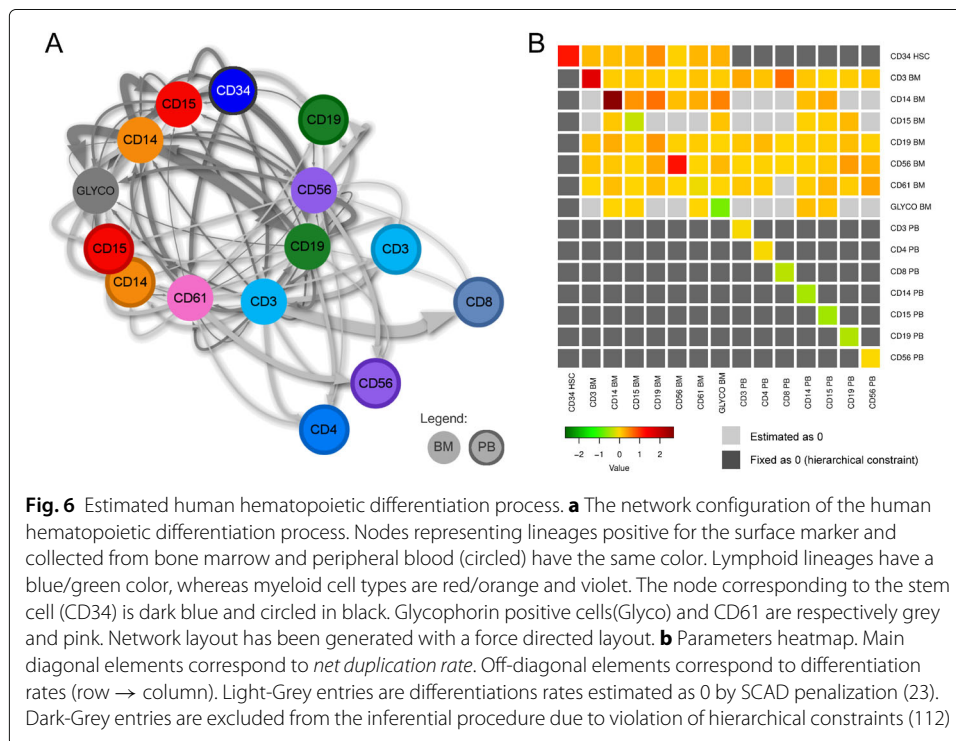
$$p'_\eta(\theta) = \eta I(\theta \leq \eta) + \frac{(\xi \eta - \theta)}{(\xi - 1)\eta} I(\theta > \eta).$$

The QP problem for PCGLS becomes:

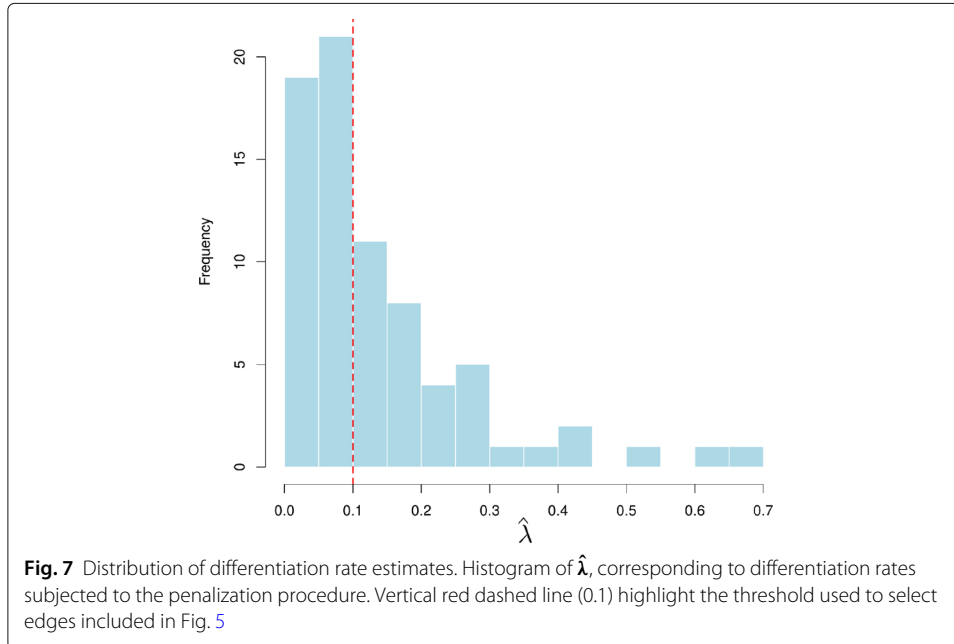
$$\hat{\theta}^*_p = \arg \min_{\theta^*} \left[\theta^{*\top} (M^{*\top} W^{*-1} M^* + P_\eta) \theta^* - 2(dx^\top W^{*-1} M^*)^\top \theta^* \right] \text{ s.t. } A\theta^* \leq \mathbf{0}_{r^*}. \tag{22}$$

The algorithm described in Algorithm 3 takes in input the parameters vector estimates $\hat{\theta}^*$ obtained from the CGLS procedure, state increments vector, dx , predictors matrix M^* and a vector of candidate values for tuning parameter η , named η . For each value in η , the initial values for parameters is set to $\hat{\theta}^*$ and then an iterative procedure composed by estimation of the covariance matrix; calculation of the penalty matrix, optimization of the objective function; is reiterated until convergence criteria on parameters vector estimates are met. For each value of η , based of the final estimates returned by the iterative procedure, the GCV statistics is calculated and stored. Finally, the rates estimates corresponding to the minimum GCV statistics are returned.

Appendix 4: Full network representation of the estimated human hematopoietic differentiation process



Appendix 5: Histogram of estimated differentiation rates



Algorithm 3: Iterative procedure for PCGLS based model selection and parameters estimation

Data: $\hat{\theta}^*, dx, M^*, \eta$

Result: Model selection based on minimum GCV criteria.

Initialization: $\text{tol} = \epsilon, \xi;$

begin

foreach η *in* η **do**

$\text{iter} = 0;$

$\hat{\theta}_{P}^{*0} = \hat{\theta}^*;$

while $\sum_{i=1}^{r^*} (|\hat{\theta}_{P}^{*iter} - \hat{\theta}_{P}^{*iter-1}|) \geq \text{tol}$ **do**

$$\hat{W}_{P}^{*iter} = \begin{bmatrix} \hat{W}_{P,t_0}^{*iter} & 0 & \dots & 0 \\ 0 & \hat{W}_{P,t_1}^{*iter} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{W}_{P,t_{S-1}}^{*iter} \end{bmatrix} \text{ where}$$

$$\hat{W}_{P,t_s}^{*iter} = V^T D(x_{t_s})(t_{s+1} - t_s) \text{Diag}(|\hat{\theta}_{P}^{*iter}|) V;$$

$\text{iter} = \text{iter} + 1;$

$$P_{\eta,\lambda}^{*iter}; P_{\eta}^{*iter} = \begin{bmatrix} P_{\gamma} & 0 \\ 0 & P_{\eta,\lambda}^{*iter} \end{bmatrix};$$

$$\hat{\theta}_{P}^{*iter} = \arg \min_{\theta^*} \{ \theta^{*\top} [M^{*\top} (\hat{W}_{P}^{*iter})^{-1} M^* + P_{\eta}^{*iter}] \theta^* -$$

$$2[dx^{\top} (\hat{W}_{P}^{*iter})^{-1} M^{*\top}] \theta^* \} \text{ s.t. } A\theta^* \leq b;$$

end while

$\text{GCV}(\eta);$

end foreach

 Return model associated to minimum $\text{GCV}(\eta);$

end

Appendix 6: Abbreviations

- BM: bone marrow
- CDP: cell differentiation process
- CGLS: constrained generalized least squares
- GCV: generalized cross-validation
- GT: gene therapy
- HSC: hematopoietic stem cells
- IS: integration site
- MACS: magnetic cell sorting
- MAD: mean absolute deviation
- NGS: next-generation sequencing
- ODE: ordinary differential equation
- PB: peripheral blood
- RHS: right-hand side
- SCAD: smoothly clipped absolute deviation
- SHDP: simulated hierarchical differentiation process
- WAS: Wiskott-Aldrich syndrome

Appendix 7: Mathematical symbols

- α, α_i : duplication rate
- δ, δ_j : death rate
- $\Delta X^l(t)$: random variable for clone l increments
- λ, λ_{ij} : differentiation rate
- γ, γ_i : net duplication rate
- $\eta > 0, \xi > 2$: SCAD tuning parameters
- θ, θ^* : parameters vector
- C_i : cell type
- $D(O)$: diagonal matrix with appropriate process component repetition
- dx, dx_i^l : increments vector
- $\frac{dP(X_t, t)}{dt}$: ODE describing evolution over time of process probability distribution
- $\frac{dE[X_i(t)]}{dt}$: ODE describing evolution over time of mean for cell type i counts
- $\frac{d\Sigma_{X_i, X_j}(t)}{dt}$: ODE describing variance-covariance evolution for cell type i constrained generalized least squares
- $E[X_i(t)], E[X_i^l(t)]$: expected value for process state at time t
- GCV: Generalized Cross Validation
- L, l : number of clones
- M, M_t^l, M^* : predictors matrix
- N : number of cell types
- P_η : parameters penalization matrix
- $P(X(t), t)$: process probability distribution
- r : total number of cell events
- S, s : total number of timepoints and related index
- t, t_s : time and timepoints
- V, V_k, v_{ij} : net effect matrix
- W, W_t^l, W^* : covariance matrix

- $X(t), X_i(t)$: stochastic process for cell differentiation process
- $x^l, x_t^l, x_{i,t}^l$: observed state for clone l
- $X^l(t), X_i^l(t)$: stochastic process for clone l dynamics
- $x_t, x_{i,t}$: observed state at time t
- $\Sigma_{X^l}(t), \Sigma_{X_i, X_j}(t)$: variance-covariance matrix

Acknowledgments

The authors acknowledge support from EU COST Action COSTNET on Statistical Network Science (CA15109).

Authors' contributions

DP and EW developed the modeling and inference procedures. CdS critically reviewed and provided useful suggestions. DP wrote the draft and EW redacted it. LB and AA provided the data and biomedical interpretation of the results. All authors read and approved the final manuscript.

Availability of data and materials

Additional information about the study, the experimental procedures and the data can be found at <http://dx.doi.org/10.1016/j.stem.2016.04.016>.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Dana-Farber, Harvard Medical School, 1 Jimmy Fund Way, 02115 Boston, MA, USA. ²San Raffaele Telethon Institute for Gene Therapy, IRCCS Ospedale San Raffaele, Milano, Italy. ³University Center for Statistics in the Biomedical Sciences, Vita-Salute San Raffaele University, Milan, Italy. ⁴Institute of Computational Science, Università della Svizzera italiana, Via G. Buffi 13, 6900 Lugano, Switzerland. ⁵Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands. ⁶Medicine, Università Vita Salute San Raffaele, Milano (Italy). ⁷Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Harvard Medical School, 02115 Boston, MA, USA. ⁸University College of London (UCL), Great Ormond Street Institute of Child Health, Faculty of Population Health Sciences, London, WC1N 1EH, United Kingdom.

Received: 19 March 2019 Accepted: 18 October 2019

Published online: 02 December 2019

References

- Abkowitz JL, Linenberger ML, Newton MA, Shelton GH, Ott RL, Guttorp P (1990) Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. *Proc Natl Acad Sci* 87(22):9062–9066
- Aiuti A, Biasco L, Scaramuzza S, Ferrua F, Cicalese MP, Baricordi C, Dionisio F, Calabria A, Giannelli S, Castiello MC, Bosticardo M, Evangelio C, Assanelli A, Casiraghi M, Di Nunzio S, Callegaro L, Benati C, Rizzardi P, Pellin D, Di Serio C, Schmidt M, Von Kalle C, Gardner J, Mehta N, Neduva V, Dow DJ, Galy A, Miniero R, Finocchi A, Metin A, Banerjee PP, Orange JS, Galimberti S, Valsecchi MG, Biffi A, Montini E, Villa A, Ciceri F, Roncarolo MG, Naldini L (2013) Lentiviral hematopoietic stem cell gene therapy in patients with wiskott-aldrich syndrome. *Science* 341(6148). <https://doi.org/10.1126/science.1233151>. <http://www.sciencemag.org/content/341/6148/1233151.full.pdf>
- Ambrosi A, Cattoglio C, Di Serio C (2008) Retroviral integration process in the human genome: is it really non-random? a new statistical approach. *PLoS Comput Biol* 4(8):1000144
- Bates D, Maechler M (2015) Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-2. <http://CRAN.R-project.org/package=Matrix>. Accessed 25 Oct 2019
- Biasco L, Ambrosi A, Pellin D, Bartholomae C, Brigida I, Roncarolo MG, Di Serio C, von Kalle C, Schmidt M, Aiuti A (2011) Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol Med* 2(5):1757–4684
- Becker A, McCulloch E, Till J (1963) Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* 197:452–454
- Biasco L, Pellin D, Scala S, Dionisio F, Basso-Ricci L, Leonardelli L, Scaramuzza S, Baricordi C, Ferrua F, Cicalese MP, et al. (2016) In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* 19:107–119
- Catlin SN, Abkowitz JL, Guttorp P (2001) Statistical inference in a two-compartment model for hematopoiesis. *Biometrics* 57(2):546–553
- Fan J (1997) Comments on wavelets in statistics: a reviews by a. antoniadis. *J Ital Stat Assoc* 6:131–138
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96:1348–1360
- Gardiner CW (1985) *Handbook of Stochastic Methods*. Springer, New York
- Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361
- Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–223
- Goyal S, Kim S, Chen IS, Chou T (2015) Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. *BMC Biol* 13(1):85
- Guennebaud G, Jacob B, et al. (2010) Eigen v3. <http://eigen.tuxfamily.org>. Accessed 25 Oct 2019
- IBM (2010) User's Manual for CPLEX. IBM ILOG CPLEX V12.1. https://public.dhe.ibm.com/software/websphere/ilog/docs/optimization/cplex/ps_usrmancomplex.pdf. Accessed 25 Oct 2019

- Kampen NGV (1981) *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam
- Kawamoto H, Ikawa T, Masuda K, Wada H, Katsura Y (2010) A map for lineage restriction of progenitors during hematopoiesis: the essence of the myeloid-based model. *Immunol Rev* 238(1):23–36
- Kawamoto H, Wada H, Katsura Y (2010) A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. *Int Immunol* 22(2):65–70
- Marciniak-Czochra A, Stiehl T (2013) Mathematical models of hematopoietic reconstitution after stem cell transplantation. In: *Model Based Parameter Estimation*. Springer, Berlin, pp 191–206
- Naldini L (2011) Ex vivo gene transfer and correction for cell-based therapies. *Nat Rev Genet* 12(5):301–15
- Pellin D, Di Serio C (2016) A novel scan statistics approach for clustering identification and comparison in binary genomic data. *BMC Bioinformatics* 17(11):320
- Purutcuoglu V, Wit E (2008) Bayesian inference for the mapk erk pathway by considering the dependency of the kinetic parameters. *Bayesian Anal* 3(4):851–886
- R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. R Foundation for Statistical Computing. <https://www.R-project.org/>. Accessed 25 Oct 2019
- Risken H (1984) *The Fokker-Planck Equation*. Springer, New York
- Romano O, Peano C, Tagliazucchi GM, Petiti L, Poletti V, Cocchiarella F, Rizzi E, Severgnini M, Cavazza A, Rossi C, et al (2016) Transcriptional, epigenetic and retroviral signatures identify regulatory regions involved in hematopoietic lineage commitment. *Sci Rep* 6:24724
- Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. *J Am Stat Assoc* 88(424):1273–1283
- Scala S, Basso-Ricci L, Dionisio F, Pellin D, Giannelli S, Salerio FA, Leonardelli L, Cicalese MP, Ferrua F, Aiuti A, et al (2018) Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat Med* 24(11):1683
- Stroustrup B (1997) *The C++ Programming Language*. 3rd edn. Addison-Wesley, Boston
- Tibshirany R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 58:267–288
- Wilkinson DJ (2006) *Stochastic Modelling for Systems Biology*. Chapman and Hall, London

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
