Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of *in vivo* hematopoiesis

Jason Xu[1], Samson Koelle[1], Peter Guttorp[1], Chuanfeng Wu[2], Cynthia E. Dunbar[2], Janis L. Abkowitz[3], Vladimir N. Minin[1,4]
[1]Department of Statistics, University of Washington
[2]National Heart, Lung, and Blood Institute, National Institutes of Health
[3]Department of Medicine, Division of Hematology, University of Washington
[4]Department of Biology, University of Washington

Abstract

Single-cell lineage tracking strategies enabled by recent experimental technologies have produced significant insights into cell fate decisions, but lack the quantitative framework necessary for rigorous statistical analysis of mechanistic models of cell division and differentiation. In this paper, we develop such a framework with corresponding moment-based parameter estimation techniques for continuous-time stochastic compartmental models that provide a probabilistic description of how cells divide and differentiate. We apply this method to *hematopoiesis*, the complex mechanism of blood cell production. Viewing compartmental models of cell division and differentiation as multi-type branching processes, we derive closed-form expressions for higher moments in a general class of such models. These analytical results allow us to efficiently estimate parameters of compartmental models of hematopoiesis that are much richer than the models used in previous statistical studies. To our knowledge, the method provides the first rate inference procedure for fitting such models to time series data generated from cellular barcoding experiments. After testing the methodology in simulation studies, we apply our estimator to hematopoietic lineage tracking data from rhesus macaques. Our analysis provides a more complete understanding of cell fate decisions during hematopoiesis in non-human primates, which may be more relevant to human biology and clinical strategies than previous findings in murine studies. The methodology is transferrable to a large class of compartmental models and multi-type branching models, commonly used in studies of cancer progression, epidemiology, and many other fields.

1 Introduction

This paper develops inferential tools for a class of hidden stochastic population processes. In particular, we present a correlation-based z-estimator for rate inference in multi-type branching process models of *hematopoiesis*, the process of blood cell production. During hematopoiesis, self-renewing hematopoietic stem cells (HSCs) specialize or *differentiate* via a series of intermediate progenitor cell stages to produce mature blood cells [Weissman, 2000]. Understanding the details of this system is a fundamental problem in biology, and progress in this area will also help shed light on other areas of basic biology. For example, further advances in hematopoiesis research will yield insights into mechanisms of cellular interactions, cell lineage programming, and characterization of cellular phenotypes during cell differentiation [Orkin and Zon, 2008]. Moreover, understanding hematopoiesis is clinically important: all blood cell diseases, including leukemias, myeloproliferative disorders and myelodysplasia are caused by malfunctions in some part of the hematopoiesis process, and hematopoietic stem cell transplantation has become a mainstay for gene therapy and cancer treatments [Whichard et al., 2010].

Hematopoiesis research was one of the earliest successes of mathematical modeling in cell biology [Becker et al., 1963, Siminovitch et al., 1963]. *Stochastic compartmental models* form one popular class of models used to study hematopoiesis, in which cells are assumed to self replicate and differentiate according to a *Markov branching process* [Kimmel and Axelrod, 2002]. While much is known about production of blood cells by progenitor cells, uncovering details of HSC and progenitor cell replication/differentiation dynamics has proven to be more difficult. Notably, experimental techniques developed to study feline hematopoiesis using X-chromosome inactivation markers have produced a series of statistical studies using a two-compartmental stochastic model of hematopoiesis [Abkowitz et al., 1990, Newton et al., 1995, Golinelli et al., 2006, Fong et al., 2009, Catlin et al., 2011]. However, this simple two-type representation cannot distinguish between stages of differentiation beyond the HSC, and results obtained from analyzing this model have not resolved long standing questions about patterns and sizes of the clones descended from individual HSC cells. It should be noted that even these simplified models capturing the clonal dynamics descended from an HSC have posed significant statistical and computational challenges.

More complex multi-compartmental models have been studied mathematically under additional assumptions—for instance, the regulatory behavior of several multi-stage models have been studied by Colijn and Mackey [2005] and Marciniak-Czochra et al. [2009]. Efforts in analyzing these more precisely specified structures have relied on deterministic modeling of the overall population with continuous-valued state variables. To study dynamics in detail at the single-cell level, continuous approximations are not suitable as HSC counts descended from a single clone are often very low and near zero, and while deterministic models may be appropriate for steady-state population level behavior, they are unsuitable to model fate decisions at the cell level which are much more sensitive to stochastic events. Indeed, studies suggest that hematopoietic dynamics are stochastic in nature [Ogawa, 1993, Kimmel, 2014]. Additionally, as one cannot completely specify all details of such a complex system in a mathematical model, a stochastic modeling approach that quantifies uncertainty provides a natural safeguard against model misspecification to some extent.

Recently emergent experimental techniques now allow researchers to track the dynamics of cell lineages descended from distinct ancestral progenitor or HSC cells. Collecting such high resolution data is made possible by lentiviral genetic barcoding coupled with modern high-throughput sequencing technologies [Gerrits et al., 2010, Lu et al., 2011, Wu et al., 2014]. Data collected from individual cell barcodes, rather than from a population descended from a mixture of indistinguishable clones, comprise independent and identically distributed time series, potentially allowing for investigation of much more realistic models of hematopoiesis. More importantly, the ability to analyze individual lineage trajectories can be very useful in characterizing patterns of cell differentiation, shedding light on the larger tree structure of the differentiation process. While these data are certainly more informative than those from previous experiments, statistical methods capable of analyzing such data are only beginning to emerge. Perié et al. [2014] model genetic barcoding data in a murine study collected at the end of the mice's lifespans, but do not account for the longitudinal aspect of the data nor read count information, instead working with a binarized simplification of the data. Goyal et al. [2015] present a neutral steady-state model of long term hematopoiesis applied to vector site integration data, but cannot infer crucial process parameters such as the rate of stem cell self-renewal. Biasco et al. [2016] manage to estimate cell differentiation rates from blood lineage tracking data, but resort to diffusion approximation and ignore experimental noise during their statistical analysis.

Wu et al. [2014] provide a preliminary analysis of their cellular barcoding data revealing important scientific insights, but lacking the ability to perform statistical tasks such as parameter estimation and model fitting/selection. This paper attempts to fill this methodological gap, developing new statistical techniques for studying the barcoded hematopoietic cell lineages from the rhesus macaque data. The difficulty lies in the partially observed nature of a complex process with a massive hidden state space. Statistical challenges arise from several facets of the experimental design so that standard techniques for hidden Markov models and continuous-time March chains

(CTMCs) cannot be readily applied, instead requiring careful modeling that at once captures the complexity of the data yet allows feasible algorithms for inference. We propose a fully generative stochastic modeling framework and an efficient method of parameter estimation that allows much richer hematopoietic structures to be statistically analyzed than previously possible, allowing for many-compartmental models that consider HSC, progenitor, and mature cell stages. The following section details the experimental design and dataset we consider, and provides an overview of the stochastic model. Next, we motivate the approach by statistically formulating our inferential goal, provide a rigorous characterization of each component of our model, and derive the necessary mathematical expressions in Section 3. We then thoroughly validate these methods via several simulation studies, and fit the models to the rhesus macaque barcoding data. Finally, we close with a discussion of these results, their implications, and avenues for future work.

2 Data and Model

We analyze single cell lineage tracking data generated by the cellular barcoding experiments in [Wu et al., 2014]. Briefly, Wu et al. [2014] start by mobilizing marrow cells from rhesus macaques into blood, selecting the subset of cells that contains HSCs and progenitors, and labelling these cells. Specifically, lentiviral vectors are created using high diversity oligonucleotides with known DNA sequences that can later be retrieved — these vector sequences each correspond to a unique ID, collectively forming a genetic *barcode library*. Next, cells are extracted and enriched for HSC and early progenitor cells. These cells are transduced, or labeled, by the lentiviruses, and are then infused back into the irradiated monkeys. Since irradiation depletes the residual blood cells, reconstitution of the blood system is supported by these extracted cells following transplantation.

Hematopoietic reconstitution is monitored indirectly by taking samples of the blood cells at observation intervals ranging from several weeks to months. All cells descended from a marked cell inherit its unique barcode ID, thereby enabling lineage tracking. We note that by lineage we mean cells that descend from a marked cell. Such cells would be denoted as a clone in the hematopoiesis literature, where the word lineage is reserved for a cell type. At each observation time, the blood sample is sorted into monocyte (Mono), granulocyte (Gr), T, B, and natural killer (NK) cell types. Next, polymerase chain reaction (PCR) is performed on purified DNA samples from each sorted cell population, and barcodes are retrieved from the PCR product using Illumina sequencing. Sequences are filtered in such a way that only barcode IDs with numbers of reads exceeding a specified minimum read threshold remain in the dataset, reducing the effect of nonlinearities and noise arising from the PCR procedure in the pool of sequences we work with. Thus, at each observation time, the experimental protocol yields a read count corresponding to each barcode ID present in each cell type sample. Together, read count data for a given barcode ID constitute an independent time series that informs us of contributions to different cell types over time. Restricting our attention to only those barcodes exceeding a threshold of 1000 read counts at any observation time similarly to [Wu et al., 2014], we arrive at the dataset consisting of over 110 million read counts across 9635 unique barcode IDs, observed at irregular time intervals over a total period of 30 months. An illustration summarizing the process after transplantation corresponding to one clone is provided in Figure 1.

The observed dataset is the collection of read counts for each mature cell type m and can be represented by $P \times J$ matrices $\mathbf{Y}_m = (\mathbf{y}_m^1, \mathbf{y}_m^2, \ldots, \mathbf{y}_m^J)$ whose columns correspond to observation times $\mathbf{t} = (t_1, \ldots, t_J)$. Each pth row encodes the read count time series corresponding to a unique barcode ID $p \in 1, \ldots, P$ among the cell type population associated with this matrix.

To analyze the data, we first assume that the hematopoietic process evolves according to a

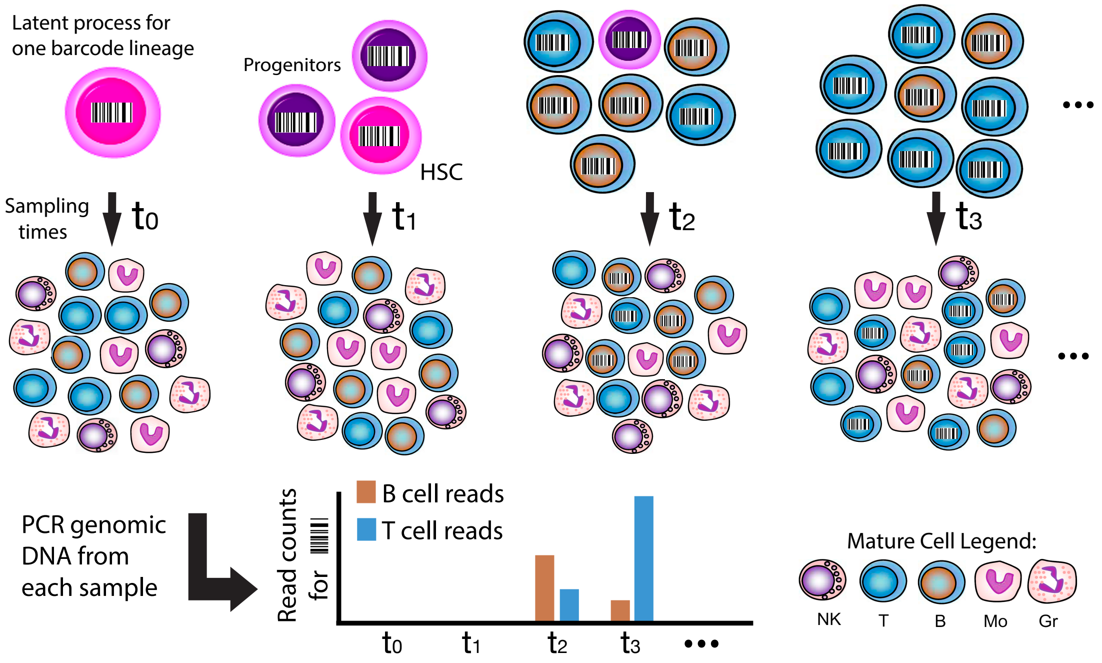


Figure 1: Illustration of experimental protocol for one single fixed barcode ID. The top panel represents the latent process starting with a single HSC (pink) at several snapshots in time $t_0, \ldots, t_3$. The second panel illustrates blood samples. Note that the barcode only becomes present in the blood when mature cells, produced by time t_2 in this example, are sampled in blood; the HSCs and early progenitors (purple) reside in the marrow and thus are unobservable. Read counts corresponding to the given barcode after PCR and sequencing reflect the number of cells sharing that barcode in the sample, which in turn reflect the barcoded population in the latent process.

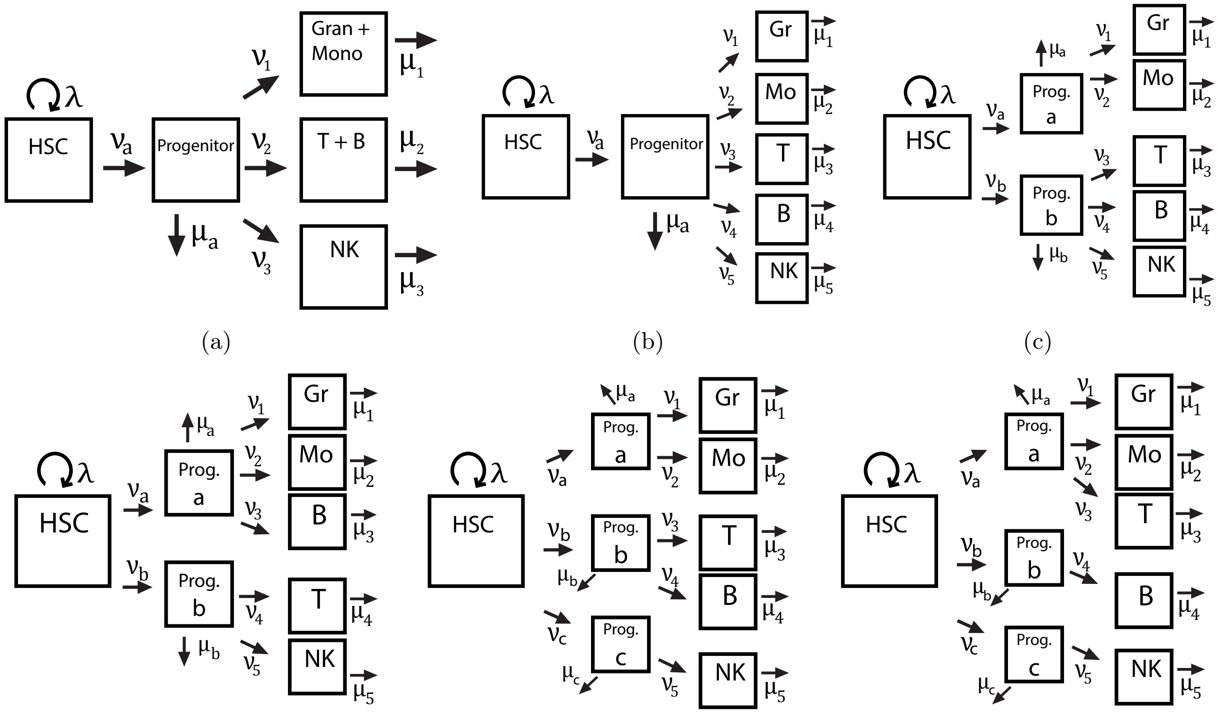
continuous-time Markov branching process. The choice of a branching process model is natural, as canonical differentiation trees that have been posited in the scientific literature follow such a structure, and equivalent stochastic models have been established and successfully studied in the statistical hematopoiesis literature [Kimmel and Axelrod, 2002, Catlin et al., 2011].

2.1 Stochastic branching model formulation

A branching process is a Markov process in which a collection of *independently acting* particles (cells) can reproduce and die according to a probability distribution. Here we consider a continuous-time, multi-type branching process taking values over a discrete state space of cell counts. In this setting, each particle type has a distinct mean lifespan and reproductive probabilities, and can give rise to particles of its own type as well as other types at its time of death.

For concreteness, notation is introduced for the branching process corresponding to Figure 2 (a). The process is a stochastic vector $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_5(t))$ taking values in state space $\Omega = \mathbb{N}^5$, where $X_i(t)$ denotes the number of type i cells at time $t \geq 0$. Each type i cell produces j type 1 particles, l type 2 particles, m type 3 particles, n type 4 particles, and m type 5 particles at *instantaneous rates* $a_i(j,k,l,m,n)$ upon completion of its lifespan. The rate of no event occurring, beginning with one type 1 particle, is defined as $\alpha_1 := a_1(1,0,0,0,0) = -\sum_{(j,k,l,m,n)\neq(1,0,0,0,0)} a_1(j,k,l,m,n)$, with α_i defined analogously, so that $\sum_{j,k,l,m,n} a_i(j,k,l,m,n) = 0$ for $i = 1, \ldots, 5$.

Particle independence implies that the process is *linear*: overall rates are multiplicative in the



(a) (b) (c)

(d) (e) (f)

Figure 2: Differentiation trees to be considered in simulation study and real data analysis. In the first two models, mature cells are descended from one common multipotent progenitor: (a) groups mature cells in a model with three total mature cell compartments, and (b) assigns each mature cell type its own compartment. Note that previous statistical studies by Catlin et al. [2001], Golinelli et al. [2006], Fong et al. [2009] have modeled only the first two compartments. Models (c)—(f) include several biologically plausible topologies featuring two or three oligopotent progenitors, each specializing to produce only particular mature cells.

number of particles. For example, in such a process, the infinitesimal probability of jumping to $\mathbf{X}(h) = (j,k,l,m,n)$ beginning with K type 1 particles over a short interval of time h is

$$\begin{aligned}\Pr_{(K,0,0,0,0),(j,k,l,m,n)}(h) &:= \Pr\left\{\mathbf{X}(h) = (j,k,l,m,n) | \mathbf{X}(0) = (K,0,0,0,0)\right\} \\ &= K \cdot a_1(j,k,l,m,n) \cdot h + o(h).\end{aligned}$$

Subsequently, offspring of each particle evolve according to the same set of instantaneous rates, and these rates $a_i(j,k,l,m,n)$ do not depend on t so that the process is *time-homogeneous*. Together these assumptions imply that each type i particle has exponentially distributed lifespan with rate $-\alpha_i$, and $\mathbf{X}(t)$ evolves over time as a CTMC [Guttorp, 1995, Chapter 3].

As depicted in Figure 2 (a), the parameters $\lambda, \nu_a, \mu_a, \nu_1, \nu_2, \nu_3, \mu_1, \mu_2$, and μ_3 define the infinitesimal rates of the process. The rates denoted ν_i correspond to differentiation, while μ_i denotes cell death/exhaustion; λ denotes HSC self-renewal. Specifying such a process classically using the infinitesimal generator or CTMC rate matrix is mathematically unwieldy, as this is an infinite matrix with no simplifying structure. However, in terms of branching process rates, these event rates can now be equivalently and compactly expressed as

$$a_1(2,0,0,0,0) = \lambda, \quad a_1(0,1,0,0,0) = \nu_a, \quad a_1(1,0,0,0,0) = -(\lambda + \nu_a), \quad a_2(0,0,0,0,0) = \mu_a,$$

$$a_2(0,1,1,0,0) = \nu_1, \quad a_2 = (0,1,0,1,0) = \nu_2, \quad a_2(0,1,0,0,1) = \nu_3, \quad a_2(0,1,0,0,0) = -(\mu_a + \nu_1 + \nu_2),$$

$$a_3(0,0,0,0,0) = \mu_1, \quad a_3(0,0,1,0,0) = -\mu_1, \quad a_4(0,0,0,0,0) = \mu_2,$$

$$a_4(0,0,0,1,0) = -\mu_2, \quad a_5(0,0,0,0,0) = \mu_3, \quad a_5(0,0,0,0,1) = -\mu_5,$$

with all other rates zero. Thus, the process is characterized by parameters $\boldsymbol{\theta} = (\lambda, \nu_a, \mu_a, \nu_1, \nu_2, \nu_3, \mu_1, \mu_2, \mu_3, \pi_a)$ containing the rates and initial distribution parameter π_a representing the probability that the clone is originally descended from a progenitor rather than from an HSC. In models with more than one progenitor compartment, the initial distribution is parametrized by a vector $\boldsymbol{\pi} = (\pi_a, \pi_b, \ldots)$.

2.2 Observation model

To complete the data generating model, it is necessary to specify the probability distribution of the barcode read counts conditional on the state of the branching process $\mathbf{X}(t)$. Read counts are observed between mature blood cells, and recall we denote these counts for cell type m corresponding to barcode p at time t by $y^p_m(t)$. Read counts are assumed proportional to the number of blood cells with barcode p *sampled* at time t, denoted $\widetilde{y}^p_m(t)$, so that $y^p_m(t) = d_m(t) \times \widetilde{y}^p_m(t)$ where constants $d_m(t)$ reflect the results of PCR amplification at time t. Such a linear representation of PCR amplification is standard after applying minimum read count thresholds that already ameliorate noise and nonlinearities in the amplification process. However, this has not accounted for uncertainty due to sampling: recall that at each observation time point, a fixed number of cells of each type is obtained from the blood sample. Within the purified DNA samples, a random number of barcodes is present, sampled in proportion to their prevalence in the cell population. Therefore, the distribution of sampled cells can be well-modeled by a multivariate hypergeometric distribution

$$\widetilde{\mathbf{y}}_m(t) \mid \mathbf{X}(t) \sim \text{mvhypergeom}(B_m, \mathbf{X}_m(t), b_m), \tag{1}$$

where b_m is the known number of sampled type m cells, B_m is the total number of barcoded cells of type m in the animal, and $\mathbf{X}_m(t)$ again represents the state of the underlying branching process, whose pth components contain the numbers of type m cells with barcode p. Note that b_m, B_m are known based on the experimental protocol, while $\mathbf{X}_m(t)$ is unknown. The pth component of the probability mass vector $\widetilde{y}_m^p(t)$ can be interpreted as the probability of drawing $\widetilde{y}_m^p$ balls of color p out of an urn containing B_m total balls, $X_m^p(t)$ of which are of color p, in a sample of size b_m. In this setting, each color corresponds to a barcode ID; the distributional choice is driven by its close mechanistic resemblance to the experimental sampling itself.

3 Methods

When feasible, likelihood methods for CTMC model-based inference are often preferable as they are most statistically efficient. However, the likelihood in our setting is intractable for two reasons: the observed data likelihood of the latent branching process is already computationally unwieldy—recent numerical techniques to compute this likelihood for multi-type branching processes and efforts to scale these techniques [Xu et al., 2015, Xu and Minin, 2015] fall short in our application due to the potential sizes of barcoded mature cell populations, which reach hundreds of thousands per type for a single barcode. Further, marginalizing over all possible configurations of unobserved compartments and underlying cell populations that are consistent with observed reads requires an additional integration step over an enormous hidden state space. Without an expression to analytically integrate out the hidden variables, alternatives such as data augmentation are notoriously difficult when the discrete hidden space is large. Although HMMs have been extensively studied, likelihood-based inference for HMMs is generally intractable when the state space of the hidden Markov process is infinite or finite but massive [Cappé et al., 2006]. On the other hand, populations of HSCs and early progenitors of a given barcode are likely to be very low and near zero, rendering approximations such as diffusion processes and other continuous-space representations ill-suited.

In lieu of feasible likelihood methods, we consider inference based on the generalized method of moments, a computationally simpler alternative to maximum likelihood estimation that yields consistent estimators. This method relies on deriving equations relating a set of population moments to the target model parameters to be estimated. Next, the discrepancy between the population and sample moments are minimized to estimate parameters of interest. Although moment-based estimators are known to be less statistically efficient than MLEs, the choice is well-motivated for our dataset consisting of thousands of barcoded clones, each acting as an independent, identically distributed realization from the model. Perhaps more appealing than their simplicity, moment-based methods feature more robustness to model misspecification than techniques relying on a completely prescribed likelihood [Wakefield, 2013]. Similar approaches have found success in application to stochastic kinetic models [Lakatos et al., 2015] and toward developing quasi- and pseudo-likelihood estimators [Chen and Hyrien, 2011].

Our estimator seeks to match pairwise empirical read count correlations across barcodes with their corresponding model-based population correlations. We derive explicit analytic forms for the first and second moments of a general class of branching models for hematopoiesis, allowing for the computation of marginal correlations between any two mature types. The advantage of working with correlations in the data is twofold: first, the observed correlation profiles between types are more time-varying and thus more informative than the mean and variance curves of read counts. Second, because correlations are scale invariant, we do not need to additionally model and estimate the effect of PCR amplification and fluctuations of absolute cell numbers on read counts in an already complex model. This robustness comes with a caveat — we may not expect all branching process rates to be identifiable with a scale free approach, instead requiring some parameters be fixed to provide scale information. This will be further discussed in Section 4.1.

With closed form moment expressions, model-based correlations can be computed very efficiently given any parameter setting, enabling the use of generic optimization methods to minimize a loss function relating the model-based correlations to observed correlations in the read count data. The following derivations apply to a rich class of models, including the candidate models displayed in Figure 2, with arbitrarily many compartments at the progenitor level and mature cell level, enabling us to investigate arbitrary groupings of cell types and candidate branching pathways.

3.1 Correlation loss function

To estimate the parameter vector $\boldsymbol{\theta}$, consisting of branching process rates and initial distribution $\boldsymbol{\pi}$, we seek to match model-based correlations closely to the empirical correlations between observed read counts. This is achieved by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta};\mathbf{Y}) = \sum_{t_j}\sum_{m}\sum_{n\neq m}\left[\psi^j_{mn}(\boldsymbol{\theta};\mathbf{Y}) - \hat{\psi}^j_{mn}(\mathbf{Y})\right]^2, \tag{2}$$

where ψ^j_{mn} represents model-based correlation between reads of type m and n cells at time t_j:

$$\psi^j_{mn}(\boldsymbol{\theta};\mathbf{Y}) := \rho(Y_m(t_j), Y_n(t_j);\boldsymbol{\theta}) = \frac{\mathrm{Cov}[Y_m(t_j), Y_n(t_j);\boldsymbol{\theta}]}{\sigma(Y_m(t_j);\boldsymbol{\theta})\sigma(Y_n(t_j);\boldsymbol{\theta})},$$

and $\hat{\psi}^j_{mn}$ denotes corresponding sample correlations across realizations $p = 1,\ldots,N$ at time t_j, where N denotes the total number of barcode IDs:

$$\hat{\psi}^j_{mn}(\mathbf{Y}) := \hat{\rho}(\mathbf{y}_m(t_j), \mathbf{y}_n(t_j) = \frac{\sum_{p=1}^N (y^p_m(t_j) - \overline{y}_m(t_j))(y^p_n(t_j) - \overline{y}_n(t_j))}{\sqrt{\sum_{p=1}^N (y^p_m(t_j) - \overline{y}_m(t_j))^2}\sqrt{\sum_{p=1}^N (y^p_n(t_j) - \overline{y}_n(t_j))^2}}.$$

Underlying (2) is the system of moment equations $\left\{\psi^j_{mn}(\boldsymbol{\theta};\mathbf{Y}) = \hat{\psi}^j_{mn}\right\}$ equating theoretical normalized moments with their sample analogs at each time t_j. Because the dataset contains more constraints than parameters to be estimated, the loss function is motivated by minimizing the residuals as a nonlinear least squares objective. The problem of estimating hematopoietic rates now translates to seeking

$$\hat{\boldsymbol{\theta}}_N = \underset{\boldsymbol{\theta}}{\mathrm{argmin}}\,\mathcal{L}(\boldsymbol{\theta};\mathbf{Y}) = \underset{\boldsymbol{\theta}}{\mathrm{argmin}}\|\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})\|_2^2, \qquad \text{where } \mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y}) := \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) - \hat{\boldsymbol{\psi}}(\mathbf{Y}),$$

and $\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}), \hat{\boldsymbol{\psi}}(\mathbf{Y})$ are vectors containing all pairwise model-based and empirical correlations at each time point, respectively. Note that N is also the number of rows in the data matrix $\mathbf{Y}$, from which the dependence of $\mathbf{G}_N$ on N arises.

If $\boldsymbol{\theta}_0$ are the true data generating parameters, then $\mathrm{E}[\mathbf{G}_N(\boldsymbol{\theta}_0;\mathbf{Y})] \to 0$ as the number of processes $N \to \infty$. Our method is therefore akin to a *z-estimator* or estimating equations approach [Van der Vaart, 2000, Chapter 5], which typically assumes $\mathrm{E}[\mathbf{G}_N(\boldsymbol{\theta}_0;\mathbf{Y})] = 0$ for all sample sizes N, and therefore yields a consistent estimator as the zero of the estimating equations $\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N;\mathbf{Y}) = \mathbf{0}$. While we minimize $\|\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})\|_2^2$ rather than root-finding and do not have unbiasedness for every N, our loss function estimator $\hat{\boldsymbol{\theta}}_N$ is also shown to be consistent. We prove the following result in the Appendix within a generalized method of moments (GMM) framework:

Theorem 3.1 *Assume the observed process $Y(t)$ has finite first and second moments, and assume the true parameter vector $\boldsymbol{\theta}_0$ is identifiable, i.e.* $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1} = O_p(1)$ *for each $\delta > 0$. Then $\left\{\hat{\boldsymbol{\theta}}_N\right\}$ converges in probability to $\boldsymbol{\theta}_0$, where $\hat{\boldsymbol{\theta}}_N = \mathrm{argmin}_{\boldsymbol{\theta}}\,\mathcal{L}(\boldsymbol{\theta};\mathbf{Y})$.*

In addition to serving as a useful context for analyzing properties of $\hat{\boldsymbol{\theta}}_N$, it is worth mentioning that the GMM framework provides a natural extension of our loss function estimator by replacing the ℓ^2 norm $\|\cdot\|_2$ by a general family of norms $\|\cdot\|_W$ induced by positive definite weight matrices $\mathbf{W}$. The estimator is now given by

$$\hat{\boldsymbol{\theta}}_W = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{G}_N(\boldsymbol{\theta}; \mathbf{Y})\|_W^2 := \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbf{G}_N(\boldsymbol{\theta}; \mathbf{Y})^T \,\hat{\mathbf{W}}\, \mathbf{G}_N(\boldsymbol{\theta}; \mathbf{Y});$$

notice minimization of $\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})$ is the special case of $\hat{\mathbf{W}} = \mathbf{I}$. The norm induced by $\mathbf{W}$ allows different moment equations to have unequal contributions to the objective function, and its estimate $\hat{\mathbf{W}}$ from the data intuitively assigns less weight to components which have higher variance and thus provide less information. GMM estimators $\hat{\boldsymbol{\theta}}_W$ enjoy asymptotic normality under additional regularity assumptions which we do not impose here [Pakes and Pollard, 1989, Van der Vaart, 2000], and are furthermore asymptotically efficient under optimal choice of $\hat{\mathbf{W}}$ [Hansen, 1982]. While many algorithms exist for estimating the weight matrix $\hat{\mathbf{W}}$, the task is nontrivial [Hansen et al., 1996]. Because we have a large enough dataset such that finite-sample efficiency is of lesser concern and we do not expect particular time points or correlation pairs to be more informative than others, we opt for the simple case with $\hat{\mathbf{W}} = \mathbf{I}$, avoiding the inclusion of many additional entries of the weight matrix as parameters to be estimated.

Having established the data generating model and estimation framework, next we derive the second moments of the latent process $\mathbf{X}(t)$ using branching process techniques. While this enables us to compute model-based correlations of the branching process, we must then relate these quantities to those in the observed process $\mathbf{Y}(\mathrm{t})$: we do so by connecting correlations of $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ via laws of iterated expectations and (co)variances.

3.2 Moments of the compartmental process

Here we derive analytic expressions for the first and second moments of the latent branching process, enabling efficient computation of model-based correlations $\psi^j(\boldsymbol{\theta}, \mathbf{Y})$ appearing in the loss function. Our approach is similar to the random variable technique introduced by Bailey [1964], but we derive expressions by way of probability generating functions rather than appealing to the cumulants. The derivation applies to the general class of models consisting of an HSC stage, progenitor stage, and mature cell stage, with arbitrary number of progenitor compartments and mature cell compartments, including all structures depicted in Figure 2. In this class, each mature cell type m is descended from only one progenitor compartment, so that its corresponding differentiation rate ν_m is unique and well-defined. The subscript 0 indicates rates relating to HSCs, and we use indices $a \in \mathcal{A}$ to denote progenitors, with mature cell types denoted by $m \in \mathcal{M}$. All intermediate progenitors are descended from the HSC compartment, and we use the notation $\{a \to m\}$ if progenitor a gives rise to type m mature cells, thus completely specifying a given branching model. The total number of compartments or cell types is denoted by C, and we use the notation $\mathbf{e}_i$ to represent the vector of length C whose ith entry equals 1 and is 0 elsewhere.

From applying the process rates to the Kolmogorov backward equations, we can write *pseudo-generating functions* defined as

$$u_i(\mathbf{s}) = \sum_{k_1} \sum_{k_2} \cdots \sum_{k_C} a_i(k_1, \ldots, k_C) s_1^{k_1} s_2^{k_2} \cdot s_C^{k_C}, \tag{3}$$

where $\mathbf{s}$ is a vector of dummy variables. For our class of models, these are given by

$$\begin{aligned}
u_0(\mathbf{s}) &= \lambda s_0^2 + \sum_{a\in\mathcal{A}} \nu_a s_a - \left(\lambda + \sum_{a\in\mathcal{A}} \nu_a\right) s_0, \\
u_a(\mathbf{s}) &= \sum_{m\in\mathcal{M}} \nu_m s_a s_m \mathbf{1}_{\{a\to m\}} + \mu_a - \left(\mu_a + \sum_{m\in\mathcal{M}} \nu_m \mathbf{1}_{\{a\to m\}}\right) s_a \qquad \forall a \in \mathcal{A}, \\
u_m(\mathbf{s}) &= u_m(s_m) = \mu_m - \mu_m s_m, \qquad\quad \forall m \in \mathcal{M}.
\end{aligned}$$

Next, we can write the probability generating function (PGF) of the process, beginning with one type 1 (HSC) particle, via a relation to the pseudo-generating function u_1 as follows:

$$\begin{aligned}
\phi_1(t;\mathbf{s}) &= \mathrm{E}\left[\prod_{j=1}^{C} s_j^{X_j(t)} | \mathbf{X}(0) = \mathbf{e}_1\right] = \sum_{k_1=0}^{\infty} \cdots \sum_{k_C=0}^{\infty} \mathrm{Pr}_{\mathbf{e}_1,(k_1,k_2,\ldots k_C)}(t) s_1^{k_1} s_2^{k_2} \cdots s_C^{k_C} \\
&= \sum_{k_1=0}^{\infty} \cdots \sum_{k_C=0}^{\infty} \left[\mathbf{1}_{\{k_1=1,k_2=\ldots=k_C=0\}} + a_1(k_1,\ldots,k_C)t + o(t)\right] s_1^{k_1} s_2^{k_2} \cdots s_C^{k_C} \\
&= s_1 + u_1(\mathbf{s})t + o(t). \qquad (4)
\end{aligned}$$

Analogously defining ϕ_i for processes beginning with one type i particle for each $i = 1,\ldots,C$, Equation (4) yields the relation

$$\frac{\partial}{\partial t}\phi_i(t,\mathbf{s}) = u_i(\phi_1(t,\mathbf{s}),\ldots,\phi_C(t,\mathbf{s})).$$

Only expressions conditioning on one initial particle are required throughout, since each latent process represents cells sharing a unique genetic barcode, which is always descended from a single marked cell. Now, let $M_{l|k}(t)$ denote the expected number of type l cells at time t, given one initial type k cell. From definition of ϕ_i, we see that we can relate the probability generating functions to these first moments via partial differentiation:

$$M_{l|k}(t) = \frac{\partial}{\partial s_l}\phi_k(t,\mathbf{s})|_{s_1=s_2=\ldots=s_C=1}.$$

Similarly, we may further differentiate the PGF to derive second moments used toward variance and covariance calculations. Define

$$U_{kl|1}(t) = \mathrm{E}\left[X_k(X_l - \mathbf{1}_{\{k=l\}})|\mathbf{X}(0) = \mathbf{e}_1\right],$$

with $U_{kl|i}(t)$ defined analogously beginning with one type i particle. Then $U_{kl|j}(t) = \left.\frac{\partial^2\phi_j}{\partial s_k \partial s_l}\right|_{\mathbf{s}=\mathbf{1}}$.

This relationship via partial differentiation enables us to write a system of differential equations governing the moments. Applying the multivariate chain rule and the Faà di Bruno formula,

$$\frac{\partial}{\partial t}M_{j|i}(t) = \frac{\partial^2\phi_i}{\partial t \partial s_j} \mid_{\mathbf{s}=\mathbf{1}} = \sum_k \frac{\partial u_i}{\partial s_k}\frac{\partial \phi_k}{\partial s_j} \mid_{\mathbf{s}=\mathbf{1}}, \qquad (5)$$

$$\frac{\partial}{\partial t}U_{jk|i}(t) = \frac{\partial^3\phi_i}{\partial t \partial s_j \partial s_k} \mid_{\mathbf{s}=\mathbf{1}} = \sum_{m=1}\left(\frac{\partial u_i}{\partial \phi_m}\frac{\partial^2\phi_m}{\partial s_j \partial s_k}\right) + \sum_{m,n=1}\left(\frac{\partial^2 u_i}{\partial \phi_m \partial \phi_n}\frac{\partial \phi_m}{\partial s_j}\frac{\partial \phi_k}{\partial s_k}\right) \mid_{\mathbf{s}=\mathbf{1}}. \qquad (6)$$

Notice equation (5) defines a system of ordinary differential equations (ODEs) determining the mean behavior, whose solutions can be plugged in to solve the second system of equations (6) governing second moments. These systems are subject to the initial conditions $M_{j|i}(0) = \mathbf{1}_{\{i=j\}}$, $U_{jk|i}(0) = 0$ for all i, j, k. For simplicity we introduce the notation $\kappa_{ij} = \frac{\partial u_i}{\partial s_j}\big|_{\mathbf{s}=\mathbf{1}}$: for instance,

$$\kappa_{00} = \lambda - \sum_{a \in \mathcal{A}} \nu_a, \quad \kappa_{aa} = -\mu_a, \quad \kappa_{mm} = -\mu_m, \quad \kappa_{0a} = \nu_a, \quad \kappa_{am} = \nu_m \mathbf{1}_{\{a \to m\}} \qquad \forall a \in \mathcal{A}, m \in \mathcal{M}.$$

The system for first moments is relatively straightforward: first, the means $M_{m|m}(t)$ where $m \in \mathcal{M}$ are simply solutions to pure death equations, so that

$$M_{m|m}(t) = e^{\kappa_{mm}t} = e^{-\mu_m t}.$$

These solutions can now be substituted into simple first moment equations conditional on beginning with a marked progenitor: from (5), these equations are given by

$$\frac{\partial}{\partial t} M_{m|a}(t) = \kappa_{aa} M_{m|a}(t) + \mathbf{1}_{\{a \to m\}} \kappa_{am} M_{m|m}(t),$$

and upon rearrangement are of the general form

$$\frac{d}{dt} M_{m|a}(t) + P(t) M_{m|a}(t) = Q(t). \tag{7}$$

Such a differential equation can be solved using the integrating factor method, multiplying both sides $e^{\int P(t)dt}$ and rearranging for $M_{m|a}(t)$. Solving, we obtain

$$M_{m|a}(t) = \mathbf{1}_{\{a \to m\}} \frac{\kappa_{am}}{\kappa_{aa} - \kappa_{mm}} \left(e^{\kappa_{aa}t} - e^{\kappa_{mm}t} \right) \quad = \quad \mathbf{1}_{\{a \to m\}} \frac{\nu_m}{\mu_m - \mu_a} \left(e^{-\mu_a t} - e^{-\mu_m t} \right).$$

Next, (5) again gives us mean equations conditional on beginning with one marked HSC:

$$\frac{\partial}{\partial t} M_{m|0}(t) = \kappa_{00} M_{m|0}(t) + \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \kappa_{0a} M_{m|a}(t),$$

which clearly is also of the form (7). Thus, we can plug in the solutions we've obtained for $M_{m|a}(t)$ and solve the system using the same technique, yielding

$$\begin{aligned} M_{m|0}(t) &= e^{\kappa_{00}t} \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \frac{\kappa_{0a}\kappa_{am}}{\kappa_{aa} - \kappa_{mm}} \left(\frac{e^{(\kappa_{aa} - \kappa_{00})t} - 1}{\kappa_{aa} - \kappa_{00}} - \frac{e^{(\kappa_{mm} - \kappa_{00})t} - 1}{\kappa_{mm} - \kappa_{00}} \right) \\ &= e^{(\lambda - \sum_a \nu_a)t} \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \frac{\nu_a \nu_m}{\mu_m - \mu_a} \left(\frac{e^{((\sum_a \nu_a) - \mu_a - \lambda)t} - 1}{(\sum_a \nu_a) - \mu_a - \lambda} - \frac{e^{((\sum_a \nu_a) - \mu_m - \lambda)t} - 1}{(\sum_a \nu_a) - \mu_m - \lambda} \right). \end{aligned}$$

These expressions characterize the mean behavior of the system, and furthermore may now be used toward solving for the second moments. We introduce for simplicity the additional notation $\kappa_{i,jk} := \frac{\partial^2 u_i}{\partial s_j \partial s_k}\big|_{\mathbf{s}=\mathbf{1}}$; for instance, $\kappa_{0,00} = 2\lambda$. Further, the equations $U_{mm|m}(t) = \kappa_{mm} U_{mm|m}(t)$, and together with the initial condition are only satisfied by the trivial solution $U_{mm|m}(t) = 0$ for all final types $m \in \mathcal{M}$. Now, many terms in equation (6) have zero contribution, and the remaining equations in the system can be simplified to yield

$$\begin{aligned} \frac{d}{dt} U_{mn|a}(t) &= \mathbf{1}_{\{a \to m\}} \mathbf{1}_{\{a \to n\}} \left(\frac{\partial u_a}{\partial s_a} \frac{\partial^2 \phi_a}{\partial s_m \partial s_n} + \frac{\partial^2 u_a}{\partial s_a \partial s_m} \frac{\partial \phi_a}{\partial s_n} \frac{\partial \phi_m}{\partial s_m} + \frac{\partial^2 u_a}{\partial s_a \partial s_n} \frac{\partial \phi_a}{\partial s_m} \frac{\partial \phi_n}{\partial s_n} \right) \\ &= \mathbf{1}_{\{a \to m\}} \mathbf{1}_{\{a \to n\}} \left(\kappa_{aa} U_{mn|a} + \kappa_{a,am} M_{n|a} M_{m|m} + \kappa_{a,an} M_{m|a} M_{n|n} \right) \qquad \forall a \in \mathcal{A}, m \neq n \in \mathcal{M}, \end{aligned}$$

$$\begin{aligned}\frac{d}{dt}U_{mn|0}(t) &= \left(\frac{\partial u_0}{\partial s_0}\frac{\partial^2\phi_0}{\partial s_m s_n} + 2\frac{\partial^2 u_0}{\partial s_0^2}\frac{\partial\phi_0}{\partial s_m}\frac{\partial\phi_0}{\partial s_n} + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\frac{\partial u_0}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m s_n}\right)\Big|_{\mathbf{s}=\mathbf{1}}\\ &= \kappa_{00}U_{mn|0} + 2\kappa_{0,00}M_{m|0}M_{n|0} + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\kappa_{0a}U_{mn|a} \qquad \forall m\neq n\in\mathcal{M}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{d}{dt}U_{mm|a}(t) &= \mathbf{1}_{\{a\to m\}}\left(\frac{\partial u_a}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m^2} + 2\frac{\partial^2 u_a}{\partial s_a\partial s_m}\frac{\partial\phi_a}{\partial s_m}\frac{\partial\phi_m}{\partial s_m} + 0\right)\\ &= \mathbf{1}_{\{a\to m\}}\left(\kappa_{aa}U_{mm|a} + 2\kappa_{a,am}M_{m|a}M_{m|m}\right) \qquad \forall a\in\mathcal{A}, m\in\mathcal{M},\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}U_{mm|0}(t) &= \left[\frac{\partial u_0}{\partial s_0}\frac{\partial^2\phi_0}{\partial s_m^2} + \frac{\partial^2 u_0}{\partial s_0^2}\left(\frac{\partial\phi_0}{\partial s_m}\right)^2 + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\frac{\partial u_0}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m^2}\right]\Big|_{\mathbf{s}=\mathbf{1}}\\ &= \kappa_{00}U_{mm|0} + \kappa_{0,00}M_{m|0}^2 + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\kappa_{0a}U_{mm|a} \qquad \forall m\in\mathcal{M}.\end{aligned}$$

Since we already have expressions for the means $M_{\cdot|\cdot}$, these equations $U_{\cdot|a}(t)$ each become a first order linear ODE and can now each be solved individually. Indeed, they again take the form (7), and we find

$$\begin{aligned}U_{mm|a}(t) &= \mathbf{1}_{\{a\to m\}}e^{\kappa_{aa}t}\int_0^t 2\cdot e^{-\kappa_{aa}x}\kappa_{a,am}M_{m|a}(x)M_{m|m}(x)\,dx,\\ U_{mn|a}(t) &= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}e^{\kappa_{aa}t}\int_0^t e^{-\kappa_{aa}x}\left(\kappa_{a,am}M_{n|a}(x)M_{m|m}(x) + \kappa_{a,an}M_{m|a}(x)M_{n|n}(x)\right)\,dx.\end{aligned}$$

Replacing $\kappa_\cdot$ with model-based rates, we integrate and simplify these expressions to obtain

$$\begin{aligned}U_{mm|a}(t) &= \mathbf{1}_{\{a\to m\}}\frac{2\nu_m^2}{\mu_m-\mu_a}e^{-\mu_a t}\left[\frac{\mu_a-\mu_m}{\mu_m(\mu_a-2\mu_m)} - \frac{e^{-\mu_m t}}{\mu_m} - \frac{e^{(\mu_a-2\mu_m)t}}{\mu_a-2\mu_m}\right]\\ U_{mn|a}(t) &= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\left\{\frac{\nu_m\nu_n}{\mu_n-\mu_a}e^{-\mu_a t}\left[\frac{\mu_a-\mu_n}{\mu_m(\mu_a-\mu_m-\mu_n)} - \frac{e^{-\mu_m t}}{\mu_m} - \frac{e^{(\mu_a-\mu_m-\mu_n)t}}{\mu_a-\mu_m-\mu_n}\right]\right.\\ &\qquad\left. + \frac{\nu_m\nu_n}{\mu_m-\mu_a}e^{-\mu_a t}\left[\frac{\mu_a-\mu_m}{\mu_n(\mu_a-\mu_m-\mu_n)} - \frac{e^{-\mu_n t}}{\mu_n} - \frac{e^{(\mu_a-\mu_m-\mu_n)t}}{\mu_a-\mu_m-\mu_n}\right]\right\}.\end{aligned}$$

Finally, we plug in these solutions into the differential equations beginning with an HSC governing $U_{\cdot|0}(t)$, which now take on the same general form and again can be solved by the integrating factor method:

$$\begin{aligned}U_{mn|0}(t) &= e^{\kappa_{00}t}\int_0^t e^{-\kappa_{00}x}\left(\kappa_{0,00}M_{n|0}(x)M_{m|0}(x) + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\kappa_{0a}U_{mn|a}(x)\right)\,dx,\\ U_{mm|0}(t) &= e^{\kappa_{00}t}\int_0^t e^{-\kappa_{00}x}\left(\kappa_{0,00}M_{m|0}^2(x) + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\kappa_{0a}U_{mm|a}(x)\right)\,dx.\end{aligned}$$

At this stage, we see that these integrals have closed form solutions as well, since their integrands only differ from the previous set of equations by including additional sums of exponentials from the

$U_{\cdot|a}(t)$ expressions. We omit the integrated forms in the general case for brevity, but remark that while they appear lengthy, they are comprised of simple terms and can be very efficiently evaluated, enabling use within iterative algorithms. For completeness, we include the explicit solutions to the simplest model in the Appendix.

With closed form moment expressions in hand, we can readily recover variance and covariance expressions and thus calculate model-based correlations. For instance,

$$\mathrm{Cov}\left[X_4(t), X_5(t)|\mathbf{X}(0) = \mathbf{e}_1\right] = U_{45|1}(t) - M_{4|1}(t)M_{5|1}(t).$$

Because the initial state is uncertain, unconditional variances and covariances between mature types can be computed by marginalizing over the initial distribution vector $\boldsymbol{\pi}$, with details in the Appendix. We thus arrive at the marginal expressions by applying the law of total (co)variance:

$$\begin{aligned}\mathrm{Var}[X_i(t)] &= \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{i|k}^2] - \sum_{k=1}^{K} \pi_k^2 (\mathrm{E}[X_{i|k}])^2) - 2\sum_{j>k} \pi_j \pi_k \mathrm{E}[X_{i|j}]\mathrm{E}[X_{i|k}] \\ &= \sum_{k=1}^{K} \pi_k [U_{ii|k}(t) + M_{i|k}(t)] - \pi_k^2 M_{i|k}(t)^2 - 2\sum_{j>k} \pi_j \pi_k M_{i|k}(t) M_{i|j}(t). \end{aligned} \tag{8}$$

$$\begin{aligned}\mathrm{Cov}[X_i(t), X_j(t)] &= \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{i|k} X_{j|k}] - \sum_{k=1}^{K} \pi_k^2 \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|k}] - \sum_{k \neq l} \pi_k \pi_l \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|l}] \\ &= \sum_{k=1}^{K} \pi_k U_{ij|k}(t) - \pi_k^2 M_{i|k}(t) M_{j|k}(t) - \sum_{k \neq l} \pi_k \pi_l M_{i|k}(t) M_{j|l}(t). \end{aligned} \tag{9}$$

3.3 Moments of observed read counts

Analytic expressions for the covariances and variances of the latent branching process enable calculation of pairwise correlations between latent mature cell populations, but it remains to relate these expressions to the correlations between read counts, $\psi_{mn}(\boldsymbol{\theta};\mathbf{Y})$, appearing in our loss function. Computing these correlations requires applying the laws of total variance and covariance to the moment expressions obtained for the latent branching process. Conditioning the previously derived expressions moment expressions on the multivariate hypergeometric sampling distribution, we obtain the following expressions comprising $\psi_{mn}(\boldsymbol{\theta};\mathbf{Y})$:

$$\begin{aligned}\mathrm{Cov}(Y_m, Y_n) &= \frac{b_m b_n}{B_m B_n} \mathrm{Cov}(X_m, X_n), \\ \mathrm{Var}(Y_m) &= \frac{b_m(B_m - b_m)}{B_m(B_m - 1)} \mathrm{E}(X_m) - \frac{b_m(B_m - b_m)}{B_m^2(B_m - 1)} \mathrm{E}(X_m^2) + \frac{b_m^2}{B_m^2} \mathrm{Var}(X_m). \end{aligned} \tag{10}$$

3.4 Implementation

We implemented our methods in R package `branchCorr`, available at `https://github.com/jasonxu90/branchCorr`. Software includes algorithms to simulate and sample from the class of stochastic compartmental models, to compute model-based moments given parameters, and to estimate parameters by optimizing the loss function objective. We provide a vignette that steps through smaller-scale versions of all simulations in this paper.

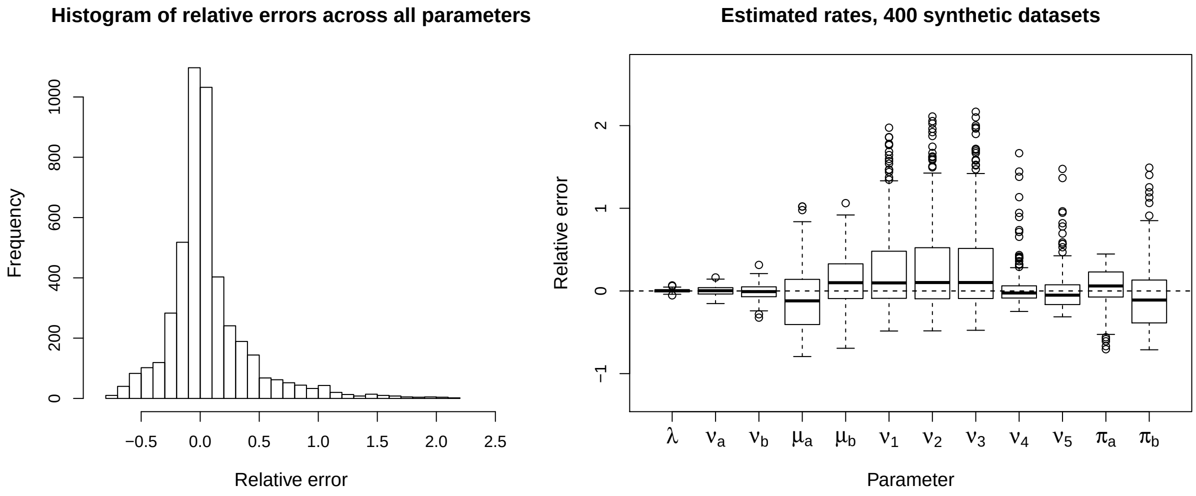


Figure 3: Performance of loss function estimator on synthetic data from model with five mature types and two progenitor compartments, i.e. model (c) or (d). While we see outlier influence, median estimates are accurate despite the parameter rich setting. Detailed medians, median absolute deviations, and standard errors corresponding to the plotted estimates are included in the Appendix.

4 Results

4.1 Simulation study

To assess our methods, we examine the performance of our loss function estimator on simulated data, generated from several hematopoietic tree structures in our branching process framework. Specifically, we consider models with three or five mature types with varying progenitor structures displayed in Figure 2. For each model, we simulate 400 independent datasets, each consisting of $20,000$ realizations representing distinct barcode IDs, from the continuous-time branching process model. True rates for simulating these processes were chosen such that summing over the $20,000$ barcodes, the total populations of each mature cell type are relatively constant after time $t = 2$, since true cell populations should be fairly constant for scientific realism. Note that while total populations are stable, individual barcode trajectories display a range of heterogeneous behaviors, with many trajectories becoming extinct and others reaching very high counts. This reflects the behavior we see in the real dataset.

From each of these synthetic datasets, we then produce an *observed dataset* by drawing samples of fixed size from the complete data according to the multivariate hypergeometric distribution, mimicking experimental sampling noise. Observations are recorded at irregular times over a two year period similar to the span and frequency of the experimental sampling schedule. Parameter estimation is then performed on these observed datasets.

To minimize the loss function objective, we use the general optimization implementation in package `nlminb`. Optimization is performed over 250 random restarts per observed dataset. We constrain rates to be non-negative, and include a simple log-barrier constraint to enforce that the overall growth of the HSC reserve is non-negative. In models with more than one progenitor cell, the initial distribution vector is constrained to a probability simplex. Rather than specifying additional hard constraints in the optimization problem, we use a multinomial logistic reparametrization

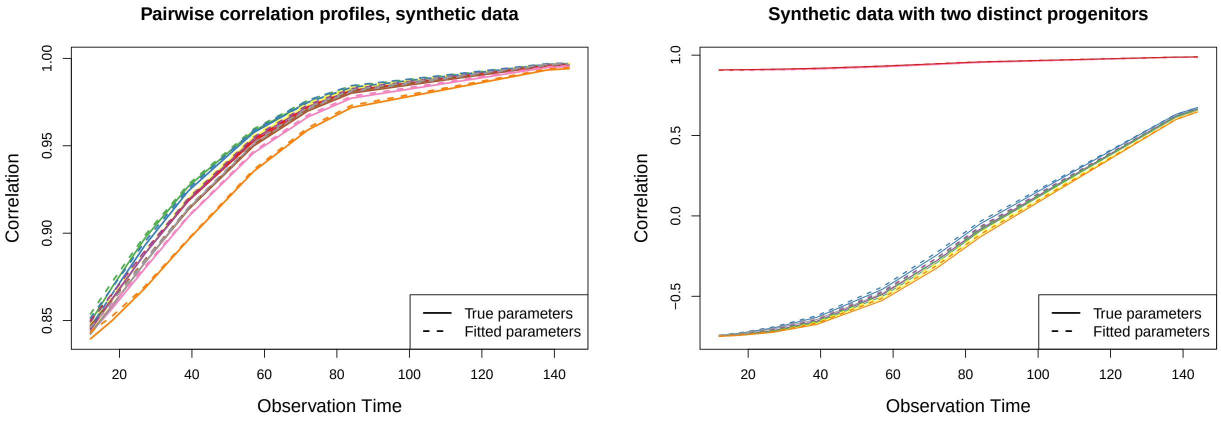


Figure 4: Pairwise correlation curves between five mature cell compartments descended from one common progenitor (left) or two distinct progenitors (right) calculated based on our point estimates. Solution curves from best fitting parameter estimates are almost indistinguishable from those corresponding to true parameters in both cases. Note that in the two-progenitor model, pairwise correlations among mature cell types display two distinct clusters of behavior, and that negative correlations are possible.

so that each initial distribution parameter varies freely in $\mathbb{R}$; see Appendix for details. Finally, we remark that optimization over all free parameters leads to mild identifiability problems—in particular, pairs of mature differentiation rates and death rates are often only identifiable up to a ratio. This is unsurprising: the correlations comprising the objective function are invariant to scale, so we would expect parameters to be distinguishable only up to a multiplicative constant. To remedy this, we choose to fix the death rates μ_i at their true value, supplying information that provides a sense of scale to infer all other parameters. Indeed, this is also justifiable in practice: mature cell types are observable in the bloodstream, and information about their behavior, i.e. average lifespans, is available in the scientific literature.

Correlation profiles from estimated parameters corresponding to the results in the tables above are displayed in Figure 4. Visually, we see the curves are very close to those corresponding to true parameters. We also note clear qualitative differences between models, with the two-progenitor model exhibiting two clear groupings of correlation profiles and exhibiting low and negative correlations.

Model misspecification In the following simulation experiments, we examine the performance of the estimator in under- and over-specified models. We do so by fitting incorrect models, assuming the data are generated from a model with one common progenitor or with three intermediate progenitors, to the data simulated from the two-progenitor model which we have fitted in the previous section. Recall that in the true model, mature types 1, 2, and 3 are descended from progenitor a, while the others are from progenitor b. Estimates reported in Figure 3 have near zero median relative error, and we note the median value of the objective function (2) at convergence was 2.78×10^{-4}, with median absolute deviation 1.31×10^{-4} and standard deviation 2.47×10^{-4}.

The fitted correlation curves in under- and over-specified progenitor structures are displayed in Figure 5, with detailed tables containing estimates again included in the Appendix. We also examine the behavior when fitting a model with fewer compartments by “lumping” similar mature types together. To this end, we consider grouping mature types 2 and 3 together, and types

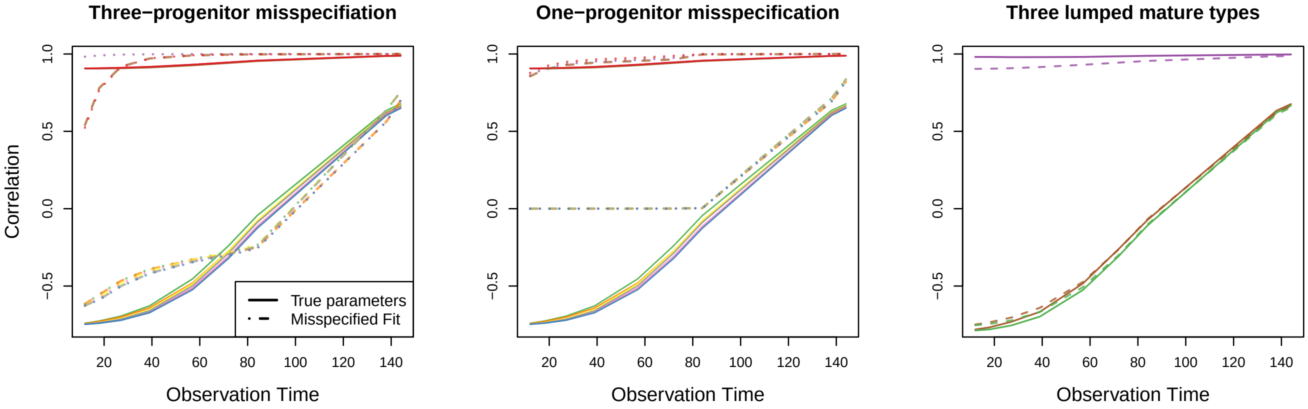


Figure 5: Fitted correlation curves corresponding to misspecified model estimates. Data are generated from a true model with two distinct progenitors and the true correlation profiles are the same as those displayed in the right panel of Figure 4. While we see a generic lack of fit in the three-progenitor model, notice that specifying one common progenitor fails to exhibit negative correlations necessary to explain the data. On the other hand, "lumping" mature compartments but properly specifying progenitor structure results in reasonable performance, as evident in the rightmost panel.

4 and 5 together, thus fitting a model with three total mature cell compartments, but with a progenitor structure consistent with the true model. Results in Figure 5 suggest it is reasonable to group cells with shared lineages together, resulting in a much milder effect on model fit than progenitor structure misspecification. Such a grouping strategy can be important toward avoiding overfitting a model to real data when some degree of model misspecification is inevitable, and will be advantageous in settings where limited data requires aggregation to fit a simpler model with fewer parameters.

4.2 Cell lineage barcoding in rhesus macaques

Having validated our method on simulated data from the model, we are ready to analyze the data from the lineage barcoding experiments from [Wu et al., 2014]. We consider barcoding data collected from a rhesus macaque over a 30 month period following transplantation. We consider only sampling times at which uncontaminated read data for each of the five cell types (granulocyte, monocyte, T, B, and Natural Killer) are available, and as in the original study, apply a filter so that we consider only clones exceeding a threshold of at least 1000 read counts at any time point. After restricting by these criteria, our dataset consists of 9635 unique barcode IDs, with read data available at eleven unevenly spaced sampling times.

As inputs to the loss function estimator, we fix death rates, reported below, at biologically realistic parameters based on previous studies [Hellerstein et al., 1999, Zhang et al., 2007, Kaur et al., 2008]. Parameters of the multivariate hypergeometric sampling distribution are informed by circulating blood cell (CBC) data recorded at sampling times. These include $B_m(t)$, the total population of type m cells in circulation at time t across all barcodes, and b_m, the constant number of type m cells in the sample at each observation time. Finally, the initial barcoding level for HSCs π_1 is informed by levels of green fluorescent protein (GFP) positivity, which stabilize after 3 months. Because only HSCs have long-term regenerative capacity, the stable GFP marking level

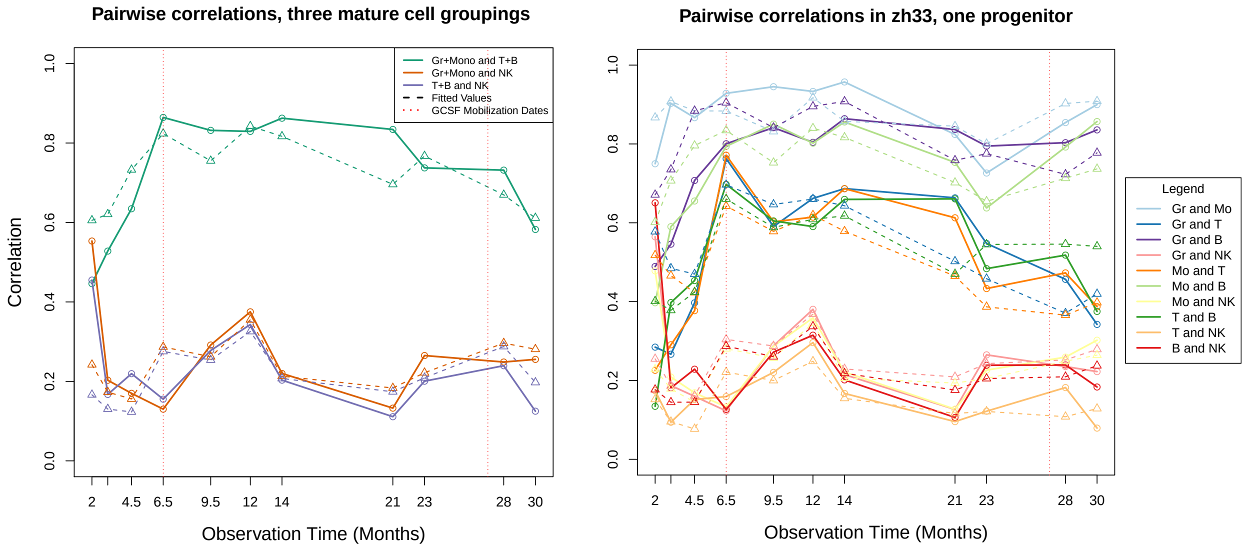


Figure 6: Dashed lines depict fitted correlations to read data in models (a) and (b) assuming one early progenitor compartment. GCSF mobilization dates are marked by vertical red lines. Solid lines connect the empirical correlations.

suggests the proportion of barcoded cells that were marked at the HSC stage as opposed to a later progenitor stage. While the GFP levels are observable and available to us, we will also infer π_1 independently of the GFP data in model (a) as additional validation.

We estimate the remaining rate parameters and initial barcoding distribution using the loss function estimator in all models displayed in Figure 2. Fitted pairwise correlation curves from estimates obtained loss function optimization with 2000 random restarts in models with one multipotent progenitor compartment are displayed in Figure 6: there are three such curves in the model with three mature compartments, with ten possible pairs among the model consisting of all five mature types in the plot on the right. The raw data correlations are also displayed as solid lines, and we comment that at a qualitative level, there is visible separation into three clusters of correlation profiles among the five mature cell groups, consistent with the choice of three lumped compartments in the simpler model (a). Notably, empirical correlations between NK cells and any other cell type are significantly lower than all other pairwise correlations. This supports the main result in the pilot clustering-based analysis in the original study [Wu et al., 2014], reporting on distinctive NK lineage behavior, from a new perspective. In both plots, fitted curves successfully follow the shape of observed correlations over time, and we observe that the largest error occurs at the 6.5 month sample, coinciding with the application of granulocyte-colony stimulating factor (GCSF), a technical intervention that perturbs normal hematopoiesis in the animal. The corresponding plots for models with multiple progenitors are included in the Appendix.

Next, we display a visual comparison of intermediate differentiation rates normalized as fate decision probabilities in Figure 7 and fitted self-renewal rates in Figure 8 across models. The complete set of parameter estimates (used to generate fitted curves in Figure 6) and their corresponding confidence intervals are reported in the Appendix. Confidence intervals are produced via 2500 bootstrap replicate datasets. Nonparametric bootstrap resampling was performed over barcode IDs as well as over read count sampling, to account for variation across stochastic realizations and from sampling noise.

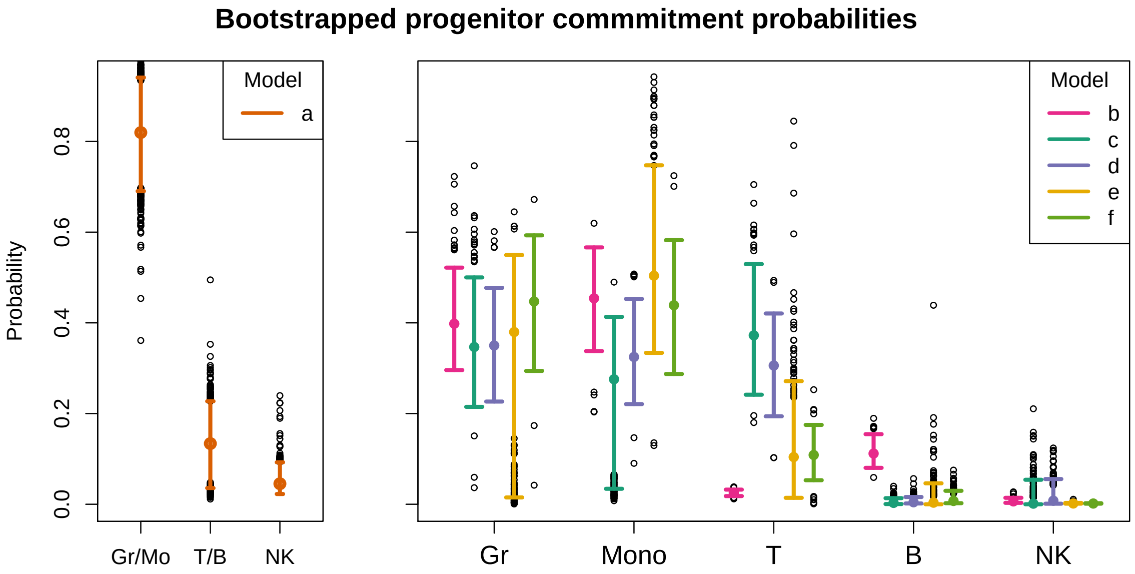


Figure 7: Comparison of fitted intermediate differentiation rates parametrized as fate decision probabilities. Displayed are the bootstrap estimates of normalized commitment rates to each mature compartment i, $\frac{\hat{\nu}_i}{\sum_j \hat{\nu}_j}$, in each model displayed in Figure 2 (a)-(f) fitted to rhesus macaque data.

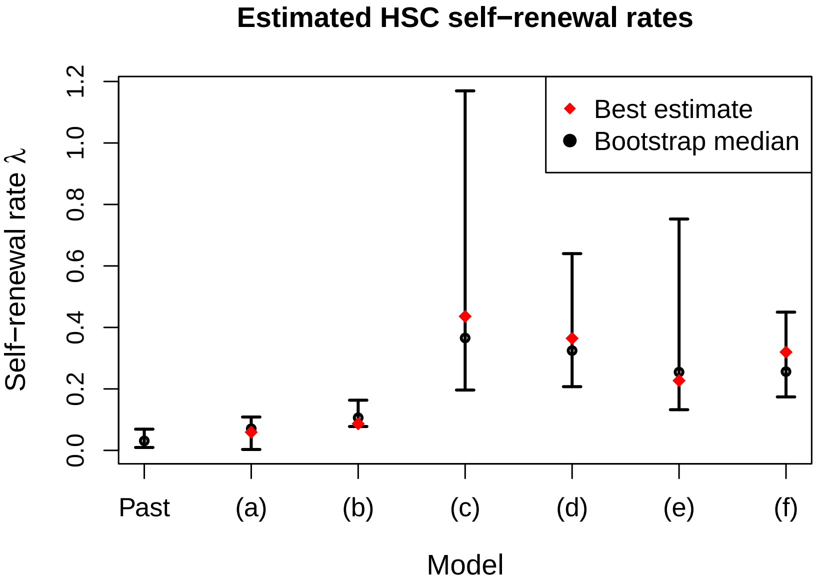


Figure 8: Comparison of fitted self-renewal rates $\hat{\lambda}$ and 95% confidence intervals across all models displayed in Figure 2 (a)-(f). Point estimates with lowest objective value (best estimates) are marked by red diamonds, while bootstrap confidence intervals and medians are plotted in black. The confidence interval around $\hat{\lambda}$ from model (a) overlaps with the interval obtained in previous telomere analyses focusing on HSC behavior in primates [Shepherd et al., 2007], while the interval from model (b) is very close and in reasonable range. The other models, which do not feature a multipotent common progenitor, result in less biologically plausible estimates.

Rate estimates are parametrized as number of events per five days: for instance, the fixed death rates $\boldsymbol{\mu} = (0.4, 0.04, 0.3)$ in the lumped model correspond to half-lives of about eight days among granulocytes and monocytes, three months for T and B cells, and two weeks in NK cells. In all models with five mature compartments, we fix death rates at $\boldsymbol{\mu} = (0.8, 0.3, 0.04, 0.08, 0.4)$.

Previous studies of HSC dynamics in nonhuman primates based on telomere analysis [Shepherd et al., 2007] estimate the HSC self-renewal rate at once every 23 weeks, with 11-75 week range, corresponding to an estimate of $\widetilde{\lambda} = 0.0310$, with interval $(0.0095, 0.0649)$ when translated to our parametrization. As we see in Figure 8, these findings coincide with our estimates and confidence intervals for $\hat{\lambda}$ in models with one multipotent progenitor compartment. While other rates pertaining to intermediate cell stages and initial barcoding level are quantities that have not been previously estimated, our results suggest that granulocytes and monocytes are produced much more rapidly than T, B and NK cells, and that individual progenitor cells are long-lived and can each produce thousands of these mature cells per day—biologically reasonable results that are newly supported from a statistical modeling perspective. Finally, we remark that the GFP data stabilize at around 13%. This level indicates the proportion of marked cells with long-term proliferative potential, suggesting the remaining 87% of barcoded cells are marked downstream at a progenitor stage. Holding out this information in Model 2(a), we estimate the initial progenitor marking level 86.1%, consistent with the GFP data as additional model validation.

In models (c)–(f) with multiple specialized, oligopotent progenitors compartments, we utilize the GFP data to fix the total progenitor marking level at 87% and estimate the proportion marked in each progenitor compartment. However, Supplementary Table C-8 shows that best estimates in these models lie on the boundary of the probability simplex. Along with wider confidence intervals, higher objective values, and less biologically plausible parameters, these results indicate a poorer model fit, reminiscent of the results discussed in the model misspecification experiments in Section 4.1.

While it may initially seem intuitive that a richer model with more compartments should result in a better fit, the models with multiple progenitors implicitly assume the loss of lineage potential by restricting the types of mature cells that can be produced by each distinct progenitor. This may be a source of model misspecification. Indeed, recent studies dispute traditional assumptions about hematopoietic structures prescribing restricted differentiation pathways. For instance, Kawamoto et al. [2010] challenge the classical notion of a specialized myeloid progenitor, showing that lymphocyte progenitors (i.e. T, B, NK) can also give rise to myeloid cells (Gr and Mono). Recent *in vitro* studies of human hematopoiesis suggest multipotence of early progenitors [Notta et al., 2016] may only occur in mature systems, and argue that oligopotent behavior is only observed in early stages of development. Such oligopotent behavior in the specialization of progenitor cells is investigated in models (c)–(f), whose lack of fit to the data support these recent findings.

We emphasize that our method enables joint estimation of the initial barcoding distribution and intermediate process rates, including those relating to unobservable intermediate progenitor stages, given scale information via known mature cell death rates. The models we can consider are much more detailed and parameter-rich than those in previous statistical studies of hematopoiesis. We also note that while estimates are biologically plausible, they are obtained with an inevitable level of model misspecification, and rigorous approaches to model selection and to goodness of fit will be crucial to having more confidence in the validity of such model-based inference attempts.

5 Discussion

Our estimation procedure is the first method to our knowledge that enables parameter estimation in stochastic models including HSC, progenitor, and mature cell stages for time series data from hematopoietic lineage tracking experiments. Further, we show via simulation that the generalized loss function approach is very accurate when applied to data simulated from this class of models. Results from fitting experimental data have scientific bearing, newly estimating parameters such intermediate differentiation rates and initial marking levels in a multistage stochastic model. Our analysis confirms the major finding of Wu et al. [2014] — a distinct differentiation history of NK cells — from a newly statistical perspective using pairwise correlations. While we do not provide a rigorous approach to model selection in this paper, our exploration of several models suggests that non-restricted multipotent progenitor compartments provide a better fit to the data than models requiring an ordered differentiation through defined intermediaries. This supports recent findings from *in vitro* studies of human hematopoiesis [Notta et al., 2016] that challenge fundamental assumptions in the classical model of hematopoiesis.

We note that there are several limitations inherent to modeling hematopoiesis with a Markov branching process model. The assumptions of linearity and rate homogeneity imply a possibility of unlimited growth, and extending analysis to allow nonlinear effects such as feedback loops modulating the regulatory behavior as the system grows near a carrying capacity is merited. Similarly, the Markov assumption can be relaxed to include arbitrary lifespan distributions—age-dependent processes are one example falling under this model relaxation, and have been applied to analyzing stress erythropoiesis in recent studies [Hyrien et al., 2015]. Further phenomena such as immigration or emigration in a random environment may be considered in future studies: it is known that some cells we study in the peripheral bloodstream move in and out of tissue, for instance. While such extensions are mathematically difficult, they are trivial modifications to implement in simulation, and various forward simulation approaches or approximate methods such as approximate Bayesian computation (ABC) [Marjoram et al., 2003, Toni et al., 2009] may provide a promising alternative. Indeed, a Bayesian framework would allow existing prior information available from previous studies about average lifespans of mature blood cells to be incorporated without fixing these parameters.

The fully generative framework and accompanying method of inference additionally enable simulation studies and sensitivity analyses, and can be adapted to developing model selection tools. The larger scientific problem of inferring the most likely lineage pathway structure directly translates to the statistical problem of model selection. Many model selection approaches essentially build on parameter estimation techniques, balancing model complexity and goodness of fit by penalizing the number of model parameters via regularization. While model selection is generally difficult to perform in a loss function minimization framework, future work can investigate various penalization strategies applied to this class of models [Tibshirani, 1996, Fan and Li, 2001], or with shrinkage priors in a Bayesian setting [Park and Casella, 2008, Griffin and Brown, 2013]. Model selection using ABC is an active and rapidly developing area of research; see for instance [Toni et al., 2009, Liepe et al., 2014, Pudlo et al., 2016].

Modeling attempts using more parameter-rich models enabling more pathways or including additional cell fate events such as asymmetric division [Fong et al., 2009] should expect to be met with challenges related to overparametrization and identifiability, as well as added computational and mathematical complexity. However, such efforts and corresponding tools for model selection will be crucial in further progress toward understanding the structure of hematopoiesis. Finally, it should be noted that our results already support recent studies that challenge canonical multi-stage models of hematopoiesis [Kawamoto et al., 2010, Perié et al., 2014, Notta et al., 2016], and exploring a general class of models with quantitative model selection tools will lend a rigorous foundation to these insights, crucial toward a detailed understanding of the structure of hematopoiesis.

The class of models we consider and their available moment expressions are general in that an arbitrary number of intermediate progenitors and mature compartments can be specified, but have several limitations. First, we feature three stages of cell development in our model, and future work may extend this to include additional stages. Second, our modeling assumptions only allow for each mature cell to be descended from one progenitor compartment, which limits the ability to investigate fully connected and nested models. Nonetheless, we are now able to perform parameter estimation in a much more detailed model than previous statistical studies, while accounting for missing information and experimental noise. Such models commonly arise in related fields such as chemical kinetics, oncology, population ecology, and epidemiology, and our methodology contributes broadly to the statistical toolbox for inference in partially observed stochastic processes, a rich area of research that still faces significant challenges.

Acknowledgements

VNM was supported by the NIH grants R01-AI107034 and U54-GM111274. JX was supported by the NSF MSPRF. CED, SK and CW were supported by the intramural program of the National Heart, Lung and Blood Institute. The authors thank Jon Wellner and Jon Wakefield for helpful discussions and guidance about GMM and z-estimation.

References

JL Abkowitz, ML Linenberger, MA Newton, GH Shelton, RL Ott, and P Guttorp. Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. Proceedings of the National Academy of Sciences, 87(22):9062–9066, 1990.

NTJ Bailey. The Elements of Stochastic Processes; with Applications to the Natural Sciences. New York: Wiley, 1964.

AJ Becker, EA McCulloch, and JE Till. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature, 197:452–454, 1963.

L Biasco, D Pellin, S Scala, F Dionisio, L Basso-Ricci, L Leonardelli, S Scaramuzza, C Baricordi, F Ferrua, MP Cicalese, S Giannelli, V Neduva, DJ Dow, M Schmidt, C Von Kalle, MG 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, 19:107–119, 2016.

O Cappé, E Moulines, and T Rydén. Inference in hidden Markov models. Springer, New York, USA, 2006.

SN Catlin, JL Abkowitz, and P Guttorp. Statistical inference in a two-compartment model for hematopoiesis. Biometrics, 57(2):546–553, 2001.

SN Catlin, L Busque, RE Gale, P Guttorp, and JL Abkowitz. The replication rate of human hematopoietic stem cells in vivo. Blood, 117(17), 2011.

R Chen and O Hyrien. Quasi- and pseudo-maximum likelihood estimators for discretely observed continuous-time Markov branching processes. Journal of Statistical Planning and ISome priors for sparse regression modellingnference, 141(7):2209–2227, 2011.

C Colijn and MC Mackey. A mathematical model of hematopoiesis—I. Periodic chronic myelogenous leukemia. Journal of Theoretical Biology, 237(2):117–132, 2005.

J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360, 2001.

Y Fong, P Guttorp, and J Abkowitz. Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. The Annals of Applied Statistics, 3(4):1695–1709, 12 2009.

A Gerrits, B Dykstra, OJ Kalmykowa, K Klauke, E Verovskaya, MJC Broekhuis, G de Haan, and LV Bystrykh. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood, 115(13):2610–2618, 2010.

D Golinelli, P Guttorp, and JA Abkowitz. Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. Mathematical Medicine and Biology, 23(3):153–172, 2006.

S Goyal, S Kim, ISY Chen, and T Chou. Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biology, 13(1):85, 2015.

JE Griffin and PJ Brown. Some priors for sparse regression modelling. Bayesian Analysis, 8(3): 691–702, 2013.

P Guttorp. Stochastic Modeling of Scientific Data. CRC Press, 1995.

LP Hansen. Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, pages 1029–1054, 1982.

LP Hansen, J Heaton, and A Yaron. Finite-sample properties of some alternative GMM estimators. Journal of Business & Economic Statistics, 14(3):262–280, 1996.

Marc Hellerstein, MB Hanley, D Cesar, S Siler, C Papageorgopoulos, E Wieder, D Schmidt, R Hoh, R Neese, D Macallan, et al. Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. Nature Medicine, 5(1):83–89, 1999.

O Hyrien, SA Peslak, NM Yanev, and J Palis. Stochastic modeling of stress erythropoiesis using a two-type age-dependent branching process with immigration. Journal of Mathematical Biology, 70(7):1485–1521, 2015.

A Kaur, M Di Mascio, A Barabasz, M Rosenzweig, HM McClure, AS Perelson, RM Ribeiro, and RP Johnson. Dynamics of T-and B-lymphocyte turnover in a natural host of simian immunodeficiency virus. Journal of Virology, 82(3):1084–1093, 2008.

H Kawamoto, H Wada, and Y Katsura. A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. International Immunology, 22(2):65–70, 2010.

M Kimmel. Stochasticity and determinism in models of hematopoiesis. In A Systems Biology Approach to Blood, pages 119–152. Springer, 2014.

M Kimmel and DE Axelrod. Branching Processes in Biology. Springer, New York, 2002.

E Lakatos, A Ale, PDW Kirk, and MPH Stumpf. Multivariate moment closure techniques for stochastic kinetic models. The Journal of Chemical Physics, 143(9), 2015.

J Liepe, P Kirk, S Filippi, T Toni, CP Barnes, and MPH Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nature Protocols, 9(2):439–456, 2014.

R Lu, NF Neff, SR Quake, and IL Weissman. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature Biotechnology, 29(10):928–933, 2011.

A Marciniak-Czochra, T Stiehl, AD Ho, W Jäger, and W Wagner. Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. Stem Cells and Development, 18(3):377–386, 2009.

P Marjoram, J Molitor, V Plagnol, and S Tavaré. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, USA, 100(26):15324–15328, 2003.

MA Newton, P Guttorp, S Catlin, R Assunção, and JL Abkowitz. Stochastic modeling of early hematopoiesis. Journal of the American Statistical Association, 90(432):1146–1155, 1995.

F Notta, S Zandi, N Takayama, S Dobson, OI Gan, G Wilson, KB Kaufmann, J McLeod, E Laurenti, CF Dunant, JD McPherson, LD Stein, Y Dror, and JE Dick. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science, page aab2116, 2016.

M Ogawa. Differentiation and proliferation of hematopoietic stem cells. Blood, 81:2844–2844, 1993.

SH Orkin and LI Zon. Hematopoiesis: An evolving paradigm for stem cell biology. Cell, 132(4): 631–644, 2008.

A Pakes and D Pollard. Simulation and the asymptotics of optimization estimators. Econometrica: Journal of the Econometric Society, pages 1027–1057, 1989.

T Park and G Casella. The Bayesian lasso. Journal of the American Statistical Association, 103 (482):681–686, 2008.

L Perié, pd Hodgkin, SH Naik, TN Schumacher, RJ de Boer, and KR Duffy. Determining lineage pathways from cellular barcoding experiments. Cell Reports, 6(4):617 – 624, 2014.

P Pudlo, JM Marin, A Estoup, JM Cornuet, M Gautier, and CP Robert. Reliable ABC model choice via random forests. Bioinformatics, 32(6):859–866, 2016.

BE Shepherd, HP Kiem, PM Lansdorp, CE Dunbar, G Aubert, A LaRochelle, R Seggewiss, P Guttorp, and JL Abkowitz. Hematopoietic stem-cell behavior in nonhuman primates. Blood, 110 (6):1806–1813, 2007.

L Siminovitch, EA McCulloch, and JE Till. The distribution of colony-forming cells among spleen colonies. Journal of Cellular and Comparative Physiology, 62(3):327–336, 1963.

R Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, 58:267–288, 1996.

T Toni, D Welch, N Strelkowa, A Ipsen, and MPH Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface, 6(31):187–202, 2009.

AW Van der Vaart. Asymptotic Statistics, volume 3. Cambridge University Press, 2000.

J Wakefield. Bayesian and Frequentist Regression Methods. Springer-Verlag, New York, 2013.

IL Weissman. Stem cells: Units of development, units of regeneration, and units in evolution. Cell, 100(1):157–168, 2000.

ZL Whichard, CA Sarkar, M Kimmel, and SJ Corey. Hematopoiesis and its disorders: a systems biology approach. Blood, 115(12):2339–2347, 2010.

C Wu, B Li, R Lu, SJ. Koelle, Y Yang, A Jares, AE Krouse, M Metzger, F Liang, K Loré, CO Wu, RE. Donahue, ISY Chen, I Weissman, and CE Dunbar. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. Cell Stem Cell, 14(4): 486–499, 2014.

J Xu and VN Minin. Efficient transition probability computation for continuous-time branching processes via compressed sensing. Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pages 952–961, 2015.

J Xu, P Guttorp, MM Kato-Maeda, and VN Minin. Likelihood-based inference for discretely observed birth-death-shift processes, with applications to evolution of mobile genetic elements. Biometrics, 71(4):1009–1021, 2015.

Y Zhang, DL Wallace, CM De Lara, H Ghattas, B Asquith, A Worth, GE Griffin, GP Taylor, DF Tough, PCL Beverley, and Macallan DC. In vivo kinetics of human natural killer cells: the effects of ageing and acute and chronic viral infection. Immunology, 121(2):258–265, 2007.

Appendix

Consistency of loss function estimator

There are many variations on proofs of consistency and asymptotic normality for z-estimators or GMM estimators [Hansen, 1982, Pakes and Pollard, 1989, Van der Vaart, 2000]. Our consistency result is perhaps most similar to a version given by Theorem 3.1 in [Pakes and Pollard, 1989], showing that consistency holds generally for any vector $\hat{\boldsymbol{\theta}}_N$ that minimizes the norm $\|\mathbf{G}(\cdot)\|$ of a random, vector-valued function under the following conditions:

(i) $\|\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N)\| \leq o_p(1) + \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{G}_N(\boldsymbol{\theta})\|$,

(ii) $\mathbf{G}_N(\boldsymbol{\theta}_0) = o_p(1)$,

(iii) $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|\mathbf{G}_N(\boldsymbol{\theta})\|^{-1} = O_p(1) \qquad$ for each $\delta > 0$.

Heres $\boldsymbol{\theta}_0$ denotes the true data-generating parameters and is assumed to provide a global minimum to $\mathbf{G}$. A set of random variables $Z_n = o_p(1)$ if Z_n converges to zero in probability, while $Z_n = O_p(1)$ if the set is stochastically bounded, i.e. for any $\varepsilon > 0$, there exists finite M such that $\Pr(|Z_n| > M) < \varepsilon$ for all n.

The first condition restricts us to estimators $\hat{\boldsymbol{\theta}}_N$ that nearly minimize $\|\mathbf{G}_N(\cdot)\|$. Condition (ii) requires that under the true value, $\mathbf{G}_N(\boldsymbol{\theta}_0)$ converges to zero, which together with (i) implies that $\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N)$ must also approach zero. Finally, condition (iii) is an identifiability assumption, stating that small values of $\|\mathbf{G}_N(\boldsymbol{\theta})\|$ can only occur near $\boldsymbol{\theta}_0$; this now forces $\hat{\boldsymbol{\theta}}_N$ to approach $\boldsymbol{\theta}_0$. Note that our consistency result must also assume the identifiability condition (iii); we remark that we do not need to impose any smoothness assumptions, nor do we require that $\mathbf{G}(\boldsymbol{\theta}_0) = \mathbf{0}$. The formulation is repeated below:

Theorem 5.1 *Assume the observed process $\mathbf{Y}(t)$ has finite first and second moments. Let $\mathbf{G}_N(\boldsymbol{\theta}) = \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) - \hat{\boldsymbol{\psi}}(\mathbf{Y})$, where $\psi(\boldsymbol{\theta};\mathbf{Y})$ is a vector of correlations defined in Section 3.1 of the main text. We assume the true parameter $\boldsymbol{\theta}_0$ is identifiable, i.e. $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1} = O_p(1)$ for each $\delta > 0$. Then $\left\{\hat{\boldsymbol{\theta}}_N\right\}$ converges in probability to $\boldsymbol{\theta}_0$, where $\hat{\boldsymbol{\theta}}_N = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta};\mathbf{Y})$, $\mathcal{L}(\boldsymbol{\theta};\mathbf{Y}) = \|\mathbf{G}_N(\boldsymbol{\theta})\|_2^2$, and N is the number of independent processes or rows in $\mathbf{Y}$.*

Proof We begin by establishing almost sure convergence of $\mathbf{G}_N(\boldsymbol{\theta}) = \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) - \hat{\boldsymbol{\psi}}(\mathbf{Y})$ over the parameter domain $\boldsymbol{\Theta}$; here the dependence on N enters as the number of rows in the data matrix $\mathbf{Y}$. Because $\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}$ are finite-length vectors, it suffices to establish convergence component-wise for any indices j, m, n. Almost sure convergence of the empirical term $\hat{\boldsymbol{\psi}}(\mathbf{Y})$ is standard as its entries are sample Pearson correlation coefficients: we first equivalently write $\hat{\psi}_{mn}$ in the form

$$\hat{\psi}^j_{mn}(\mathbf{Y}) = \frac{\sum_{p=1}^N y^p_m(t_j) y^p_n(t_j) - \overline{y}_m(t_j)\overline{y}_n(t_j)}{\sqrt{\sum_{p=1}^N (y^p_m(t_j) - \overline{y}_m(t_j))^2}\sqrt{\sum_{p=1}^N (y^p_n(t_j) - \overline{y}_n(t_j))^2}}.$$

The Strong Law of Large Numbers ensures almost sure convergence of each term appearing above: $\sum_{p=1}^N y^p_m(t_j) y^p_n(t_j) \xrightarrow{\text{a.s.}} \mathrm{E}\left[Y_m(t_j)Y_n(t_j)\right]$, and for any n as well as m, $\overline{y}_m(t_j) \xrightarrow{\text{a.s.}} \mathrm{E}\left[Y_m(t_j)\right]$, $\sum_{p=1}^N (y^p_m(t_j) - \overline{y}_m(t_j))^2 \xrightarrow{\text{a.s.}} \mathrm{Var}\left[Y_m(t_j)\right]$. Applying the Continuous Mapping Theorem with the function $f(a_1, a_2, a_3, a_4, a_5) = \frac{a_1 - a_2 a_3}{\sqrt{a_4}\sqrt{a_5}}$ yields that $\hat{\psi}^j_{mn}(\mathbf{Y}) \xrightarrow{\text{a.s.}} \rho(Y_m(t_j)Y_n(t_j))$, where $\rho(Y_m(t_j)Y_n(t_j))$

denotes the true correlation between $Y_m(t_j), Y_n(t_j)$. Next, the remaining term $\boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y})$ consists of model-based correlations at each observation time t_j parametrized by $\boldsymbol{\theta}$ (both suppressed in the notation below), and recall from the main text that each correlation in its components is comprised of the expressions

$$\begin{aligned}\mathrm{Cov}(Y_m, Y_n) &= \frac{b_m b_n}{B_m B_n}\mathrm{Cov}(X_m, X_n),\\ \mathrm{Var}(Y_m) &= \frac{b_m(B_m - b_m)}{B_m(B_m-1)}\mathrm{E}(X_m) - \frac{b_m(B_m-b_m)}{B_m^2(B_m-1)}\mathrm{E}(X_m^2) + \frac{b_m^2}{B_m^2}\mathrm{Var}(X_m).\end{aligned}$$

While these expressions are formally independent of the number of processes N, the constant B_m should grow with N as it denotes the total number of barcoded type m cells; this can be understood as the multivariate hypergeometric sampling distribution limiting to a multinomial distribution. In this case, algebraic manipulations yield the pointwise limit

$$\psi_{mn} \to \frac{b_m b_n \mathrm{Cov}(X_m, X_n)}{\sqrt{b_m(b_m \mathrm{Var}(X_m) - \mathrm{E}(X_m^2))}\sqrt{b_n(b_n \mathrm{Var}(X_n) - \mathrm{E}(X_n^2))}}.$$

Thus, the random vector $\mathbf{G}_N$ as a whole converges almost surely to a deterministic limit function $\mathbf{G}$. Notice that implicitly if $\boldsymbol{\theta}_0$ are the true data-generating parameters under a correctly specified model, then

$$\hat{\psi}^j_{mn}(\mathbf{Y}) \xrightarrow{\text{a.s.}} \psi^j_{mn}(\boldsymbol{\theta_0}; \mathbf{Y}) = \rho(Y_m(t_j)Y_n(t_j)),$$

so that $\mathbf{G}(\boldsymbol{\theta}_0) = \mathbf{0}$. However, nowhere do we need the assumption that $\mathbf{G}(\boldsymbol{\theta}_0) = \mathbf{0}$, and so we will refer to the limiting value as $\mathbf{G}(\boldsymbol{\theta}_0)$ since the proof applies in non-idealized settings where the minimum of $\|\mathbf{G}(\boldsymbol{\theta})\|$, $\|\mathbf{G}(\boldsymbol{\theta}_0)\| > 0$.

Because $\mathbf{G}_N \xrightarrow{\text{a.s.}} \mathbf{G}$ on $\boldsymbol{\Theta}$, Egorov's theorem implies that $\mathbf{G}_N$ converges *uniformly* to $\mathbf{G}$ almost everywhere: that is, for any $\delta > 0$, there exists a set $S_\delta \in \boldsymbol{\Theta}$ such that $P(S_\delta) < \delta$, and $\sup\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus S_\delta} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})\| \xrightarrow{\text{P}} 0$. Therefore, since δ is arbitrarily small,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})\| \xrightarrow{\text{P}} 0.$$

Now, we are given an optimization procedure that produces estimators $\hat{\boldsymbol{\theta}}_N$ minimizing $\|\mathbf{G}_N\|$ by assumption, allowing us to write the first inequality in the series of relations

$$\|\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N)\| \le o_p(1) + \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{G}_N(\boldsymbol{\theta})\| \le o_p(1) + \|\mathbf{G}_N(\boldsymbol{\theta}_0)\| = o_p(1) + \|\mathbf{G}(\boldsymbol{\theta}_0)\|. \tag{A-1}$$

The second inequality holds as the infimum ranges over $\boldsymbol{\theta}_0$. Next, since $\mathbf{G}_N$ converges uniformly, it is certainly true that in particular $\mathbf{G}_N(\boldsymbol{\theta}_0) \xrightarrow{\text{P}} \mathbf{G}(\boldsymbol{\theta}_0)$, establishing the final equality. Subtracting the leftmost and rightmost sides of (A-1) from $\|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\|$ yields

$$\|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\| - \|\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N)\| \ge \|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\| - \|\mathbf{G}(\boldsymbol{\theta}_0)\| - o_p(1);$$

rearranging and again invoking uniform convergence of $\mathbf{G}_N$ reveals

$$\begin{aligned}\|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\| - \|\mathbf{G}(\boldsymbol{\theta}_0)\| &\le \|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\| - \|\mathbf{G}_N(\hat{\boldsymbol{\theta}}_N)\| + o_p(1)\\ &\le \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})\| + o_p(1) \xrightarrow{\text{P}} 0.\end{aligned} \tag{A-2}$$

Finally, the identifiability assumption allows us to find a number M such that for $\delta > 0$ and $\varepsilon > 0$,

$$\limsup \Pr \left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1} > M \right] < \varepsilon.$$

From (A-2) and since $\|\mathbf{G}(\boldsymbol{\theta}_0)\| \leq \|\mathbf{G}(\hat{\boldsymbol{\theta}}_N)\|$, we may write

$$\Pr \left[\|\mathbf{G}(\hat{\boldsymbol{\theta}}_N) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1} > M \right] \to 1,$$

and thus for large enough N we have with probability at least $1 - 2\varepsilon$,

$$\|\mathbf{G}(\hat{\boldsymbol{\theta}}_N) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1} > M \geq \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}_0)\|^{-1},$$

guaranteeing that $\hat{\boldsymbol{\theta}}_N$ is within δ of $\boldsymbol{\theta}_0$. Explicitly,

$$\limsup_{N\to\infty} \Pr \left[\|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\| > \delta \right] \leq 2\varepsilon.$$

Since δ, ε can be chosen to be arbitrarily small, the result follows that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{P}} \boldsymbol{\theta}_0$. ■

We remark that identifiability was assumed, rather than proved, because a formal assessment of identifiability is nontrivial due to the nonlinear moment expressions $\boldsymbol{\psi}(\cdot)$ arising in our estimator. We instead provide strong empirical evidence in simulation studies supporting identifiability .

Derivation of second moments

Here we explicitly derive the second moments of the simplest instance in our class of branching models of hematopoiesis. The derivation considers a four-type model with one progenitor and two mature types (Figure 2 (a) ignoring the third mature compartment). We also derive the marginalized moment expressions after incorporating the sampling distribution.

From applying the process rates to the Kolmogorov backward equations, we can write *pseudo-generating functions* defined

$$u_i(s_1, s_2, s_3, s_4) = \sum_j \sum_k \sum_l \sum_m a_i(j, k, l, m) s_1^j s_2^k s_3^l s_4^m. \tag{A-3}$$

For the model depicted in Figure 2 (b), these are given by

$$\begin{aligned} u_1(s_1, s_2) &= \lambda s_1^2 + \nu_0 s_2 - (\lambda + \nu_0) s_1, \\ u_2(s_2, s_3, s_4) &= \nu_1 s_2 s_3 + \nu_2 s_2 s_4 + \mu_0 - (\mu_0 + \nu_1 + \nu_2) s_2, \\ u_3(s_3) &= \mu_1 - \mu_1 s_3; \qquad u_4(s_4) = \mu_2 - \mu_2 s_4. \end{aligned}$$

Next, we can write the probability generating function (PGF) of the process, beginning with one type 1 particle, which is related to the pseudo-generating function u_1 as follows:

$$\begin{aligned}\phi_1(t; s_1, s_2, s_3, s_4) &= \mathrm{E}\left[\prod_{j=1}^{4} s_j^{X_j(t)} | \mathbf{X}(0) = (1,0,0,0)\right] \\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \Pr_{(1,0,0,0),(k,l,m,n)} s_1^k s_2^l s_3^m s_4^n \\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \left[\mathbf{1}_{\{k=1,l=m=n=0\}} + a_1(k,l,m,n)t + o(t)\right] s_1^k s_2^l s_3^m s_4^n \\ &= s_1 + u_1(s_1, s_2, s_3, s_4)t + o(t). \qquad \text{(A-4)}\end{aligned}$$

We may analogously define ϕ_i for processes beginning with one type i particle, for each $i = 1, \ldots, 4$. We have from Equation (A-4) the relation

$$\frac{\partial}{\partial t}\phi_i(t, s_1, \ldots, s_4) = u_i(\phi_1(t, s_1, \ldots, s_4), \ldots, \phi_4(t, s_1, \ldots, s_4)).$$

Now, let $M_{l|k}(t)$ denote the expected number of type l cells at time t, given one initial type k cell. From definition of ϕ_i, we see that we can relate the probability generating functions to these first moments via partial differentiation:

$$M_{l|k}(t) = \frac{\partial}{\partial s_l}\phi_k(t, s_1, \ldots, s_4)|_{s_1=s_2=s_3=s_4=1}.$$

Similarly, we may further differentiate the PGF to derive second moments used toward variance and covariance calculations. Define

$$U_{kl|1}(t) = \mathrm{E}\left[X_k(X_l - \mathbf{1}_{\{k=l\}})|\mathbf{X}(0) = (1,0,0,0)\right],$$

with $U_{kl|i}(t)$ defined analogously beginning with one type i particle. Then $U_{kl|j}(t) = \left.\frac{\partial^2 \phi_j}{\partial s_k \partial s_l}\right|_{\mathbf{s}=1}$, and by the Faà di Bruno formula,

$$\frac{\partial^3 \phi_i}{\partial t \partial s_j \partial s_k} = \sum_{m=1}^{4}\left(\frac{\partial u_i}{\partial \phi_m}\frac{\partial^2 \phi_m}{\partial s_j \partial s_k}\right) + \sum_{m,n=1}^{4}\left(\frac{\partial^2 u_i}{\partial \phi_m \partial \phi_n}\frac{\partial \phi_m}{\partial s_j}\frac{\partial \phi_k}{\partial s_k}\right).$$

This relation allows us to write a system of non-homogeneous, linear ordinary differential equations

(ODEs) governing second order moments:

$$\begin{aligned}
\frac{\partial}{\partial t}U_{33|1}(t) &= (\lambda-\nu_0)U_{33|1}(t)+\nu_0 U_{33|2}(t)+(2\lambda)M_{3|1}^2(t),\\
\frac{\partial}{\partial t}U_{44|1}(t) &= (\lambda-\nu_0)U_{44|1}(t)+\nu_0 U_{44|2}(t)+(2\lambda)M_{4|1}^2(t),\\
\frac{\partial}{\partial t}U_{34|1}(t) &= (\lambda-\nu_0)U_{34|1}(t)+\nu_0 U_{34|2}(t)+(2\lambda)M_{3|1}(t)M_{4|1}(t),\\
\frac{\partial}{\partial t}U_{34|2}(t) &= -\mu_0 U_{34|2}(t)+\nu_1 M_{4|2}(t)M_{3|3}(t)+\nu_2 M_{3|2}(t)M_{4|4}(t),\\
\frac{\partial}{\partial t}U_{33|2}(t) &= -\mu_0 U_{33|2}(t)+\nu_1 U_{33|3}(t)+2\nu_1 M_{3|2}(t)M_{3|3}(t),\\
\frac{\partial}{\partial t}U_{44|2}(t) &= -\mu_0 U_{44|2}(t)+\nu_2 U_{44|4}(t)+2\nu_2 M_{4|2}(t)M_{4|4}(t),\\
\frac{\partial}{\partial t}U_{33|3}(t) &= -\mu_1 U_{33|3}(t),\\
\frac{\partial}{\partial t}U_{44|4}(t) &= -\mu_2 U_{44|4}(t),
\end{aligned}$$

all with initial conditions $(\cdot)_{k,l}(0)=0$. We immediately see that $U_{33|3}(t)=U_{44|4}(t)=0$, and upon a series of solutions and substitutions, we successively solve the system of ODEs, yielding the following explicit solutions:

$$\begin{aligned}
U_{33|2}(t) &= 2\frac{\nu_1^2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-2\mu_2 t}}{\mu_0-2\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-2\mu_2)}\right],\\
U_{44|2}(t) &= 2\frac{\nu_2^2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-2\mu_2 t}}{\mu_0-2\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-2\mu_2)}\right],\\
U_{34|2}(t) &= \frac{\nu_1\nu_2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_1)t}}{\mu_1}-\frac{e^{-(\mu_1+\mu_2)t}}{\mu_0-\mu_1-\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_1(\mu_0-\mu_1-\mu_2)}\right]\\
&\quad+\frac{\nu_1\nu_2}{(\mu_1-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-(\mu_1+\mu_2)t}}{\mu_0-\mu_1-\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-\mu_1-\mu_2)}\right],\\
U_{33|1}(t) &= e^{(\lambda-\nu_0)t}\Bigg\{2\frac{\nu_0\nu_1^2}{\mu_1-\mu_0}\Bigg[\frac{(\mu_0-\mu_1)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_1(\mu_0-2\mu_1)(\nu_0-\lambda-\mu_0)}-\frac{e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{\mu_1(\nu_0-\lambda-\mu_0-\mu_1)}-\frac{e^{(\nu_0-\lambda-2\mu_1)t}}{(\mu_0-2\mu_1)(\nu_0-\lambda-2\mu_1)}\\
&\qquad+\frac{\mu_1-\mu_0}{\mu_1(\mu_0-2\mu_1)(\nu_0-\lambda-\mu_0)}+\frac{1}{\mu_1(\nu_0-\lambda-\mu_0-\mu_1)}+\frac{1}{(\mu_0-2\mu_1)(\nu_0-\lambda-2\mu_1)}\Bigg]\\
&\quad+\frac{2\lambda\nu_0^2\nu_1^2}{(\mu_1-\mu_0)^2}\Bigg[\frac{e^{(\nu_0-\lambda-2\mu_0)t)}}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)}-\frac{2e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)}\\
&\qquad+\frac{2(\mu_0-\mu_1)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0)^2}+\frac{e^{(\nu_0-\lambda-2\mu_1)t}}{(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-2\mu_1)}+\frac{2(\mu_1-\mu_0)e^{-\mu_1 t}}{\mu_1(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)}\\
&\qquad+\frac{(\mu_1-\mu_0)^2e^{(\lambda-\nu_0)t}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)^2}-\frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)}\\
&\qquad+\frac{2}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)}-\frac{2(\mu_0-\mu_1)}{\mu_0(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0)^2}\\
&\qquad-\frac{1}{(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-2\mu_1)}-\frac{2(\mu_1-\mu_0)}{\mu_1(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)}\\
&\qquad-\frac{(\mu_1-\mu_0)^2}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)^2)}\Bigg]\Bigg\},
\end{aligned}$$

$$
\begin{aligned}
U_{44|1}(t) = e^{(\lambda-\nu_0)t}\Bigg\{ & 2\frac{\nu_0\nu_2^2}{\mu_2-\mu_0}\Bigg[\frac{(\mu_0-\mu_2)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_2(\mu_0-2\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{\mu_2(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{e^{(\nu_0-\lambda-2\mu_2)t}}{(\mu_0-2\mu_2)(\nu_0-\lambda-2\mu_2)} \\
& \quad + \frac{\mu_2-\mu_0}{\mu_2(\mu_0-2\mu_2)(\nu_0-\lambda-\mu_0)} + \frac{1}{\mu_2(\nu_0-\lambda-\mu_0-\mu_2)} + \frac{1}{(\mu_0-2\mu_2)(\nu_0-\lambda-2\mu_2)}\Bigg] \\
& + \frac{2\lambda\nu_0^2\nu_2^2}{(\mu_2-\mu_0)^2}\Bigg[\frac{e^{(\nu_0-\lambda-2\mu_0)t)}}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} - \frac{2e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} \\
& + \frac{2(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)^2} + \frac{e^{(\nu_0-\lambda-2\mu_2)t}}{(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-2\mu_2)} + \frac{2(\mu_2-\mu_0)e^{-\mu_2 t}}{\mu_2(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)} \\
& + \frac{(\mu_2-\mu_0)^2e^{(\lambda-\nu_0)t}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)^2} - \frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} \\
& + \frac{2}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{2(\mu_0-\mu_2)}{\mu_0(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)^2} \\
& - \frac{1}{(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-2\mu_2)} - \frac{2(\mu_2-\mu_0)}{\mu_2(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)} \\
& - \frac{(\mu_2-\mu_0)^2}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)^2)}\Bigg]\Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
U_{34|1}(t) = e^{(\lambda-\nu_0)t}\Bigg\{ & \frac{\nu_0\nu_1\nu_2}{\mu_2-\mu_0}\cdot\Bigg[\frac{(\mu_0-\mu_2)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_1(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_0)t}}{\mu_1(\nu_0-\lambda-\mu_1-\mu_0)} \\
& - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{\mu_2-\mu_0}{\mu_1(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{1}{\mu_1(\nu_0-\lambda-\mu_1-\mu_0)} + \frac{1}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)}\Bigg] \\
& + \frac{\nu_0\nu_1\nu_2}{\mu_1-\mu_0}\Bigg[\frac{(\mu_0-\mu_1)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_2(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_2-\mu_0)t}}{\mu_2(\nu_0-\lambda-\mu_2-\mu_0)} \\
& - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{\mu_1-\mu_0}{\mu_2(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{1}{\mu_2(\nu_0-\lambda-\mu_2-\mu_0)} + \frac{1}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)}\Bigg] \\
+ & \frac{2\lambda\nu_0^2\nu_1\nu_2}{(\mu_1-\mu_0)(\mu_2-\mu_0)}\cdot\Bigg[\frac{e^{(\nu_0-\lambda-2\mu_0)t}}{(\nu_0-\lambda-2\mu_0)(\nu_0-\lambda-\mu_0)^2} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} \\
& + \frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_2)} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)} \\
& + \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{(\mu_2-\mu_0)e^{-\mu_1 t}}{\mu_1(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} \\
+ & \frac{(\mu_0-\mu_1)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)} + \frac{(\mu_1-\mu_0)e^{-\mu_2 t}}{\mu_2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{(\mu_1-\mu_0)(\mu_2-\mu_0)e^{(\lambda-\nu_0)t)}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)} - \frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} \\
+ & \frac{1}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{\mu_0-\mu_2}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_2)} \\
+ & \frac{1}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)} - \frac{1}{(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} \\
+ & \frac{\mu_0-\mu_2}{\mu_1(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} + \frac{\mu_1-\mu_0}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)} \\
+ & \frac{\mu_0-\mu_1}{\mu_2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{(\mu_1-\mu_0)(\mu_2-\mu_0)}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)}\Bigg]\Bigg\}.
\end{aligned}
$$

Marginalized variance and covariance derivation

Because the initial state is uncertain, the variances and covariances of X_3, X_4 can now be computed by marginalizing over the initial barcoding state. The marginalized means follow trivially by linearity and the law of total expectation: for instance,

$$\mathrm{E}[X_3(t)] = \pi \mathrm{E}\left[X_3(t)|\mathbf{X}(0) = (1,0,0,0)\right] + (1-\pi)\mathrm{E}\left[X_3(t)|\mathbf{X}(0) = (0,1,0,0)\right] = \pi M_{3|1} + (1-\pi)M_{3|2}.$$

Dropping the dependence on t for notational simplicity, we use the law of total variance and law of total covariance to obtain the marginalized variance expressions

$$\begin{aligned}
\mathrm{Cov}(X_3, X_4) &= \pi^2(U_{34} - M_{3|1}M_{4|1}) + (1-\pi)^2(V_{34} - M_{3|2}M_{4|2}) \\
&\quad + \pi(1-\pi)(U_{34|1} + U_{34|2} - M_{3|2}M_{4|1} - M_{3|1}M_{4|2}) \\
\mathrm{Var}(X_3) &= \pi(U_{33|1} + M_{3|1}) + (1-\pi)(U_{33|2} + M_{3|2}) \\
&\quad - \pi^2 M_{3|1}^2 - (1-\pi)^2 M_{3|2}^2 - 2\pi(1-\pi)M_{3|1}M_{3|2} \\
\mathrm{Var}(X_4) &= \pi(U_{44|1} + M_{4|1}) + (1-\pi)(U_{44|2} + M_{4|2}) \\
&\quad - \pi^2 M_{4|1}^2 - (1-\pi)^2 M_{4|2}^2 - 2\pi(1-\pi)M_{4|1}M_{4|2}.
\end{aligned} \tag{A-5}$$

We now include the details behind Equation (A-5) and derive the expressions in the general case with K progenitors. Applying the law of iterated variance, the total variance for a type i mature cell population is given by

$$\mathrm{Var}[X_i(t)] = \underbrace{\mathrm{E}[\mathrm{Var}[X_i(t)|\mathbf{X}(0)]]}_{(1)} + \underbrace{\mathrm{Var}[\mathrm{E}[X_i(t)|\mathbf{X}(0)]]}_{(2)}.$$

We drop the dependence on t in intermediate steps for simplicity, and adopt the notation

$$\mathrm{E}(X_{i|1}^2) = \mathrm{E}\left[X_i^2|\mathbf{X}(0) = (1,0,0,\ldots,0)\right],$$

and similarly use $\mathrm{E}[X_{i|j}]$ for expectations of $X_i(t)$ conditional on beginning with one initial type j particle at $t = 0$. With these conventions, the outer expectation over initial barcoding probability (1) simplifies to

$$\begin{aligned}
\mathrm{E}[\mathrm{Var}[X_i|\mathbf{X}(0)]] &= \mathrm{E}\left\{\mathrm{E}[X_i^2|\mathbf{X}(0)] - [\mathrm{E}[X_i|\mathbf{X}(0)]]^2\right\} \\
&= \pi_1\mathrm{E}(X_{i|1}^2) + \ldots + \pi_K\mathrm{E}(X_{i|K}^2) - \pi_1\left[\mathrm{E}(X_{i|1})\right]^2 + \ldots + \pi_K\left[\mathrm{E}(X_{i|K})\right]^2 \\
&= \sum_{k=1}^{K}\pi_k\mathrm{E}\left(X_{i|k}^2\right) - \sum_{k=1}^{K}\pi_k\left[\mathrm{E}\left(X_{i|k}\right)\right]^2.
\end{aligned}$$

Next, it is straightforward to expand (2) as

$$\begin{aligned}
\mathrm{Var}[\mathrm{E}[X_i|\mathbf{X}(0)]] &= \mathrm{E}\{\mathrm{E}[X_i|\mathbf{X}(0)]\}^2 - (\mathrm{E}\{\mathrm{E}[X_i|\mathbf{X}(0)]\})^2 \\
&= \sum_{k=1}^{K}\pi_k\left[\mathrm{E}\left(X_{i|k}\right)\right]^2 - \left[\sum_{k=1}^{K}\pi_k\mathrm{E}(X_{i|k})\right]^2.
\end{aligned}$$

Combining these simplifications (1) + (2), we arrive at the total variance expression marginalized over initial state:

$$\mathrm{Var}[X_i(t)] = \sum_{k=1}^{K}\pi_k\mathrm{E}[X_{i|k}^2] - \sum_{k=1}^{K}\pi_k^2\mathrm{E}\left[\left(X_{i|k}\right)\right]^2 - 2\sum_{j>k}\pi_j\pi_k\mathrm{E}[X_{i|j}]\mathrm{E}[X_{i|k}]. \tag{A-6}$$

Analogously to (A-5) for the four-type model, this expression is directly related to the closed form solutions we obtain from solving the systems of moment differential equations. In terms of moment expressions, (A-6) becomes

$$\mathrm{Var}[X_i(t)] = \sum_{k=1}^{K} \pi_k [U_{ii|k}(t) + M_{i|k}(t)] - \sum_{k=1}^{K} \pi_k^2 M_{i|k}(t)^2 - 2\sum_{j>k} \pi_j \pi_k M_{i|k}(t) M_{i|j}(t).$$

The marginal covariance expressions are then obtained exactly analogously, applying the law of total covariance instead of the law of total variance. The covariances are given by

$$\begin{aligned}\mathrm{Cov}[X_i(t), X_j(t)] &= \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{i|k} X_{j|k}] - \sum_{k=1}^{K} \pi_k^2 \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|k}] - \sum_{k\neq l} \pi_k \pi_l \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|l}] \\ &= \sum_{k=1}^{K} \pi_k U_{ij|k}(t) - \sum_{k=1}^{K} \pi_k^2 M_{i|k}(t) M_{j|k}(t) - \sum_{k\neq l} \pi_k \pi_l M_{i|k}(t) M_{j|l}(t).\end{aligned}$$

Given these marginalized variance and covariance expressions, incorporating the hypergeometric sampling distribution to obtain covariance and variance between read data $\mathbf{Y}$ applies identically by the equations in the main paper.

Unconstrained parametrization of initial barcoding vector:

For models with multiple progenitor types, the initial barcoding probabilities must be represented as a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ where π_1 denotes the probability of starting as an HSC, and π_i denotes the probability of starting as a type i progenitor for $i = 2, \ldots, K$. These parameters π_i are naturally constrained to a probability simplex, but in practice we reparameterize by borrowing from a technique used in multinomial logistic regression by defining a set of variables $\gamma_i := \ln(\pi_i/\pi_K)$ for $i = 1, \ldots, K-1$. Then notice $\pi_i = \pi_K e^{\gamma_i}$ for all $i \leq K-1$, and letting $\pi_K = \frac{1}{1+\sum_{i=1}^{K-1} e^{\gamma_i}}$, we ensure the simplex constraint that $\sum_{i=1}^{K} \pi_i = 1$. This enables us to equivalently consider the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K-1})$ as parameters instead of $\boldsymbol{\pi}$, and because γ_i vary freely in $\mathbb{R}$, we no longer need to add a constraint to the optimization problem.

Detailed simulation results

Here, we include detailed tables of true parameters used to initiate simulation as well as median estimates, median absolute deviations, and standard deviations corresponding to the simulation study design discussed in section 4.1 for all model structures depicted in Figure 2. Some models depicted in Figure 2 are identical in simulation study — for instance, models (c) and (d) have no difference when final types are arbitrary. We also note that estimates reported in Table C-3 correspond to the results plotted in Figure 3 in the main text.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	π_a
True	0.0280	0.0200	0.0080	36	15	7	0.9000
Median	0.0283	0.0194	0.0086	34.84	14.18	6.624	0.8959
MAD	0.0008	0.0009	0.0021	6.31	2.797	1.167	0.0201
SD	0.0008	0.0010	0.0021	10.33	4.623	1.993	0.0199

Table C-1: Results of estimation on synthetic data from a model with three mature types and one common progenitor compartment, i.e. Model (a) in Figure 2 of the main text, in terms of medians, standard deviations (SD), and median absolute deviations (MAD). With fixed death rates at $\mu_1 = 0.24, \mu_2 = 0.14, \mu_3 = 0.09$, estimates are very close to true parameters used to simulate the data. Recall π_a denotes the proportion barcoded as progenitors, while $\pi_1 = 1 - \pi_a$ is the proportion marked at the HSC stage.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	ν_4	ν_5	π_a
True	0.0285	0.0200	0.0080	36.00	15.00	10.00	20.00	7.000	0.9000
Median	0.0284	0.0200	0.0076	37.16	15.54	10.35	20.69	7.246	0.9021
MAD	0.0007	0.0011	0.0016	5.851	2.568	1.693	3.399	1.178	0.0153
SD	0.0025	0.0019	0.2800	11.84	3.568	2.504	4.574	1.994	0.0465

Table C-2: Model with five mature types and one common progenitor compartment, i.e. Model (b) in Figure 2. Death rates fixed at $\mu_1 = 0.26, \mu_2 = 0.13, \mu_3 = 0.11, \mu_4 = 0.16, \mu_5 = 0.09$.

	λ	ν_a	ν_b	μ_a	μ_b	ν_1	ν_2	ν_3	ν_4	ν_5	π_a	π_b
True	0.0285	0.0130	0.0070	0.0050	0.0040	36	15	10	20	7	0.60	0.30
Med.	0.0286	0.0130	0.0069	0.0045	0.0043	38.01	16.29	10.92	19.64	6.65	0.6333	0.2706
MAD	0.0005	0.0008	0.0006	0.0021	0.0013	13.35	5.826	3.894	2.240	1.277	0.1399	0.1194
SD	0.0006	0.0007	0.0007	0.0019	0.0012	17.61	7.828	5.241	5.347	1.925	0.1388	0.1255

Table C-3: Model with five mature types and two distinct progenitor compartments, i.e. Model (c) in Figure 2. In this model, progenitor a gives rise to type 1 and 2 mature cells, and b produces type $3, 4$, and 5 type cells. Estimates remain accurate in this parameter rich setting with multiple progenitor compartments. These correspond to estimates plotted in Figure 3 in the main text.

Model misspecification experiments

Tables C-5 and C-6 display the estimates obtained under over specified and misspecified models, along with objective values of the loss function at converged estimates; these correspond to total ℓ_2 loss between fitted and observed correlations. Note that these estimates correspond to the correlation plots displayed in Figure 5 in the main text.

	λ	ν_a	ν_b	ν_c	μ_a	μ_b	μ_c	ν_1
True	0.0500	0.0280	0.0140	0.0070	0.0080	0.0060	0.0020	40.0000
Median	0.0539	0.0303	0.0150	0.0075	0.0091	0.0058	0.0034	40.7977
MAD	0.0081	0.0047	0.0032	0.0016	0.0038	0.0072	0.0041	11.8020
SD	0.0143	0.0080	0.0052	0.0024	0.0037	0.0060	0.0053	18.0492
	ν_2	ν_3	ν_4	ν_5	π_a	π_b	π_c	
True	18.0000	14.0000	20.0000	8.0000	0.5500	0.2000	0.1500	
Median	18.1527	17.7127	26.4716	10.6550	0.5595	0.2017	0.1578	
MAD	5.0599	7.8044	9.4919	5.7547	0.0412	0.0120	0.0106	
SD	7.0998	6.6657	8.8583	157.5674	0.0369	0.0159	0.0137	

Table C-4: Synthetic data from a model with five mature types and three oligopotent and unipotent progenitors, i.e. Model (f) in Figure 2. Death rates fixed at $\boldsymbol{\mu} = (0.24, 0.13, 0.12, 0.18, 0.1)$. While the standard deviation reveals influence of extreme outliers on the estimate or ν_4, median estimates are again accurate in a parameter rich model, and reasonably stable in terms of MAD.

	λ	ν_a	ν_b	ν_c	μ_a	μ_b	μ_c	ν_1
Med.	0.19365	0.05938	0.05475	0.00002	0.01136	0.19085	0.00056	56.89686
MAD	0.06633	0.02942	0.02433	0.00003	0.01683	0.28294	0.00083	38.64162
SD	0.07195	0.03583	0.03360	0.00015	0.59207	0.95299	0.00226	238.53322
	ν_2	ν_3	ν_4	ν_5	π_a	π_b	π_c	Objective
Med.	22.10742	16.70475	33.70443	11.65951	0.00121	0.00038	0.83830	2.90319
MAD	14.61860	9.40493	19.63489	5.62197	0.00179	0.00057	0.07087	0.75504
SD	11.94989	7.67183	15.40404	5.04462	0.01796	0.00948	0.08250	1.08521

Table C-5: Model fit in overspecified case with three progenitors: note that the objective value is higher than the correct specification, and note that the estimates seem more spread apart than the correctly specified inference while representative of the overall shape of true correlation profiles.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	ν_4	ν_5	π_a	Objective
Med.	0.131	0.00468	0.0332	71.1	30.0	20.6	0.000	0.000	1.000	21.358
MAD	0.0091	0.0041	0.0123	29.9	12.7	7.87	0.00000	0.00000	0.00001	0.221
SD	0.0096	0.0078	0.0149	23.7	9.91	6.43	0.00000	0.00000	0.00001	0.227

Table C-6: Underspecified model fit. Interestingly, this model seems to correctly identify that types $1, 2, 3$ are linked from a common progenitor, but because one shared progenitor is not compatible with the observed correlations, and in particular cannot explain negative correlations between types from distinct lineages, the model assigns almost zero mass to rates ν_4, ν_5 of producing the other mature types. The solution seems to be strongly a boundary solution with all barcoded cells starting in the progenitor compartment, resulting in a very poor objective function value.

	λ	ν_a	ν_b	μ_a	μ_b	ν_1	ν_2	ν_3	π_2	π_3	Obj.
Med.	0.0286	0.0130	0.0080	0.0077	0.0014	31.43	21.32	46.57	0.533	0.358	9.093×10^{-5}
MAD	0.0007	0.0009	0.0011	0.0022	0.0020	6.595	5.097	29.47	0.1096	0.1009	4.629×10^{-5}
SD	0.0067	0.0043	0.0028	0.0026	0.0022	22.31	21.54	63.07	0.1791	0.1943	6.950×10^{-5}

Table C-7: Results corresponding to three grouped mature cell compartments with correctly specified progenitor structure. Note the objective value here is orders of magnitude lower than the five-type models with misspecified progenitor structures, suggesting that lumping mature types is a justifiable model simplification compared to the tradeoff of specifying a richer model with flawed assumptions on the intermediate structure.

Tables of complete estimated parameters fitted to lineage barcoding data

Par	(a)	(b)	(c)	(d)	(e)	(f)
$\hat{\lambda}$	0.0593	0.0867	0.4360	0.3644	0.2271	0.3198
$\hat{\nu}_a$	1.00e-6	1.80e-7	0.4090	0.3521	0.1446	0.0033
$\hat{\nu}_b$			0.0257	0.0121	0.0725	0.3131
$\hat{\nu}_c$					0.0101	0.0033
$\hat{\mu}_a$	7.95e-6	0.0367	1.150	4.096	0.7037	0.1449
$\hat{\mu}_b$			4.023	3.699	4.022	1.253e-3
$\hat{\mu}_c$					3.602	1.434
$\hat{\nu}_1$	2042.0	1486.3	1305.5	866.1	1896.8	1959.09
$\hat{\nu}_2$	434.7	1764.3	201.4	391.3	221.3	560.4
$\hat{\nu}_3$	147.4	74.0	113.6	264.4	112.3	127.5
$\hat{\nu}_4$		326.4	448.7	299.5	417.1	287.9
$\hat{\nu}_5$		17.9	17.0	54.1	79.3	104.2
$\hat{\pi}_a$	0.861	0.87*	0.870	0.870	0.0	0.0
$\hat{\pi}_b$			0.0	0.0	0.0	0.870
$\hat{\pi}_c$					0.870	0.0
Loss	0.4071	1.653	3.465	3.330	3.836	2.91

Table C-8: Parameter estimates for all models displayed in Figure 2. Model (a) has fixed deaths $(0.6, 0.04, 0.4)$. All other models have fixed death rates $(0.8, 0.3, 0.04, 0.08, 0.4)$.

Par	(a)	(b)	(c)	(d)	(e)	(f)
$\hat{\lambda}$	$(0.003, 0.109)$	$(0.077, 0.163)$	$(0.196, 1.168)$	$(0.207, 0.640)$	$(0.132, 0.752)$	$(0.174, 0.449)$
$\hat{\nu}_a$	$(0.0, 0.004)$	$(0.0, 0.001)$	$(0.085, 1.148)$	$(0.131, 0.611)$	$(0.014, 0.617)$	$(0.074, 0.396)$
$\hat{\nu}_b$			$(0.006, 0.168)$	$(0.006, 0.112)$	$(0.015, 0.486)$	$(0.021, 0.154)$
$\hat{\nu}_c$					$(0.000, 0.019)$	$(0.000, 0.010)$
$\hat{\mu}_a$	$(0.0, 0.002)$	$(0.028, 0.046)$	$(0.000, 3.879)$	$(0.267, 3.603)$	$(0.0, 2.854)$	$(0.246, 2.651)$
$\hat{\mu}_b$			$(0.434, 4.022)$	$(0.437, 4.102)$	$(0.447, 4.023)$	$(0.811, 4.543)$
$\hat{\mu}_c$					$(0.102, 4.103)$	$(0.283, 4.100)$
$\hat{\nu}_1$	$(956.0, 2239.9)$	$(830.1, 1838.6)$	$(627.7, 1482.2)$	$(613.3, 1487.3)$	$(600.9, 1495.0)$	$(615.7, 1474.2)$
$\hat{\nu}_2$	$(52.0, 488.7)$	$(1021.9, 2055.3)$	$(131.8, 521.4)$	$(135.6, 344.6)$	$(187.9, 477.2)$	$(255.9, 448.7)$
$\hat{\nu}_3$	$(39.7, 148.2)$	$(60.7, 99.8)$	$(30.6, 294.6)$	$(134.5, 305.0)$	$(4.279, 291.3)$	$(126.3, 297.8)$
$\hat{\nu}_4$		$(275.2, 470.7)$	$(146.2, 558.7)$	$(126.4, 297.4)$	$(127.8, 321.7)$	$(128.2, 295.7)$
$\hat{\nu}_5$		$(10.1, 44.65)$	$(3.786, 9.559)$	$(1.137, 10.63)$	$(6.488, 84.9)$	$(26.4, 74.0)$
$\hat{\pi}_a$	$(0.017, 0.861)$	$(0.87, 0.87)$	$(0.0, .0.599)$	$(0.0, .598)$	$(0.0, 0.038)$	$(0.000, 0.001)$
$\hat{\pi}_b$			$(0.0, 0.999)$	$(0.0, 0.999)$	$(0.0, 1.0)$	$(0.999, 1.0)$
$\hat{\pi}_c$					$(0.0, 1.0)$	$(0.000, 0.000)$
Loss	$(0.352, 0.696)$	$(1.485, 2.341)$	$(2.591, 4.771)$	$(2.472, 4.489)$	$(3.092, 5.763)$	$(2.566, 4.777)$

Table C-9: Corresponding 95% confidence intervals produced via nonparametric bootstrap of 2500 replicate datasets. Recall sum of progenitor barcoding proportions fixed to be 0.87 for models (b)-(f).

Additional fitted correlation profiles fitted to lineage barcoding dataset

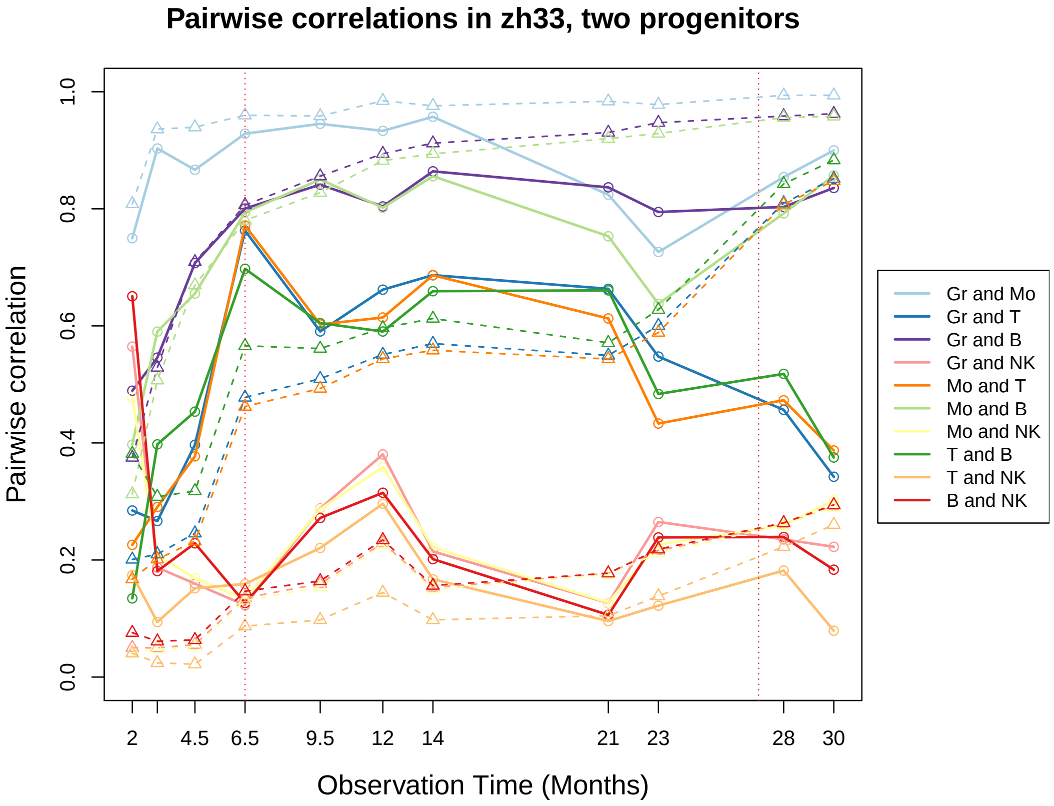


Figure C-9: Fitted curves for real data to model with two progenitors, corresponding to model (c) displayed in Figure 2. The "misgrouped" fitted curves apparent after 23 months visually suggest the misspecification in designating specialized oligopotent progenitors.

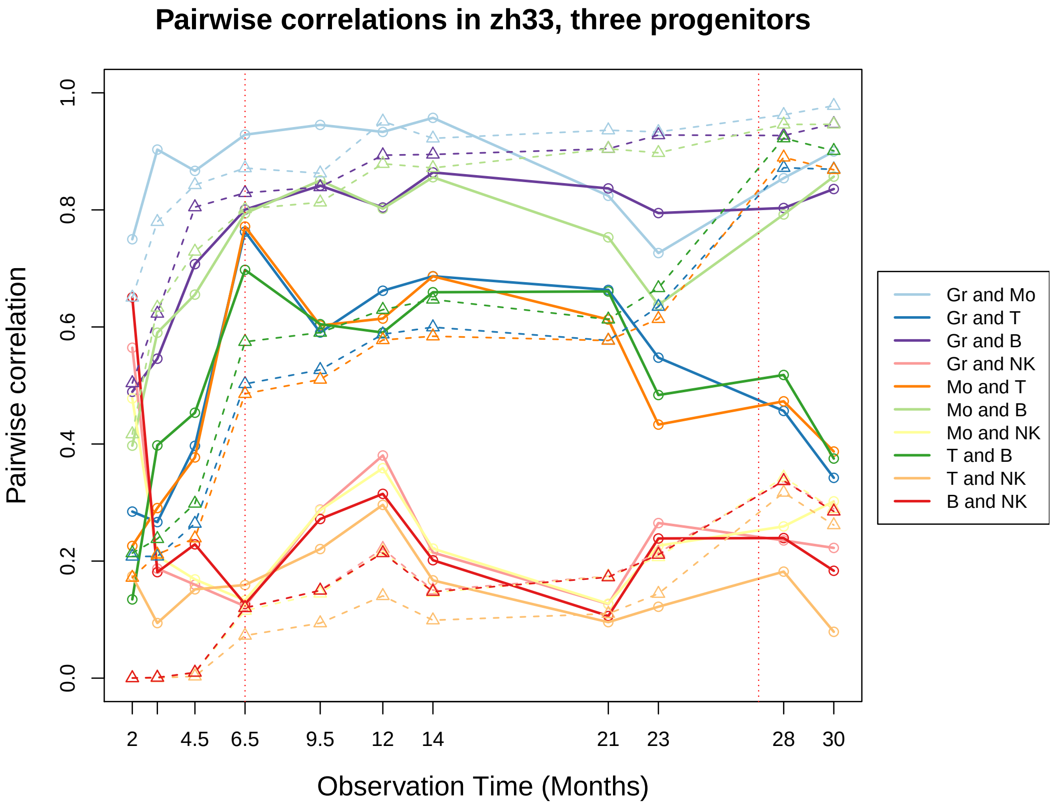


Figure C-10: Fitted curves for real data in model with three specialized progenitors, i.e. model (e) in Figure 2. Again, a misgrouping is visually apparent in fitted curves after 23 months