The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2091–2119
https://doi.org/10.1214/19-AOAS1272

STATISTICAL INFERENCE FOR PARTIALLY OBSERVED BRANCHING PROCESSES WITH APPLICATION TO CELL LINEAGE TRACKING OF *IN VIVO* HEMATOPOIESIS

BY JASON XU[1,*], SAMSON KOELLE[3,†], PETER GUTTORP[†], CHUANFENG WU[3,‡], CYNTHIA DUNBAR[3,‡], JANIS L. ABKOWITZ[†] AND VLADIMIR N. MININ[2,§]

*Duke University**, University of Washington*[†], National Institutes of Health*[‡] and University of California, Irvine*[§]

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 describing cell division and differentiation. In this paper, we develop such a framework with corresponding moment-based parameter estimation techniques for continuous-time, multi-type branching processes. Such processes provide a probabilistic model of how cells divide and differentiate, and we apply our method to study *hematopoiesis*, the mechanism of blood cell production. 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 much richer statistical models of hematopoiesis than those 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 validating 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 nonhuman primates, which may be more relevant to human biology and clinical strategies than previous findings from murine studies. For example, in addition to previously estimated hematopoietic stem cell self-renewal rate, we are able to estimate fate decision probabilities and to compare structurally distinct models of hematopoiesis using cross validation. These estimates of fate decision probabilities and our model selection results should help biologists compare competing hypotheses about how progenitor cells differentiate. The methodology is transferrable to a large class of stochastic compartmental and multi-type branching models, commonly used in studies of cancer progression, epidemiology and many other fields.

Received May 2018; revised March 2019.

[1]Supported in part by NSF Grant MSPRF #1606177.

[2]Supported by NIH Grants R01-AI107034 and U54-GM111274.

[3]Supported by National Heart, Lung, and Blood Institute intramural program.

Key words and phrases. Stochastic compartmental models, generalized method of moments, Markov jump processes, mechanistic modeling, stem cells.

1. Introduction. This paper develops inferential tools for a class of hidden stochastic population processes. In particular, we present a correlation-based M-estimator for rate inference in multi-type branching process models of *hematopoiesis*—a mechanism during which self-renewing hematopoietic stem cells (HSCs) specialize, or differentiate, to produce mature blood cells. Understanding the details of this process is a fundamental problem in systems biology, and progress in uncovering these details 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)).

An HSC can give rise to any mature blood cell. In order to generate new mature blood cells (e.g., granulocytes, monocytes, T, B and natural killer (NK) cells) an HSC first becomes a multipotent progenitor cell. This cell then further differentiates into progenitors with more limited potential. An HSC can also divide or self-renew, giving rise to two daughter HSCs. Cells make fate decisions by a carefully orchestrated change in gene expression, but the details of these decision making processes are still not fully understood (Laslo et al. (2008), Whichard et al. (2010)). Mathematically, hematopoiesis can be represented as a *stochastic compartmental model* in which cells are assumed to self replicate and differentiate according to a Markov branching process (Becker, McCulloch and Till (1963), Kimmel and Axelrod (2002), Siminovitch, McCulloch and Till (1963)).

Although this mathematical representation of hematopoiesis is more than fifty years old (Till, McCulloch and Siminovitch (1964)), fitting branching process models to experimental data remains highly nontrivial. The main difficulty stems from the fact that estimating parameters of a partially observed stochastic process usually leads to intractable computational algorithms. One way to avoid this intractability is to base inference on deterministic models of hematopoiesis, as has been done by Colijn and Mackey (2005) and Marciniak-Czochra et al. (2009), for example. However, deterministic compartmental models are not suitable when cell counts are low in some of the compartments (Kimmel (2014)), which is frequently the case in many experimental protocols (e.g., bone marrow transplantation followed by blood cell reconstitution). Although working within the stochastic modeling framework is challenging, researchers were able to fit a two-compartmental stochastic model to X-chromosome inactivation marker data (Abkowitz et al. (1990), Fong, Guttorp and Abkowitz (2009), Golinelli, Guttorp and Abkowitz (2006), Catlin et al. (2011)). Such studies have produced important insights, but this simple two-compartmental model 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 cell lineages descended from individual HSCs. It should be noted that even these simplified models capturing the clonal dynamics descended from an HSC have posed significant statistical and computational challenges.

Recently emergent experimental techniques now allow researchers to track the dynamics of cell lineages descended from distinct ancestral progenitor cells or HSCs. 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)). Each cell descended from an original barcoded population inherits the unique identifier of its ancestor. The data thus enable us to distinguish individual lineages, and comprise independent and identically distributed time series. This marked departure from previous batch experimental data, in which observations were coming from the population of cells descended from a mixture of indistinguishable cells, potentially allows for investigation of much more realistic models of hematopoiesis. 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 barcoding data are certainly more informative than those from previous experiments, statistical methods capable of analyzing such data are only beginning to emerge. Buchholz et al. (2013) develop a model fitting technique for *in vivo* fate mapping data that is closely related to our work, but their approach lacks a cell sampling model needed to analyze barcoding data. 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. They also do not fully take advantage of the information in the read count data, instead working with binary indicators of barcode presence. 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 all variation arising from experimental design in their analysis.

Wu et al. (2014) provide a preliminary analysis of their cellular barcoding data that reveals important scientific insights (Koelle et al. (2017)), but lacks the ability to perform statistical tasks such as parameter estimation and model 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. We propose a fully generative stochastic model and efficient method of parameter estimation that enables much richer hematopoietic structures to be statistically analyzed than previously possible. The following section describes the model and experimental design producing the dataset we will analyze. Next, we motivate our approach by statistically formulating our inferential goal

and deriving the necessary mathematical quantities in Section 2.3. We then thoroughly validate these methods via several simulation studies, fit the models to the rhesus macaque barcoding data, and compare the fitted models via cross validation. Finally, we close with a discussion of these results, their implications, and avenues for future work.

2. Methods.

2.1. *Data and experimental setup*. During hematopoiesis, self-renewing hematopoietic stem cells specialize or *differentiate* via a series of intermediate progenitor cell stages to produce mature blood cells (Weissman (2000)). A challenge in studying this system *in vivo* is that only the mature cells are observable, as they can be sampled from the blood. We will model hematopoiesis as a continuous-time stochastic process whose state $\mathbf{X}(t)$ is a vector of cell counts of different types (e.g., HSCs, progenitors, T, B cells). We will provide mathematical formulation of the stochastic process after a complete description of the dataset $\mathbf{Y}$. In contrast to previous studies, the single cell lineage tracking dataset we will analyze opens the possibility of inferring intermediate progenitor behavior. We briefly describe the cellular barcoding experiment that makes this possible.

Wu et al. (2014) extract HSCs and progenitor cells from the marrow of a rhesus macaque, and use lentiviral vectors to insert unique DNA sequences into the cells that will each act as an identifying "barcode." After the extracted cells are labeled in this way, the macaque is irradiated so that its residual blood cells are depleted. Next, the labeled cells are transplanted back into the marrow of the animal; reconstitution of its entire blood system is supported from this initial labeled population of extracted cells. All cells descended from a marked cell—its *lineage*—inherit its unique barcode ID; we remark that what we call a lineage is often referred to as a clone in the hematopoiesis literature. We assume that barcoded lineages act independently from each other and that each barcoded lineage $p \in \{1, \ldots, N\}$ is a realization $\mathbf{X}^p(t)$ of our stochastic process model of hematopoiesis.

Hematopoietic reconstitution is monitored indirectly over time at discrete observation times t_j. At each t_j, the experimental protocol consists of sampling blood from the macaque and separating the sample by cell type, followed by retrieving the barcodes via DNA sequencing from each sorted population. Specifically, the blood sample is sorted into five mature cell categories: monocyte (Mono), granulocyte (Gr), T, B and natural killer (NK) cell types. These type-sorted samples will be denoted $\widetilde{\mathbf{y}}_m(t_j)$ for each mature cell type m, and are of fixed size b_m at all observation times. That is, each entry $\widetilde{y}_m^p(t_j)$ is the number of type m cells with barcode p present in the sample, and $\sum_p \widetilde{y}_m^p(t_j) = b_m$ at every t_j. The random number of barcodes present in the samples is proportional to their prevalence in the total population of labeled type m cells in the population, denoted $B_m(t_j) = \sum_p X_m^p(t_j)$, where $X_m^p(t_j)$ denotes the true blood count of type m cells from lineage p at

time t_j. Therefore, the distribution of sampled cells can be modeled by a multivariate hypergeometric distribution

$$\widetilde{\mathbf{y}}_m(t) \mid \mathbf{X}(t) \sim \text{mvhypergeom}\big(B_m(t), \mathbf{X}_m(t), b_m\big). \tag{2.1}$$

Put another way, $\Pr(\widetilde{y}_m^p(t) = z)$ is the probability of drawing z 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. Recall that the sample sizes b_m are fixed and known from the experimental protocol. While the latent processes are unknown, the values of their sum $B_m(t_j)$ are observed: the total circulating blood cell (CBC) counts are recorded at each sampling time. We do not consider potential measurement error in the CBC data, and therefore do not model $B_m(t_j)$ as random variables throughout, instead treating $B_m(t_j)$ as external known constants.

Next, individual barcodes must be retrieved or *read* via sequencing. DNA is extracted from each of the sorted samples, and polymerase chain reaction (PCR) is performed to generate many copies of the DNA segments. This step aids barcode retrieval by increasing detectability of DNA segments present in the sample during sequencing. It is commonly assumed that PCR amplification preserves the proportion of barcodes present. We disregard experimental noise that may cause negligible departures from this standard assumption in order to avoid modeling PCR itself as an additional stochastic process. The read count $y_m^p(t) = d_m(t_j) \times \widetilde{y}_m^p(t_j)$ is obtained by sequencing this amplified PCR product; here $d_m(t)$ is an unknown constant representing the linear effect of PCR amplification. Thus, at each observation time t_j, the experiment yields a count $y_m^p(t_j)$ denoting the number of times barcode ID p was read after sequencing the type m cell sample.. An illustration summarizing the process for one lineage is provided in Figure 1.

Our assumption that PCR amplification is linear may be less suitable for barcoded populations with low counts, as any noise we have chosen not to model from this process may have a larger relative effect. We therefore further filter the data similarly to (Wu et al. (2014)) to include only barcode IDs exceeding 1000 reads. Altogether, our observed dataset consists of over 110 million read counts across $N = 9635$ unique barcode IDs, obtained at irregularly spaced times over a total period of $t_J = 30$ months. This collection of read counts can be viewed as a three-dimensional array, where the first array index m corresponds to mature cell type m. Fixing this index results in a $N \times J$ matrix $\mathbf{Y}_m = (\mathbf{y}_m(t_1), \mathbf{y}_m(t_2), \ldots, \mathbf{y}_m(t_J))$. The second array index, columns of each such matrix described above, correspond to observation (sampling) times $\mathbf{t} = (t_1, \ldots, t_J)$. The third array dimension indexes barcodes: $\mathbf{y}_m^p$, the pth row of $\mathbf{Y}_m$, encodes the read count time series corresponding to a unique barcode ID $p \in \{1, \ldots, N\}$ among the population of type m mature blood cells.

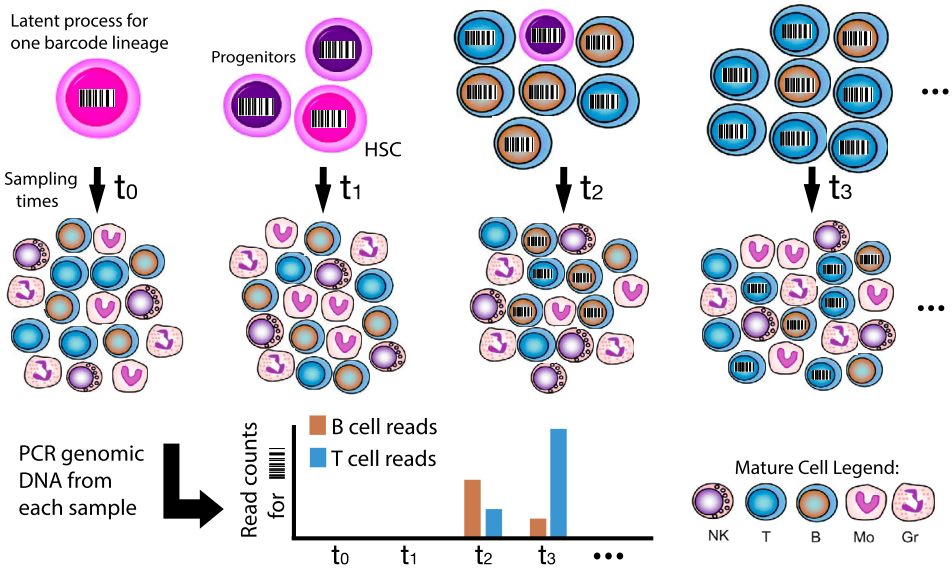

FIG. 1. *Illustration of experimental protocol for one single fixed barcode ID. The top panel represents the latent process starting with a single HSC (pink) at several snapshots in time* $t_0, \ldots, t_3$. *The second panel illustrates blood samples. Note that the barcode only becomes present in the blood when mature cells, which first appear 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.*

2.2. *Multi-type branching model of latent process.* The data $\mathbf{Y}$ form a *partial observation* of a collection of p IID continuous-time latent processes, each evolving according to the stochastic model $\mathbf{X}(t)$. We now provide a biological description of the underlying hematopoietic process we wish to model by $\mathbf{X}(t)$, followed by mathematical details of our proposed class of branching process models.

Hematopoiesis begins with bone marrow residing HSCs, which have the capacity to self-renew (give rise to another HSC) or differentiate into more specialized progenitor cells. Biologists have not reached a consensus about how many types of progenitors exist in this intermediate stage, but agree that intermediate progenitor cells lose the ability to proliferate, and each progenitor type can produce one or several types of mature blood cells before exhausting its own lifespan (cell death). These mature blood cells exist at the last phase of development, are found mainly in the bloodstream and do not give birth to any further cells. Based on this biological understanding of hematopoiesis, a multi-type branching process taking values over a discrete state space of cell counts in continuous time provides a natural modeling choice. Canonical differentiation trees that have been posited in the scientific literature follow such a structure, and such stochastic models have estab-

lished their place in the statistical hematopoiesis literature (Kimmel and Axelrod (2002), Catlin et al. (2011)).

A continuous-time branching process is a Markov jump process in which a collection of independently acting particles (cells) can reproduce and die according to a probability distribution. Each cell type has a distinct mean lifespan and reproductive probabilities, and can give rise to cells of its own type as well as other types at its time of death. Our branching process models consist of an HSC stage, progenitor stage, and mature cell stage, and allow for an arbitrary number of progenitor and mature cell types to be specified. We use alphabetic subscripts $a, b \ldots \in \mathcal{A}$ to denote progenitors, with mature cell types indexed numerically by $m = 1, 2, \ldots \mathcal{M}$. The subscript 0 indicates quantities relating to HSCs. In our models, HSCs self-renew with rate λ, or become type a progenitor cells with differentiation rates ν_a. Progenitor cells exhaust their lifespan with rates μ_a, and produce type m mature blood cells with rates ν_m. Each mature cell type m is descended from only one progenitor type, so that its corresponding production rate ν_m is unique and well defined. Finally, these mature cells exhaust with rates μ_m. Figure 2 depicts several example structures contained in this class. In a given branching model, let $C = 1 + |\mathcal{A}| + \mathcal{M}$ be the total number of cell types. The process state is a length C random vector $\mathbf{X}(t) = (X_0(t), X_a(t), \ldots, X_{|\mathcal{A}|}(t), X_1(t), \ldots, X_{\mathcal{M}}(t))$ taking values in the countably infinite state space $\Omega = \mathbb{N}^C$, whose components represent sizes of the cell populations at time $t \geq 0$. Recall the read count data $\mathbf{Y}$ are obtained by sequencing blood samples from the mature populations $X_1(t), \ldots, X_{\mathcal{M}}(t)$; the early stage populations $X_0(t), X_a(t), \ldots, X_{|\mathcal{A}|}(t)$ are entirely unobserved.

The behavior of $\mathbf{X}(t)$ is then defined by specifying a set of length C *instantaneous rate vectors*. To introduce the remaining notation, we focus on the simplest model displayed in Figure 2(a) with $C = 5$ total cell types for concreteness. Model 2(a) features one progenitor type and three mature cell types. The instantaneous rate $\alpha_0(n_0, n_a, n_1, n_2, n_3)$ (Dorman, Sinsheimer and Lange (2004), Lange (2010)) contains the rate of an event occurring in which an HSC cell produces n_0 HSCs, n_a progenitors, and n_m of each type m mature cells. These rate vectors are analogous for other parent cell types: for instance, $\alpha_a(n_0, n_a, n_1, n_2, n_3)$ denotes the same rates of production from one type a progenitor cell rather than an HSC cell. The offspring descended from each cell subsequently behave according to the same set of rate vectors, which do not change with t—the process is *time-homogeneous*.

The assumption that cells act independently implies that the process rates are *linear*: overall event rates at the population level are multiplicative in the number of cells. Together, these assumptions imply that the lifespan of each HSC follows an exponential distribution with parameter $-\beta_0 := \sum_{(n_0,n_a,n_1,n_2,n_3)\neq(1,0,0,0,0)} \alpha_0(n_0, n_a, n_1, n_2, n_3)$. After this exponential waiting time, the probability that the cell is replaced by $(n_0, n_a, n_1, n_2, n_3)$ cells of each respective type is given by normalizing the corresponding rate: $\alpha_0(n_0, n_a, n_1, n_2, n_3)/\beta_0$. The same holds analogously

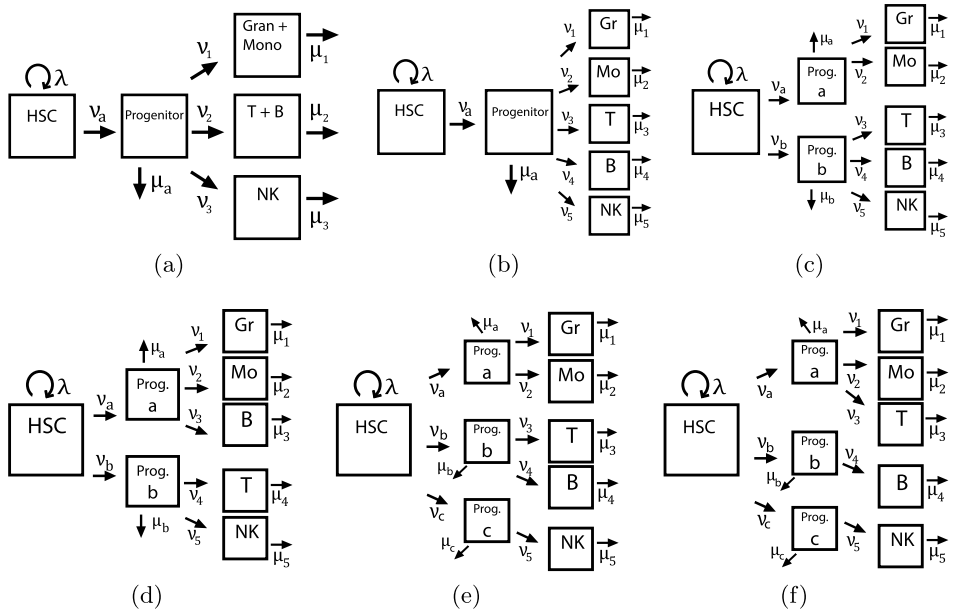

FIG. 2. *Differentiation trees to be considered in simulation study and real data analysis. In the first two models, mature cells are descended from one common multipotent progenitor*: (*a*) *groups similar mature cells* (*i.e., T and B cells are not distinguishable*), *leading to a model with three mature cell types, and* (*b*) *models each observed mature cell type separately. Note that previous statistical studies by Catlin, Abkowitz and Guttorp* (2001), *Fong, Guttorp and Abkowitz* (2009), *Golinelli, Guttorp and Abkowitz* (2006) *model only two compartments* (*types*), *containing HSC and "other" cells. Models* (*c*)–(*f*) *include several biologically plausible topologies featuring two or three oligopotent progenitors, each specializing to produce only particular mature cells.*

for all cells after replacing subscripts appropriately, and therefore $\mathbf{X}(t)$ evolves over time as a continuous-time Markov chain (CTMC) (Guttorp (1995), Chapter 3).

As an example, we see from Figure 2 that model 2(a) is characterized by the parameters $\boldsymbol{\theta} = (\lambda, \nu_a, \mu_a, \nu_1, \nu_2, \nu_3, \mu_1, \mu_2, \mu_3)$. Specifying such a process as a CTMC classically using the rate matrix (infinitesimal generator) is mathematically unwieldy—this is an infinite matrix with no simplifying structure for these models. However, the process can be compactly specified using the following instantaneous rate vectors of a branching process:

$$\alpha_0(2,0,0,0,0) = \lambda, \qquad \alpha_0(0,1,0,0,0) = \nu_a,$$
$$\alpha_0(1,0,0,0,0) = -(\lambda + \nu_a), \qquad \alpha_a(0,0,0,0,0) = \mu_a,$$
$$\alpha_a(0,1,1,0,0) = \nu_1, \qquad \alpha_a = (0,1,0,1,0) = \nu_2,$$
$$\alpha_a(0,1,0,0,1) = \nu_3, \qquad \alpha_a(0,1,0,0,0) = -(\mu_a + \nu_1 + \nu_2 + \nu_3),$$
$$\alpha_1(0,0,0,0,0) = \mu_1, \qquad \alpha_1(0,0,1,0,0) = -\mu_1,$$

$$\alpha_2(0,0,0,0,0) = \mu_2, \qquad \alpha_2(0,0,0,1,0) = -\mu_2,$$
$$\alpha_3(0,0,0,0,0) = \mu_3, \qquad \alpha_3(0,0,0,0,1) = -\mu_3,$$

with all other rates zero. An equivalent representation in terms of chemical kinetic rate notation is provided in Appendix A.1 in the Supplementary Material (Xu et al. (2019)). Given that the instantaneous branching rates of each model can be specified this way from the parameter vector $\boldsymbol{\theta}$, methods of inference in the next section target $(\boldsymbol{\theta}, \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is an initial distribution vector $\boldsymbol{\pi} = (\pi_0, \pi_a, \pi_b, \ldots)$. The components π_a represents the probability that a lineage is originally descended from a transplanted progenitor rather than from a transplanted HSC: this is unknown since the initial barcoding is applied to a heterogeneous transplanted cell population containing HSCs and early progenitors.

2.3. *Parameter estimation procedure*. We estimate model parameters using the generalized method of moments, a computationally simpler alternative to maximum likelihood estimation that yields consistent estimators. 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)). The choice is well motivated when a large number of samples is available, as is the case for our dataset consisting of thousands of IID barcoded lineages. The 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 is minimized to estimate parameters of interest. Our estimator seeks to match pairwise empirical read count correlations across barcodes with their corresponding model-based population correlations.

We find explicit analytic forms for the first and second moments of the models presented in Section 2.2, allowing for fast computation of the marginal correlations between any two mature cell type counts. Our derivations hold for all instances of a rich class of models, including those displayed in Figure 2, enabling arbitrary groupings of cell types and candidate branching pathways to be investigated. To perform model selection, we use cross-validation with respect to the objective function introduced below. Further, nonparametric bootstrap confidence intervals can be obtained by repeating the estimation procedure on resampled data.

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, the scale invariance of correlations allows us to avoid modeling PCR amplification on top of an already complex model, as the amplification constants $d_m(t)$ cancel out. This robustness also comes with a caveat—we may not expect all rates to be identifiable. For instance, we cannot detect a proportional increase in the differentiation rate into a mature cell compartment and its death rate without scale information. Indeed, the same phenomenon was reported in fitting simpler models of

hematopoiesis (Catlin, Abkowitz and Guttorp (2001)). Fortunately, because mature blood cells are observable in the bloodstream, their lifespans have been well studied separately in the biological literature (Kaur et al. (2008), Zhang et al. (2007)). We fix the death rates μ_i at these known quantities, providing scale information for our estimator. Doing so allows the remaining free parameters to be well estimated by the same procedure. This is further discussed in the following section, and empirically shown to be an effective strategy in Section 3.1.

2.4. *Correlation loss function*. To estimate the parameter vector $\boldsymbol{\theta}$ containing process rates and initial distribution $\boldsymbol{\pi}$, we seek to closely match model-based correlations to the empirical correlations between observed read counts. We overload the notation so that if any components of the parameter vector (such as death rates) are assumed fixed at their true values, then $\boldsymbol{\theta}$ is understood to denote the remaining free parameters of the model. We estimate free parameters $\boldsymbol{\theta}$ via minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{t_j} \sum_{m} \sum_{n \neq m} \big[\psi_{mn,j}(\boldsymbol{\theta}; \mathbf{Y}) - \hat{\psi}_{mn,j}(\mathbf{Y})\big]^2, \tag{2.2}$$

where $\psi_{mn,j}$ represents model-based correlation between reads of type m, n mature cells at time t_j:

$$\psi_{mn,j}(\boldsymbol{\theta}; \mathbf{Y}) := \rho\big(Y_m(t_j), Y_n(t_j); \boldsymbol{\theta}\big) = \frac{\operatorname{Cov}[Y_m(t_j), Y_n(t_j); \boldsymbol{\theta}]}{\sigma(Y_m(t_j); \boldsymbol{\theta})\sigma(Y_n(t_j); \boldsymbol{\theta})},$$

and $\hat{\psi}_{mn,j}$ denotes the corresponding sample correlations across barcodes $p = 1, \ldots, N$ at time t_j:

$$\begin{aligned}\hat{\psi}_{mn,j}(\mathbf{Y}) &:= \hat{\rho}\big(\mathbf{y}_m(t_j), \mathbf{y}_n(t_j)\big) \\ &= \frac{\sum_{p=1}^{N} (y_m^p(t_j) - \overline{y}_m(t_j))(y_n^p(t_j) - \overline{y}_n(t_j))}{\sqrt{\sum_{p=1}^{N} (y_m^p(t_j) - \overline{y}_m(t_j))^2}\sqrt{\sum_{p=1}^{N} (y_n^p(t_j) - \overline{y}_n(t_j))^2}}.\end{aligned}$$

The problem of estimating hematopoietic rates now translates to seeking

$$\hat{\boldsymbol{\theta}}_N = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \big\|\mathbf{G}_N(\boldsymbol{\theta}; \mathbf{Y})\big\|_2^2, \qquad \mathbf{G}_N(\boldsymbol{\theta}; \mathbf{Y}) := \boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y}) - \hat{\boldsymbol{\psi}}(\mathbf{Y}),$$

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. Again, if a subset of parameters is fixed, $\hat{\boldsymbol{\theta}}_N$ is understood to contain the estimates of all free parameters in the model. Denoting the true data generating parameters by $\boldsymbol{\theta}_0$, the law of large numbers implies that $\mathrm{E}[\mathbf{G}_N(\boldsymbol{\theta_0}; \mathbf{Y})] \to 0$ as the number of processes $N \to \infty$. Our method is therefore an *M-estimator*, also known as generalized method of moments (GMM) (van der Vaart (1998), Chapter 5). M-estimators are known to be consistent under general conditions as summarized in the following theorem. We

note that the regularity conditions, detailed in the Appendix, are often assumed rather than proven—in particular, identifiability is notoriously difficult to establish in nonlinear models. However, in Section 3 we provide strong empirical evidence that fixing the death rates ensures identifiability of the remaining parameters.

THEOREM 2.1. *Under regularity conditions A1–A3 (see Appendix A.2), the sequence* $\{\hat{\boldsymbol{\theta}}_N\}$ *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 processes or rows in* $\mathbf{Y}$.

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

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

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 moment equations to have unequal contributions to the objective function; $\hat{\mathbf{W}}$ intuitively may assign 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 (Pakes and Pollard (1989), van der Vaart (1998)), and are asymptotically efficient under the optimal $\hat{\mathbf{W}}$ (Hansen (1982)). While many algorithms exist for estimating $\hat{\mathbf{W}}$, the task is nontrivial (Hansen, Heaton and Yaron (1996)). Because we have a large enough dataset such that finite-sample efficiency is of lesser concern, 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.

2.5. *Moments of the multi-type branching process.* Here we derive the analytic expressions for the first and second moments of the latent branching processes defined in Section 2.2, enabling efficient computation of model-based correlations $\psi_{mn,j}(\boldsymbol{\theta}, \mathbf{Y})$ appearing in the loss function. For quantities relating to all types, we will use the common index $i = 0, a, b, \ldots, 1, \ldots, \mathcal{M}$, and adopt the notation $\mathbf{e}_i$ to represent the vector of length C (denoting total number of cell types) with a 1 in the type i component and is 0 elsewhere. The indicator $\mathbf{1}_{\{a \to m\}}$ equals 1 if mature cell type m is descended from progenitor type a in a given model, and 0 otherwise.

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. We begin by writing the *pseudo-generating functions*, also called progeny generating functions (Dorman, Sinsheimer and Lange (2004)), defined as

$$u_i(\mathbf{s}) = \sum_{k_0} \sum_{k_a} \cdots \sum_{k_{\mathcal{M}}} a_i(k_0, \ldots, k_{\mathcal{M}}) s_0^{k_0} s_a^{k_a} \cdot s_{\mathcal{M}}^{k_{\mathcal{M}}}, \tag{2.3}$$

where $\mathbf{s}$ is a vector of dummy variables confined to the $[0, 1]$ interval. For our class of models, these are given by

$$\begin{aligned} u_0(\mathbf{s}) &= \lambda s_0^2 + \sum_{a \in \mathcal{A}} \nu_a s_a - \Bigg(\lambda + \sum_{a \in \mathcal{A}} \nu_a\Bigg) s_0, \\ u_a(\mathbf{s}) &= \sum_{m=1}^{\mathcal{M}} \nu_m s_a s_m \mathbf{1}_{\{a \to m\}} + \mu_a - \Bigg(\mu_a + \sum_{m=1}^{\mathcal{M}} \nu_m \mathbf{1}_{\{a \to m\}}\Bigg) s_a \quad \forall a \in \mathcal{A}, \\ u_m(\mathbf{s}) &= u_m(s_m) = \mu_m - \mu_m s_m \quad \forall m = 1, \ldots, \mathcal{M}. \end{aligned} \tag{2.4}$$

Next, we can write the probability generating function (PGF) of the process, beginning with one HSC, via a relation to the pseudo-generating function u_0 as follows:

$$\begin{aligned} \phi_0(t; \mathbf{s}) &= \mathrm{E}\Bigg[\prod_i s_i^{X_i(t)} \Big| \mathbf{X}(0) = \mathbf{e}_0\Bigg] \\ &= \sum_{k_0=0}^{\infty} \cdots \sum_{k_{\mathcal{M}}=0}^{\infty} \Pr_{\mathbf{e}_0, (k_0, k_a, \ldots k_{\mathcal{M}})}(t) s_0^{k_0} s_a^{k_a} \cdots s_{\mathcal{M}}^{k_{\mathcal{M}}} \\ &= \sum_{k_0=0}^{\infty} \cdots \sum_{k_{\mathcal{M}}=0}^{\infty} \big[\mathbf{1}_{\{k_0=1, k_a=\cdots=k_{\mathcal{M}}=0\}} \\ &\quad + a_0(k_0, \ldots, k_{\mathcal{M}}) t + o(t)\big] s_0^{k_0} s_a^{k_a} \cdots s_{\mathcal{M}}^{k_{\mathcal{M}}} \\ &= s_0 + u_0(\mathbf{s}) t + o(t). \end{aligned} \tag{2.5}$$

Analogously defining ϕ_i for processes beginning with one type i cell ($i = 1, \ldots, C$), equation (2.5) yields the relations

$$\frac{\partial}{\partial t} \phi_i(t, \mathbf{s}) = u_i\big(\phi_0(t, \mathbf{s}), \ldots, \phi_{\mathcal{M}}(t, \mathbf{s})\big). \tag{2.6}$$

That is, the right hand side of each ordinary differential equation (ODE) in the Kolmogorov backward equations (2.6) takes the form of the right hand sides of (2.4) with s_i everywhere replaced by $\phi_i(t, \mathbf{s})$. Though the system is generally not solvable except in simple models, the full solution is not necessary as our method uses only the moments rather than all of the transition probability information. In fact, we only require moments conditional on one initial cell, since each latent process represents barcode lineages descended from a single marked cell. 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:

$$M_{m|i}(t) = \frac{\partial}{\partial s_m}\phi_i(t,\mathbf{s})\Big|_{s_0=s_a=\cdots=s_{\mathcal{M}}=1}.$$

Similarly, we may further differentiate the PGF to derive second moments used toward variance and covariance calculations. The relationship $U_{mn|i}(t) = \frac{\partial^2\phi_i}{\partial s_m \partial s_n}|_{\mathbf{s}=\mathbf{1}}$ holds, where

$$U_{mn|i}(t) := \mathrm{E}\big[X_m(X_n - \mathbf{1}_{\{m=n\}})|\mathbf{X}(0) = \mathbf{e}_i\big].$$

These identities 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,

$$\frac{\partial}{\partial t}M_{m|i}(t) = \frac{\partial^2\phi_i}{\partial t\,\partial s_m}\Big|_{\mathbf{s}=\mathbf{1}} = \sum_k \frac{\partial u_i}{\partial s_k}\frac{\partial \phi_k}{\partial s_m}\Big|_{\mathbf{s}=\mathbf{1}}, \tag{2.7}$$

$$\begin{aligned}\frac{\partial}{\partial t}U_{mn|i}(t) &= \frac{\partial^3\phi_i}{\partial t\,\partial s_m\,\partial s_n}\Big|_{\mathbf{s}=\mathbf{1}} \\ &= \sum_{j=1}\left(\frac{\partial u_i}{\partial \phi_j}\frac{\partial^2\phi_j}{\partial s_m\,\partial s_n}\right) + \sum_{j,k=1}\left(\frac{\partial^2 u_i}{\partial\phi_j\,\partial\phi_k}\frac{\partial\phi_j}{\partial s_m}\frac{\partial\phi_k}{\partial s_n}\right)\Big|_{\mathbf{s}=\mathbf{1}}.\end{aligned} \tag{2.8}$$

Notice equation (2.7) 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 (2.8) governing second moments. These systems are subject to the initial conditions $M_{m|i}(0) = \mathbf{1}_{\{m=i\}}$ and $U_{mn|i}(0) = 0$. We introduce the notation $\kappa_{ij} = \frac{\partial u_i}{\partial s_j}|_{\mathbf{s}=\mathbf{1}}$ for brevity; as an example, for all $a \in \mathcal{A}$, $m = 1,\ldots,\mathcal{M}$,

$$\kappa_{00} = \lambda - \sum_{a\in\mathcal{A}}\nu_a, \qquad \kappa_{aa} = -\mu_a, \qquad \kappa_{mm} = -\mu_m,$$

$$\kappa_{0a} = \nu_a, \qquad \kappa_{am} = \nu_m \mathbf{1}_{\{a\to m\}}.$$

While the number of equations increases in going from PGFs to conditional moments, surprisingly, the new system admits closed form solutions. To provide some intuition, we may solve the equations (2.7) beginning with the mature cells separately, and successively back-substitute to obtain solutions for the means conditional on beginning with a progenitor, and in turn with an HSC. These mean expressions can then be plugged into the system of second moment equations (2.8), where the strategy can be repeated. While the equations become more and more complicated, they retain the same general form owing to linearity of the branching process—if the model had feedback loops from interaction between cells, this approach would fall apart. Finally, we arrive at the unconditioned moments by

marginalizing over $\boldsymbol{\pi}$, accounting for uncertainty of initial cell type. Because mature cells cannot give rise to additional types, their ODEs are pure death equations, which are solved straightforwardly to obtain

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

These solutions are substituted into moment expressions conditional on beginning with a marked progenitor: by (2.7), 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

$$\frac{d}{dt} M_{m|a}(t) + P(t) M_{m|a}(t) = Q(t). \tag{2.9}$$

Upon inspection, this differential equation remains first order and can be solved using the integrating factor method: multiplying both sides by $e^{\int P(t)\,dt}$ and rearranging for $M_{m|a}(t)$ allows us to solve by applying the product rule in reverse, yielding

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

These expressions are intuitive: a higher rate of differentiation ν_m leads to an increase in the mean population of type m cells, while a larger death rate μ_m relative to the death rate of progenitors μ_a producing the type m cells decreases their mean population. Next, (2.7) 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 (2.9). Thus, we can plug in the solutions we've obtained for $M_{m|a}(t)$ and solve the system using the same technique, yielding

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

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 $\kappa_{i,jk} := \frac{\partial^2 u_i}{\partial s_j \partial s_k}|_{\mathbf{s}=\mathbf{1}}$; 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. Now, many terms in equation (2.8) have zero contribution, and the remaining equations in the system can be simplified to yield

$$\begin{aligned}
\frac{d}{dt}U_{mn|a}(t) &= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\bigg(\frac{\partial u_a}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m\partial s_n} + \frac{\partial^2 u_a}{\partial s_a\partial s_m}\frac{\partial\phi_a}{\partial s_n}\frac{\partial\phi_m}{\partial s_m} \\
&\qquad + \frac{\partial^2 u_a}{\partial s_a\partial s_n}\frac{\partial\phi_a}{\partial s_m}\frac{\partial\phi_n}{\partial s_n}\bigg) \\
&= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}(\kappa_{aa}U_{mn|a} + \kappa_{a,am}M_{n|a}M_{m|m} + \kappa_{a,an}M_{m|a}M_{n|n}), \\
\frac{d}{dt}U_{mn|0}(t) &= \bigg(\frac{\partial u_0}{\partial s_0}\frac{\partial^2\phi_0}{\partial s_m s_n} + 2\frac{\partial^2 u_0}{\partial s_0^2}\frac{\partial\phi_0}{\partial s_m}\frac{\partial\phi_0}{\partial s_n} \\
&\qquad + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\frac{\partial u_0}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m s_n}\bigg)\bigg|_{\mathbf{s}=\mathbf{1}} \\
&= \kappa_{00}U_{mn|0} + 2\kappa_{0,00}M_{m|0}M_{n|0} + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\kappa_{0a}U_{mn|a}.
\end{aligned}$$

Similarly,

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

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 (2.9), and we find

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

Replacing $\kappa_{\cdot}$ with explicit rates, we integrate and simplify these expressions to obtain

$$U_{mm|a}(t) = \mathbf{1}_{\{a\to m\}}\frac{2\nu_m^2}{\mu_m-\mu_a}e^{-\mu_a t}\left[\frac{\mu_a-\mu_m}{\mu_m(\mu_a-2\mu_m)}-\frac{e^{-\mu_m t}}{\mu_m}-\frac{e^{(\mu_a-2\mu_m)t}}{\mu_a-2\mu_m}\right],$$

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

It is worth noting here that the product $\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}$ is zero for any pair of types m, n not descended from the same progenitor type, which may occur in models with specialized oligopotent progenitors. Recall that $\mathrm{Cov}[X_m(t), X_n(t)|\mathbf{X}(0) = \mathbf{e}_a)] = U_{mn|a}(t) - M_{m|a}(t)M_{n|a}(t)$. We see that in this case, U_{mn} becomes zero, leading to lower values of the model-based covariance. In particular, $\mathrm{Cov}[X_m(t), X_n(t)|\mathbf{X}(0) = \mathbf{e}_a)]$ may be negative when m, n do not share a progenitor type. This gives some intuition on the substantial effect of progenitor structure, which becomes apparent in the results presented in Section 3.

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:

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

These integrals have closed form solutions since their integrands only differ from the previous set of equations by additional exponentials contributed 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 computed within iterative algorithms. For completeness, we include the explicit solutions to the simplest model in the Appendix A.3.

2.6. *Marginalized moments*. With closed form moment expressions in hand, we can readily recover variance and covariance expressions for mature cells and

thus calculate model-based correlations. For instance,

$$\operatorname{Cov}\bigl[X_m(t), X_n(t)|\mathbf{X}(0) = \mathbf{e}_j\bigr] = U_{mn|j}(t) - M_{m|j}(t)M_{n|j}(t).$$

Because the initial state is uncertain, unconditional variances and covariances between mature types can be computed by marginalizing over the initial distribution vector $\boldsymbol{\pi}$. Derivations for all of the following expressions in this section are included in the Appendix A-3. We arrive at the marginal expressions by applying the law of total (co)variance:

$$\begin{aligned}
&\operatorname{Var}\bigl[X_m(t)\bigr] \\
&\quad= \sum_{k=1}^{K} \pi_k \mathrm{E}\bigl[X_{m|k}^2\bigr] - \sum_{k=1}^{K} \pi_k^2 \bigl(\mathrm{E}[X_{m|k}]\bigr)^2) \\
&\qquad - 2\sum_{j>k} \pi_j \pi_k \mathrm{E}[X_{m|j}]\mathrm{E}[X_{m|k}] \\
&\quad= \sum_{k=1}^{K} \pi_k \bigl[U_{mm|k}(t) + M_{m|k}(t)\bigr] - \pi_k^2 M_{m|k}(t)^2 \\
&\qquad - 2\sum_{j>k} \pi_j \pi_k M_{m|k}(t) M_{m|j}(t),
\end{aligned} \tag{2.10}$$

$$\begin{aligned}
&\operatorname{Cov}\bigl[X_m(t), X_n(t)\bigr] \\
&\quad= \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{m|k}X_{n|k}] - \sum_{k=1}^{K} \pi_k^2 \mathrm{E}[X_{m|k}]\mathrm{E}[X_{n|k}] \\
&\qquad - \sum_{k\neq l} \pi_k \pi_l \mathrm{E}[X_{m|k}]\mathrm{E}[X_{n|l}] \\
&\quad= \sum_{k=1}^{K} \pi_k U_{mn|k}(t) - \pi_k^2 M_{m|k}(t) M_{n|k}(t) \\
&\qquad - \sum_{k\neq l} \pi_k \pi_l M_{m|k}(t) M_{n|l}(t).
\end{aligned} \tag{2.11}$$

It remains to relate these expressions to the correlations between read counts $\psi_{mn}(\boldsymbol{\theta}; \mathbf{Y})$. Applying the law of total (co)variance again with respect to the multivariate hypergeometric sampling distribution, we obtain the following expressions:

$$\operatorname{Cov}\bigl[Y_m(t), Y_n(t)\bigr] = \frac{b_m b_n}{B_m(t) B_n(t)} \operatorname{Cov}\bigl[X_m(t), X_n(t)\bigr], \tag{2.12}$$

$$\begin{aligned}\operatorname{Var}[Y_m(t)] &= \frac{b_m(B_m(t)-b_m)}{B_m(t)(B_m(t)-1)}\operatorname{E}[X_m(t)] \\ &\quad - \frac{b_m(B_m(t)-b_m)}{B_m(t)^2(B_m(t)-1)}\operatorname{E}[X_m^2(t)] \\ &\quad + \frac{b_m^2}{B_m(t)^2}\operatorname{Var}[X_m(t)].\end{aligned} \tag{2.13}$$

We note that the last set of variance and covariance expressions is an approximation, because we treated $B_m(t)$ as a constant. In Appendix A.3 we provide a justification for this approximation, and in our empirical evaluation did not observe any negative effects of this approximation.

2.7. *Implementation*. We implemented these methods in the R package `branchCorr`, available at https://github.com/jasonxu90/branchCorr. Software includes algorithms to simulate and sample from this class of branching process 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 reproductions of all simulations in this paper.

3. Results.

3.1. *Simulation study.* We examine performance of the 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. Under each model, we simulate 400 independent datasets, each consisting of 20,000 realizations representing distinct barcode IDs, from the specified continuous-time branching process. Since we observe fairly constant *in vivo* cell populations in the real data, 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 $t = 2$. 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 what we see in the real data.

From each synthetic dataset, we then sample an *observed dataset* according to the multivariate hypergeometric distribution, mimicking the noise from blood sampling. 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 the simulated datasets. To minimize the loss function, we use the general optimization implementation in R package `nlminb`. Optimization is performed over 250 random restarts per dataset. We constrain rates to be nonnegative, and include a log-barrier constraint to ensure that the overall growth of the

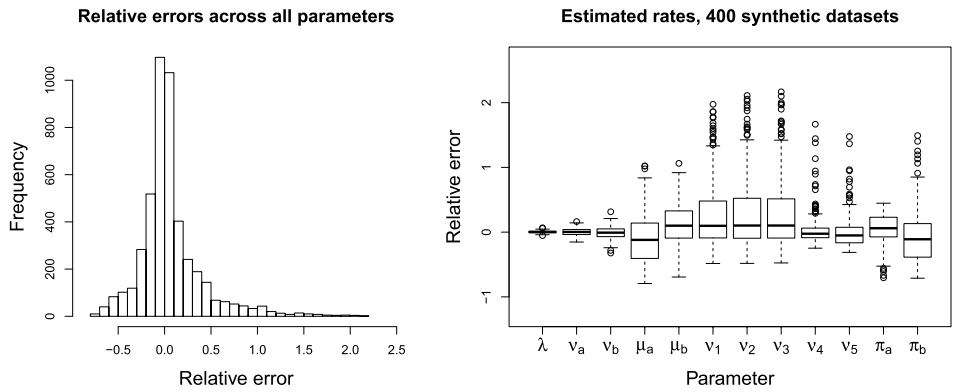

FIG. 3. *Performance of loss function estimator on synthetic data from model with five mature types and two progenitor types, that is, model* (c). *Parameters are successfully estimated despite the parameter rich setting. Additional plots assessing identifiability along with medians, median absolute deviations and standard errors appear in Appendix A*-5.

HSC reserve is nonnegative. The initial distribution vector is constrained to a probability simplex via a multinomial logistic reparametrization. Finally, we remark that optimizating over all free parameters leads to mild identifiability problems as anticipated. As mentioned in Section 2.3, correlations in the objective function are invariant to scale, and we expect some parameters to be distinguishable only up to a ratio. Intuitively, additional information to fix a sense of scale is required. Our simulation studies show that fixing the death rates μ_i at their true value remedies this problem: the corresponding intermediate rates ν_i, together with all other free parameters, can be recovered successfully. Results displayed in Figure 3 illustrate that estimates have low median error and are stable in that variation in the estimates and objective value is lower across random restarts of the optimization algorithm than across independent simulated datasets. The analogous findings for other models as well as detailed tables are reported in Appendix A-5. For real data, the strategy of fixing death rates is actionable since average lifespans of mature blood cells are 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 fitted 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 distinct groupings of correlation profiles, featuring 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 incorrectly assuming the data are generated from a model with one common progenitor or with three intermediate progenitors, and fitting these models to data simulated from the two-progenitor model (c) considered in the previous section.

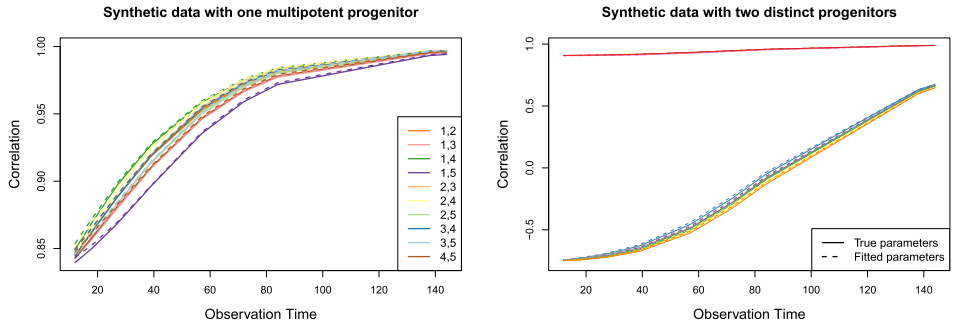

FIG. 4. *Pairwise correlation curves between five mature cell types descended from one common progenitor (left) or two distinct progenitors (right) calculated from our point estimate* $\hat{\boldsymbol{\theta}}$. *Solution curves from best fitting parameter estimates are almost indistinguishable from those corresponding to true parameters in both cases. Note that in the two-progenitor model, pairwise correlations among mature cell types display two clusters of behavior, and that negative correlations are possible.*

Figure 3 shows that the median over relative errors $\frac{\hat{\theta}_i - \theta_i}{\theta_i}$ of each component in the estimated parameter vector $\hat{\boldsymbol{\theta}}$ is near zero, and we note the median value of the objective function (2.2) 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 under misspecified progenitor structures are displayed in Figure 5, with detailed tables containing estimates again included in the Appendix A.5. We also examine the behavior when fitting a model with fewer types by "lumping" similar mature cells together. To this end, we consider grouping mature types 2 and 3 together, and types 4 and 5 together, thus fitting a model with three total mature types, but with a progenitor structure consistent with the

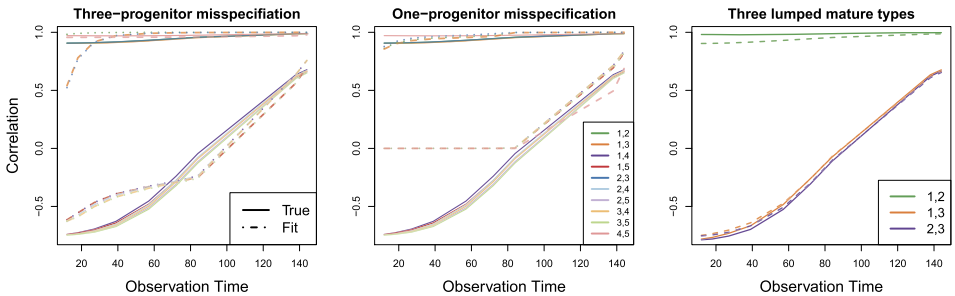

FIG. 5. *Fitted correlation curves corresponding to misspecified model estimates. Data are generated from a true model with two distinct progenitors and the true correlation profiles are the same as those displayed in the right panel of Figure 4. While we see a generic lack of fit in the three-progenitor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

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 to avoid overfitting a model to real data when some degree of model misspecification is inevitable, and is advantageous in settings where limited data suggest aggregation to reliably estimate fewer model parameters.

3.2. *Cell lineage barcoding in rhesus macaques*. Having validated our method on data simulated from the model, we turn to analyze the lineage barcoding data from Wu et al. (2014). We consider barcoding data collected from a rhesus macaque over a 30 month period following bone marrow transplantation. We include 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 only lineages exceeding a 1000 read count threshold at some time point are considered. After these restrictions, our dataset consists of 9635 unique barcode IDs, with read data available at 11 unevenly spaced sampling times.

As inputs to the loss function estimator, we fix death rates at biologically realistic parameters based on previous studies (Hellerstein et al. (1999), Zhang et al. (2007), Kaur et al. (2008)), reported below. Parameters of the multivariate hypergeometric sampling distribution are treated as known constants based on 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. Though we do model the measurement error process in CBC data, the measurement noise allows observed correlation profiles to be nonmonotonic, behavior that is also captured in the fitted model.

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 via loss function optimization with 2000 random restarts in models with one multipotent progenitor type are displayed in Figure 6. There are three curves in model (a) with three mature types, and 10 curves corresponding to possible pairs among the five mature types in model (b) plotted on the right. The empirical correlations from raw data are displayed as solid lines. On a qualitative level, there is visible separation into three clusters of correlation profiles among the five mature cell groups, consistent with the simpler lumped model (a). Cells that are "out of sync"—the scale of their differentiation and death rates are quite different—exhibit lower pairwise correlations. For instance, the abundant and short-lived granulocytes and monocytes have a high pairwise correlation compared to other pairs. 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,

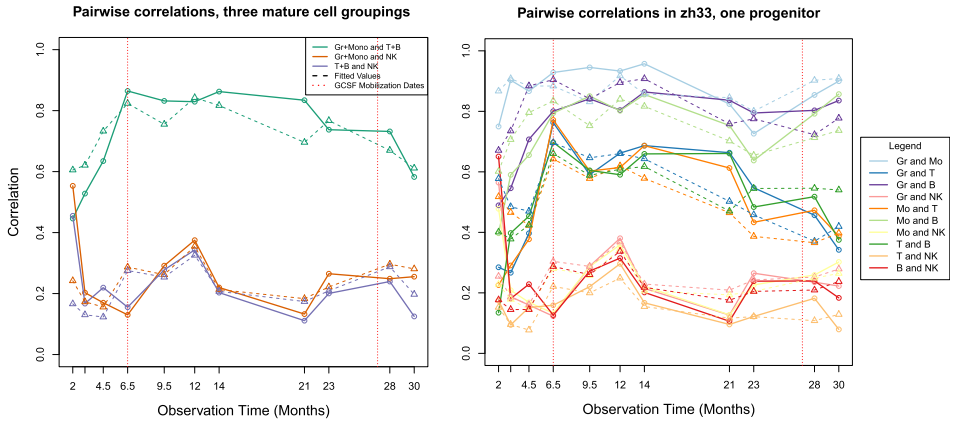

FIG. 6. *Dashed lines depict fitted correlations to read data in models* (*a*) *and* (*b*) *assuming one early progenitor type. GCSF mobilization dates are marked by vertical red lines. Solid lines connect the empirical correlations.*

from a new perspective. In both plots, fitted curves 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.

Next, we display a 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

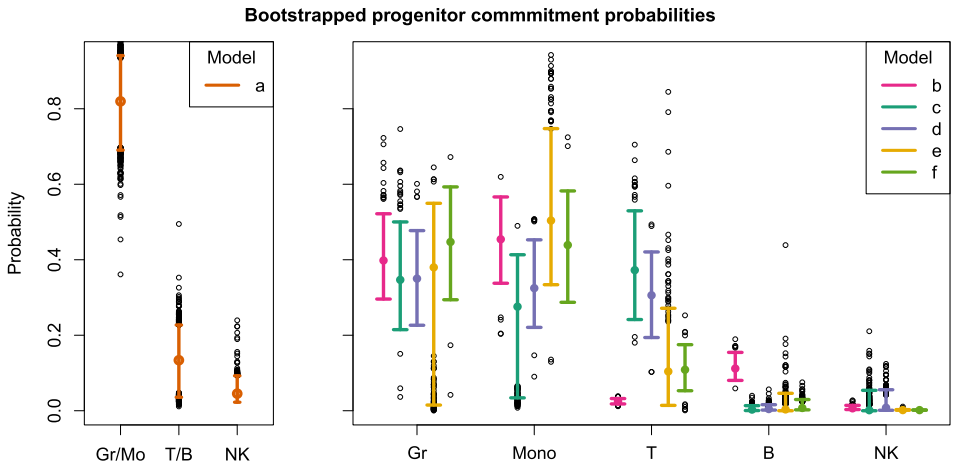

FIG. 7. *Comparison of fitted intermediate differentiation rates parametrized as fate decision probabilities. Displayed are the bootstrap percentile confidence intervals of normalized commitment rates to each mature type* i, $\frac{\hat{\nu}_i}{\sum_j \hat{\nu}_j}$, *in each model displayed in Figure* 2(*a*)–(*f*).

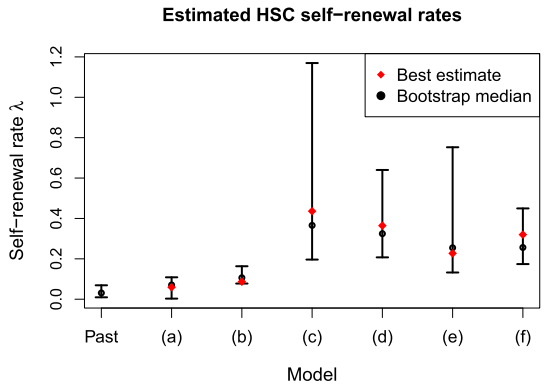

FIG. 8. *Comparison of fitted self-renewal rates* $\hat{\lambda}$ *and* 95% *confidence intervals across all models displayed in Figure* 2(*a*)–(*f*). *Point estimates with lowest objective value* (*best estimates*) *are marked by red diamonds, while bootstrap confidence intervals and medians are plotted 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 close and in reasonable range. Other models without a multipotent progenitor result in less biologically plausible estimates.*

curves in Figure 6) and their corresponding confidence intervals are reported in Appendix A-5. Rate estimates are parametrized as number of events per five days: for instance, 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 types, we fix death rates at $\boldsymbol{\mu} = (0.8, 0.3, 0.04, 0.08, 0.4)$.

These results are easier to visualize in terms of fate decision probabilities, and estimates along with confidence intervals are reported in Figure 7. 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 both variation across stochastic realizations of barcode count time series and from sampling noise. We used bootstrap percentile confidence intervals and report them in Figures 7 and 8. We find that granulocytes and monocytes are produced much more rapidly than T, B and NK cells. Converting back to the original time scale, the estimates indicate that individual progenitor cells are long-lived and can each produce thousands of these mature cells per day. These findings pertaining to the dynamics of intermediate cell stages are biologically realistic and have not been previously estimated. The dynamics of HSCs, on the other hand, have been studied in nonhuman primates via telomere analysis (Shepherd et al. (2007)), which estimates the HSC self-renewal rate at once every 23 weeks with 11–75 week range. This corresponds to an estimate of $\widetilde{\lambda} = 0.0310$ with interval (0.0095, 0.0649) in our parametrization. As we see in Figure 8, these findings coincide with our estimates and confidence intervals for $\hat{\lambda}$ in models with

TABLE 1

Model selection via fivefold cross-validation, where the first row refers to models illustrated in Figure 2. The multipotent progenitor model (b) results in a better objective value evaluated on held-out data than those with multiple oligopotent progenitors

Model	(b)	(c)	(d)	(e)	(f)
Number of progenitors	1	2	2	3	3
CV Objective	4.34	7.49	6.61	9.15	8.25

one multipotent progenitor. The initial barcoding percentage of HSCs is estimated at 13% in model (a), depicted in Figure 2. Since we experienced numerical instabilities while fitting models (b)–(f), we used model (a) estimate and fixed the total progenitor marking percentage at 87% in these more complex models. However, estimates of $\hat{\boldsymbol{\pi}}$ in models with multiple progenitors lie on the boundary of the probability simplex, even when fixing $\hat{\pi}_0$ (see Table A-9, Appendix). Along with higher objective values and less biologically plausible parameters, these results suggest a poorer model fit, reminiscent of the behavior in the model misspecification experiments in Section 3.1. mature types. We quantify this lack of fit by performing model selection via fivefold cross-validation (CV). We divide the dataset into five random subsets of equal size and fit each model to the training data consisting of four of the subsets while holding one subset out as test data to assess predictive performance. We then compute the objective function (2.2) using parameters obtained from the training data and empirical correlations computed using the test data. These cross-validated objective function values for models (b)–(f), displayed in Table 1, are the average of the objectives evaluated across the five sets of training and test data. Model (a), not displayed in the table, achieves a CV objective value of 1.72, but is not directly comparable as its loss function is comprised of fewer correlation terms since there are only three mature blood cell types. The CV objective value for the multipotent progenitor model (b) is noticeably lower, favoring this simple single progenitor model over more complex alternatives. In addition, models with oligopotent progenitors (c)–(f) visually fit the data worse than the multipotent progenitor model (b) when fitted and empirical correlations are plotted together (see Appendix A-5). We also considered additional oligopotent progenitor models (e.g., switching B and T in models (d) and (f) and switching NK and T as well as NK and B in (e)), but these additional models also performed worse that the multipotent progenitor model (b), so we do not report these results here.

4. Discussion. We propose a novel modeling framework and parameter estimation procedure for analyzing hematopoietic lineage tracking experiments. To our knowledge, this is the first such method for fitting time series counts from cell

lineage tracking data to continuous-time stochastic models featuring HSC, progenitor and mature stages of cell development. Detailed simulation studies show that the loss function estimator yields accurate inference when applied to data generated from this class of models. Our analysis of *in vivo* experimental data yields estimates of HSC self-renewal rates, intermediate cell differentiation rates, and progenitor death rates. We are the first to estimate most of these parameters in a large primate system. Moreover, our methodology opens the door for statistically rigorous selection of models describing the hierarchy of hematopoietic cell specialization and differentiation.

Our exploration of several models suggests that a model with one unrestricted multipotent progenitor provides a better fit to the data than models requiring an ordered hierarchical differentiation. This result may seem counterintuitive, but one needs to remember that even though our models with multiple progenitors are more complex, they are also more restrictive in the sense that they include loss of lineage potential by limiting the types of mature cells that can descend from each distinct progenitor. If in reality progenitors never fully lose their potential to produce all mature cell types, this restriction leads to model misspecification. Indeed, recent studies dispute traditional assumptions about hematopoietic structures prescribing restricted differentiation pathways. For instance, Kawamoto, Wada and Katsura (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. In light of this and other recent studies, our model selection results are supportive of emerging experimental data (Velten et al. (2017)).

Several limitations remain when modeling hematopoiesis as a Markov branching process. The assumptions of linearity and rate homogeneity imply a possibility of unlimited growth, and extending analysis to allow for nonlinear regulatory behavior as the system grows near a carrying capacity is merited. Similarly, the Markov assumption may 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: for instance, it is known some cells in the peripheral bloodstream move in and out of tissue. While such extensions are mathematically challenging, they are straightforward to simulate, 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 viable alternative. Indeed, a Bayesian framework would further allow existing prior information available from previous studies about average lifespans of mature blood cells to be incorporated without fixing some of the model parameters.

Our fully generative framework and accompanying estimator immediately 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 differentiation 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. While we perform model selection by loss function cross-validation, 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, a well studied and active area of research (Liepe et al. (2014), Toni et al. (2009), Pudlo et al. (2016)), is also applicable to our modeling framework.

Finally, the class of models we consider and derivations for their moment expressions are general in that an arbitrary number of intermediate progenitors and mature cell types can be specified. Nonetheless, these models 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 assumptions only allow for each mature cell to be descended from one progenitor type, which limits the ability to investigate fully connected and nested models. Nonetheless, we have enabled parameter estimation in much more detailed models than previous statistical studies, while accounting for missing information and experimental noise. These 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.

Acknowledgments. We thank Jon Wellner and Jon Wakefield for helpful discussions about GMM and M-estimation, and the anonymous reviewers and Editor for their constructive comments and insightful suggestions.

SUPPLEMENTARY MATERIAL

Appendix (DOI: 10.1214/19-AOAS1272SUPP; .pdf). Additional tables, figures, and details are included in the Appendix included in the Supplementary Material.

REFERENCES

ABKOWITZ, J. L., LINENBERGER, M. L., NEWTON, M. A., SHELTON, G. H., OTT, R. L. and GUTTORP, P. (1990). Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. *Proc. Natl. Acad. Sci. USA* **87** 9062–9066.

BAILEY, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley, New York. MR0165572

BECKER, A. J., MCCULLOCH, E. A. and TILL, J. E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197** 452–454.

BIASCO, L., PELLIN, D., SCALA, S., DIONISIO, F., BASSO-RICCI, L., LEONARDELLI, L., SCARAMUZZA, S., BARICORDI, C., FERRUA, F. et al. (2016). In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19** 107–119.

BUCHHOLZ, V. R., FLOSSDORF, M., HENSEL, I., KRETSCHMER, L., WEISSBRICH, B., GRÄF, P., VERSCHOOR, A., SCHIEMANN, M., HÖFER, T. et al. (2013). Disparate individual fates compose robust CD8+ T cell immunity. *Science* **340** 630–635.

CATLIN, S. N., ABKOWITZ, J. L. and GUTTORP, P. (2001). Statistical inference in a two-compartment model for hematopoiesis. *Biometrics* **57** 546–553. MR1855690

CATLIN, S. N., BUSQUE, L., GALE, R. E., GUTTORP, P. and ABKOWITZ, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood* **117** 4460–4466.

COLIJN, C. and MACKEY, M. C. (2005). A mathematical model of hematopoiesis. I. Periodic chronic myelogenous leukemia. *J. Theoret. Biol.* **237** 117–132. MR2205754

DORMAN, K. S., SINSHEIMER, J. S. and LANGE, K. (2004). In the garden of branching processes. *SIAM Rev.* **46** 202–229. MR2114452

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FONG, Y., GUTTORP, P. and ABKOWITZ, J. (2009). Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. *Ann. Appl. Stat.* **3** 1695–1709. MR2752154

GERRITS, A., DYKSTRA, B., KALMYKOWA, O. J., KLAUKE, K., VEROVSKAYA, E., BROEKHUIS, M. J. C., DE HAAN, G. and BYSTRYKH, L. V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115** 2610–2618.

GOLINELLI, D., GUTTORP, P. and ABKOWITZ, J. A. (2006). Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. *Math. Med. Biol.* **23** 153–172.

GOYAL, S., KIM, S., CHEN, I. S. Y. and CHOU, T. (2015). Mechanisms of blood homeostasis: Lineage tracking and a neutral model of cell populations in rhesus macaques. *BMC Biol.* **13** 85.

GRIFFIN, J. E. and BROWN, P. J. (2013). Some priors for sparse regression modelling. *Bayesian Anal.* **8** 691–702. MR3102230

GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data. Stochastic Modeling Series*. CRC Press, London. MR1358359

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.

HELLERSTEIN, M., HANLEY, M. B., CESAR, D., SILER, S., PAPAGEORGOPOULOS, C., WIEDER, E., SCHMIDT, D., HOH, R., NEESE, R. et al. (1999). Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. *Nat. Med.* **5** 83–89.

HYRIEN, O., PESLAK, S. A., YANEV, N. M. and PALIS, J. (2015). Stochastic modeling of stress erythropoiesis using a two-type age-dependent branching process with immigration. *J. Math. Biol.* **70** 1485–1521. MR3343931

KAUR, A., DI MASCIO, M., BARABASZ, A., ROSENZWEIG, M., MCCLURE, H. M., PERELSON, A. S., RIBEIRO, R. M. and JOHNSON, R. P. (2008). Dynamics of T-and B-lymphocyte turnover in a natural host of simian immunodeficiency virus. *J. Virol.* **82** 1084–1093.

KAWAMOTO, H., WADA, H. and KATSURA, Y. (2010). A revised scheme for developmental pathways of hematopoietic cells: The myeloid-based model. *Int. Immunol.* **22** 65–70.

KIMMEL, M. (2014). Stochasticity and determinism in models of hematopoiesis. In *A Systems Biology Approach to Blood* 119–152. Springer, New York.

KIMMEL, M. and AXELROD, D. E. (2002). *Branching Processes in Biology*. *Interdisciplinary Applied Mathematics* **19**. Springer, New York. MR1903571

KOELLE, S. J., ESPINOZA, D. A., WU, C., XU, J., LU, R., LI, B., DONAHUE, R. E. and DUNBAR, C. E. (2017). Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* **129** 1448–1457.

LANGE, K. (2010). *Applied Probability*. Springer, New York.

LASLO, P., PONGUBALA, J. M. R., LANCKI, D. W. and SINGH, H. (2008). Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin. Immunol.* **20** 228–235.

LIEPE, J., KIRK, P., FILIPPI, S., TONI, T., BARNES, C. P. and STUMPF, M. P. H. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* **9** 439–456.

LU, R., NEFF, N. F., QUAKE, S. R. and WEISSMAN, I. L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29** 928–933.

MARCINIAK-CZOCHRA, A., STIEHL, T., HO, A. D., JÄGER, W. and WAGNER, W. (2009). Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. *Stem Cells Dev.* **18** 377–386.

MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.

NOTTA, F., ZANDI, S., TAKAYAMA, N., DOBSON, S., GAN, O. I., WILSON, G., KAUFMANN, K. B., MCLEOD, J., LAURENTI, E. et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* aab2116.

ORKIN, S. H. and ZON, L. I. (2008). Hematopoiesis: An evolving paradigm for stem cell biology. *Cell* **132** 631–644.

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057. MR1014540

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

PERIÉ, L., HODGKIN, P., NAIK, S. H., SCHUMACHER, T. N., DE BOER, R. J. and DUFFY, K. R. (2014). Determining lineage pathways from cellular barcoding experiments. *Cell Rep.* **6** 617–624.

PUDLO, P., MARIN, J. M., ESTOUP, A., CORNUET, J. M., GAUTIER, M. and ROBERT, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics* **32** 859–866.

SHEPHERD, B. E., KIEM, H. P., LANSDORP, P. M., DUNBAR, C. E., AUBERT, G., LAROCHELLE, A., SEGGEWISS, R., GUTTORP, P. and ABKOWITZ, J. L. (2007). Hematopoietic stem-cell behavior in nonhuman primates. *Blood* **110** 1806–1813.

SIMINOVITCH, L., MCCULLOCH, E. A. and TILL, J. E. (1963). The distribution of colony-forming cells among spleen colonies. *J. Cell. Comp. Physiol.* **62** 327–336.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TILL, J. E., MCCULLOCH, E. A. and SIMINOVITCH, L. (1964). A stochastic model of stem cell prolifiration, based on the growth of spleen colony-forming cells. *Proc. Natl. Acad. Sci. USA* **51** 29–36.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VELTEN, L., HAAS, S. F., RAFFEL, S., BLASZKIEWICZ, S., ISLAM, S., HENNIG, B. P., HIRCHE, C., LUTZ, C., BUSS, E. C. et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19** 271.

WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods*. *Springer Series in Statistics*. Springer, New York. MR3025225

WEISSMAN, I. L. (2000). Stem cells: Units of development, units of regeneration, and units in evolution. *Cell* **100** 157–168.

WHICHARD, Z. L., SARKAR, C. A., KIMMEL, M. and COREY, S. J. (2010). Hematopoiesis and its disorders: A systems biology approach. *Blood* **115** 2339–2347.

WU, C., LI, B., LU, R., KOELLE, S. J., YANG, Y., JARES, A., KROUSE, A. E., METZGER, M., LIANG, F. et al. (2014). Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14** 486–499.

XU, J., KOELLE, S., GUTTORP, P., WU, C., DUNBAR, C., ABKOWITZ, J. L. and MININ, V. N. (2019). Supplement to "Statistical inference for partially observed branching processes with application to cell lineage tracking of *in vivo* hematopoiesis." DOI:10.1214/19-AOAS1272SUPP.

ZHANG, Y., WALLACE, D. L., DE LARA, C. M., GHATTAS, H., ASQUITH, B., WORTH, A., GRIFFIN, G. E., TAYLOR, G. P., TOUGH, D. F. et al. (2007). In vivo kinetics of human natural killer cells: The effects of ageing and acute and chronic viral infection. *Immunology* **121** 258–265.

J. XU
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
214 OLD CHEMISTRY BUILDING, BOX 90251
DURHAM, NORTH CAROLINA 27705
USA
E-MAIL: jason.q.xu@duke.edu
URL: https://stat.duke.edu/people/jason-qian-xu

S. KOELLE
P. GUTTORP
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
4060 E STEVENS WAY NE
SEATTLE, WASHINGTON 98195
USA
E-MAIL: sjkoelle@uw.edu
guttorp@uw.edu

C. WU
C. DUNBAR
NATIONAL HEART, LUNG,
AND BLOOD INSTITUTE
NATIONAL INSTITUTES OF HEALTH
10 CENTER DRIVE
BETHESDA, MARYLAND 20814
USA
E-MAIL: chuanfeng.wu@nih.gov
dunbarc@nhlbi.nih.gov

J. L. ABKOWITZ
DIVISION OF HEMATOLOGY
UNIVERSITY OF WASHINGTON
SCHOOL OF MEDICINE
825 EASTLAKE AVE E
SEATTLE, WASHINGTON 98109
USA
E-MAIL: janabk@u.washington.edu

V. N. MININ
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, IRVINE
BREN HALL 2019
IRVINE, CALIFORNIA 92697
USA
E-MAIL: vminin@uci.eu