Stochastic Compartmental Modeling and Inference with Biological Applications

Jason Xu
Department of Biomathematics, UCLA

Challenges in the Statistical Modeling of Stochastic Processes for the Natural Sciences Workshop

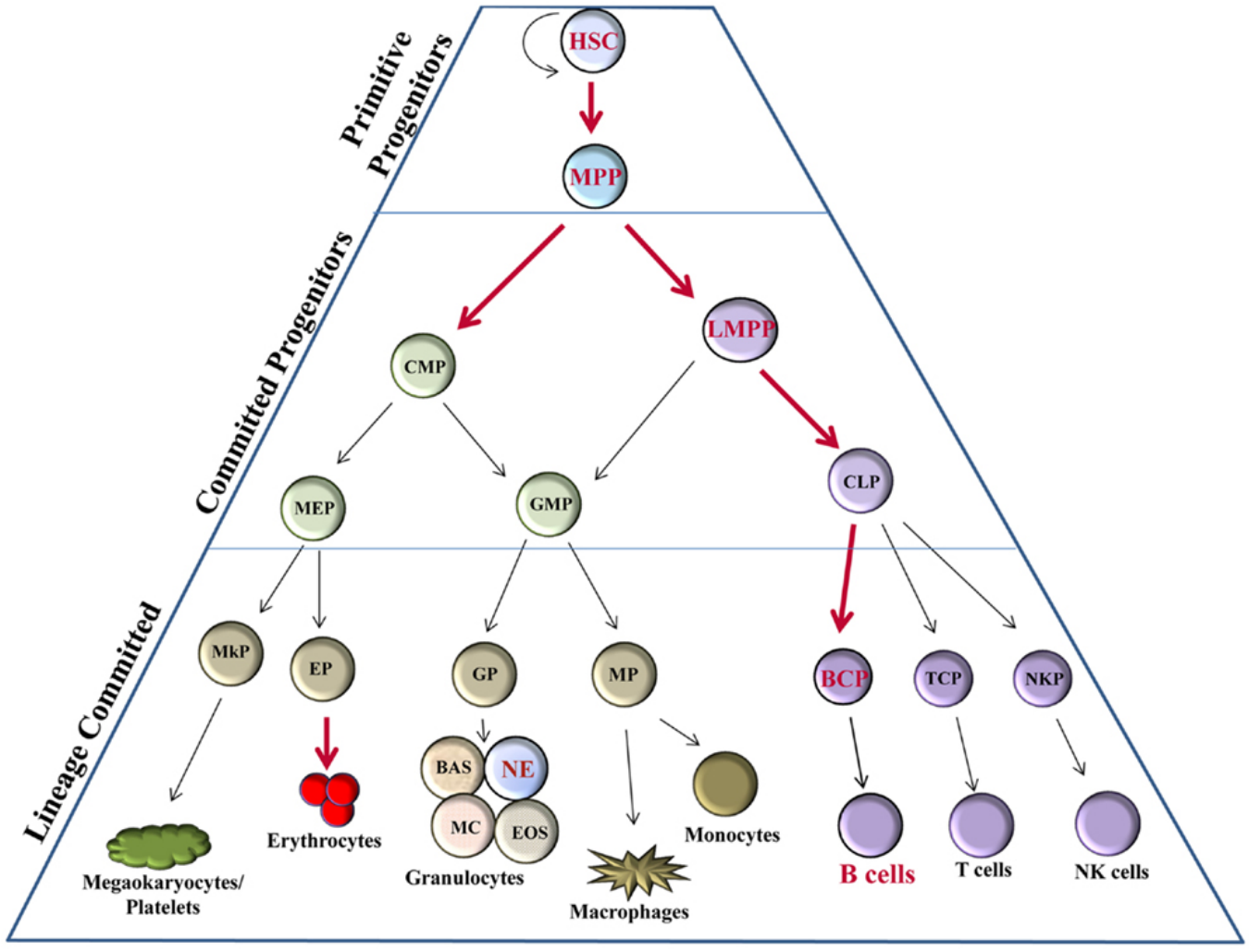
Banff International Research Station, July 11, 2017

Review: hematopoiesis

A complex mechanism in which self-renewing hematopoietic stem cells (HSCs) differentiate via a series of intermediate progenitor cell stages to produce blood cells

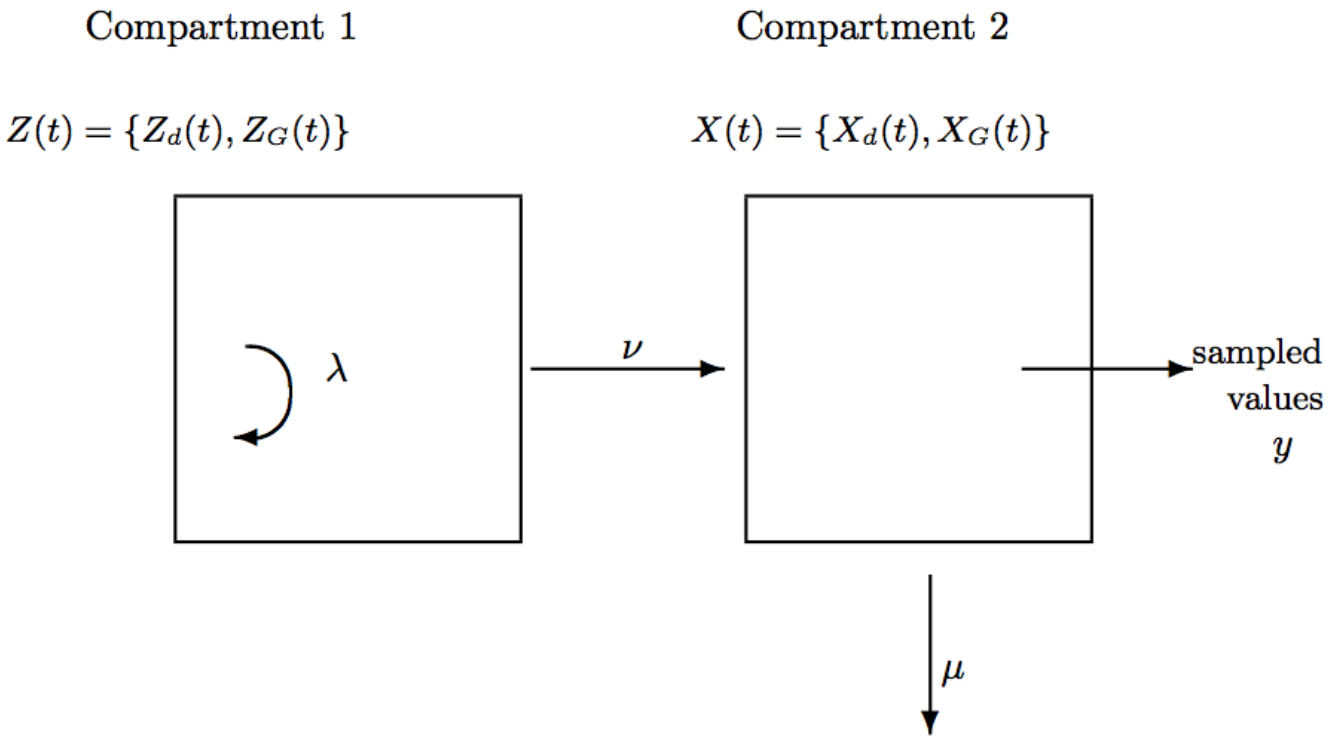
- Multi-stage process in bone marrow, lymphatic and circulatory systems: difficult to observe *in vivo*
- Dynamics and structure are largely unknown
- Clinical relevance: stem cell transplantation is a mainstay of cancer therapy; all blood cell diseases are caused by malfunctions in the hematopoietic process
- Stochastic modeling efforts provide quantitative basis to answer questions about dynamics (parameter inference) and structure (model selection)

Branching structure of hematopoiesis



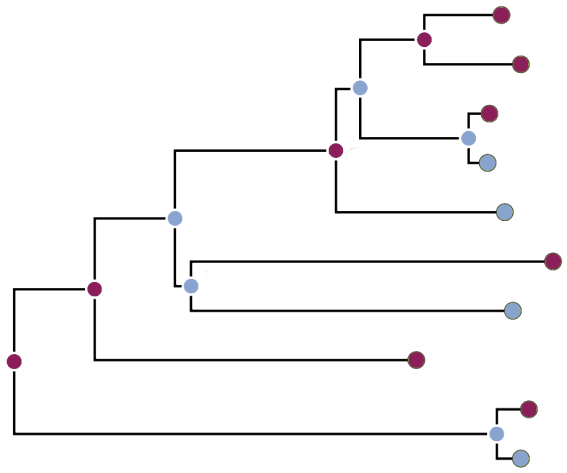
Many versions of the hematopoietic tree have been proposed

Two-type compartmental model



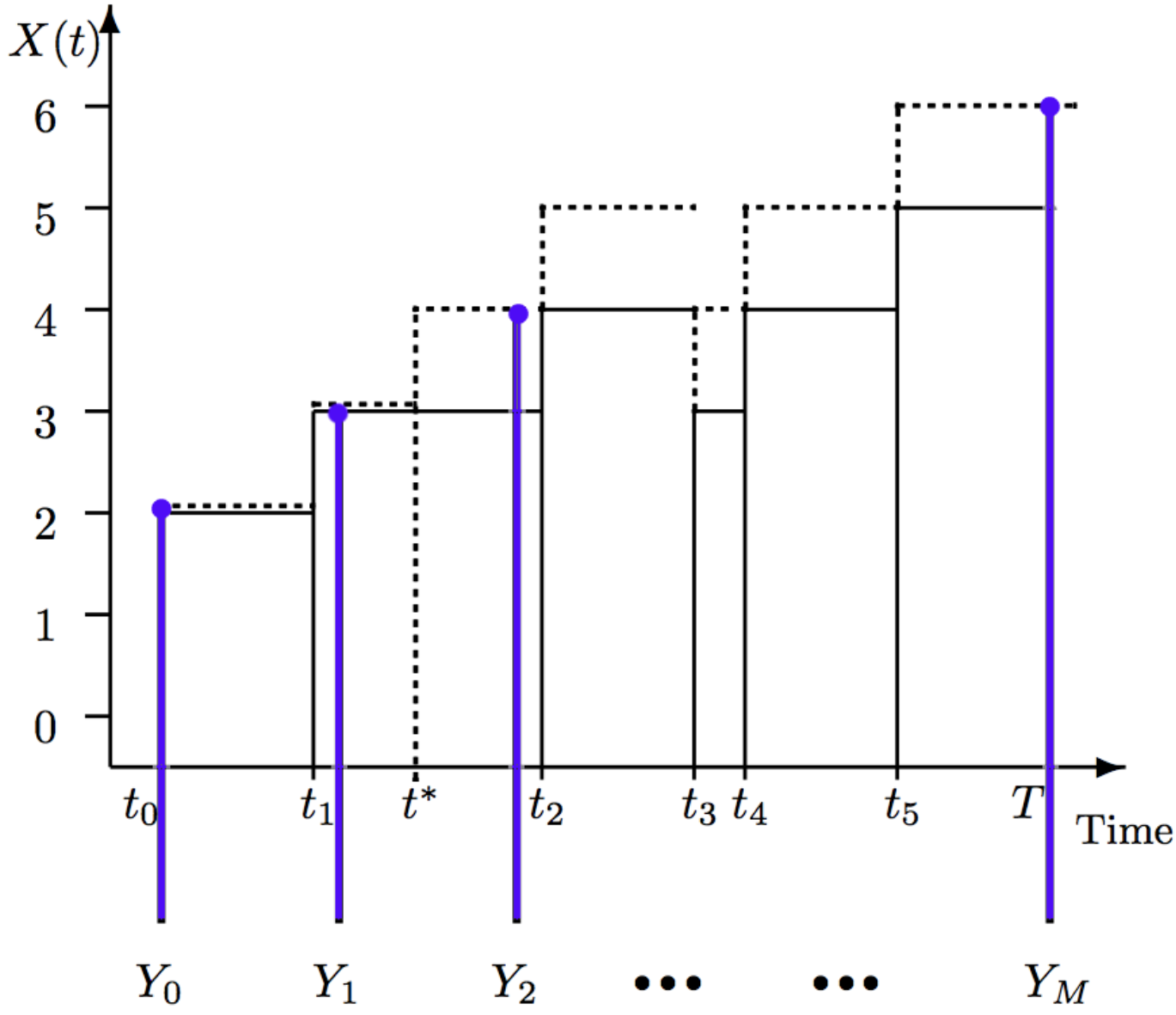
- Series of statistical studies targeting HSC dynamics [Abkowitz et al 1990, Golinelli et al 2009, Catlin et al 2011]
- Cannot resolve questions about later stages of differentiation
- Past studies: intensive simulation study, estimating equations, reversible-jump MCMC
- Can be equivalently treated as a Markov branching process

Multi-type Markov branching processes



- Random vector $\mathbf{X}(t)$; $X_i(t)$ denotes type i population at time t
- Cells act independently: can die, reproduce, create other cells
- Independence $\Rightarrow$ linearity: overall rates are multiplicative in number of cells
- Time-homogeneity: jump rates are constant over time
- A class of continuous-time Markov chains (CTMCs): memoryless, exponential times between events

Challenges: discretely observed data



The discretely-observed data likelihood

$$\ell_o(\mathbf{Y}|\boldsymbol{\theta}) = \sum_{p=1}^{m} \sum_{i=0}^{n(p)-1} \log p_{\mathbf{X}^p(t_{p,i}),\mathbf{X}^p(t_{p,i+1})}(t_{p,i+1} - t_{p,i}|\boldsymbol{\theta})$$

In particular, need finite-time transition probabilities:

$$p_{\mathbf{x},\mathbf{y}}(t) = \Pr\left(\mathbf{X}(t+s) = \mathbf{y}|\mathbf{X}(s) = \mathbf{x}\right)$$

- Classical matrix exponentiation for CTMCs is $\mathcal{O}(|\Omega|^3)$

$$\mathbf{P}(t) := \left\{p_{\mathbf{x},\mathbf{y}}(t)\right\}_{\mathbf{x},\mathbf{y}\in\Omega} = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{(\mathbf{Q}t)^k}{k!}.$$

- When only partially observed (latent process), compounded by additional marginalization over hidden states

Using the probability generating function ϕ

$$\phi_{jk}(t, s_1, s_2; \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left(s_1^{X_1(t)} s_2^{X_2(t)} | X_1(0) = j, X_2(0) = k\right)$$
$$= \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{(jk),(lm)}(t; \boldsymbol{\theta}) s_1^l s_2^m; \qquad |s_i| \leq 1$$

- PGF ϕ_{jk} computed by solving Kolmogorov forward/backward ODEs
- Transition probabilities related via differentiation, but impractical

$$p_{(jk),(lm)}(t) = \frac{1}{l!m!}\frac{\partial^l}{\partial s_1}\frac{\partial^m}{\partial s_2}\phi_{jk}(t)\Big|_{s_1=s_2=0}$$

- Transform $s_1 = e^{2\pi i w_1}, s_2 = e^{2\pi i w_2} \Rightarrow \phi$ becomes a Fourier series:

$$\phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{(jk),(lm)}(t) e^{2\pi i l w_1} e^{2\pi i m w_2}$$

From differentiation to integration: a spectral trick

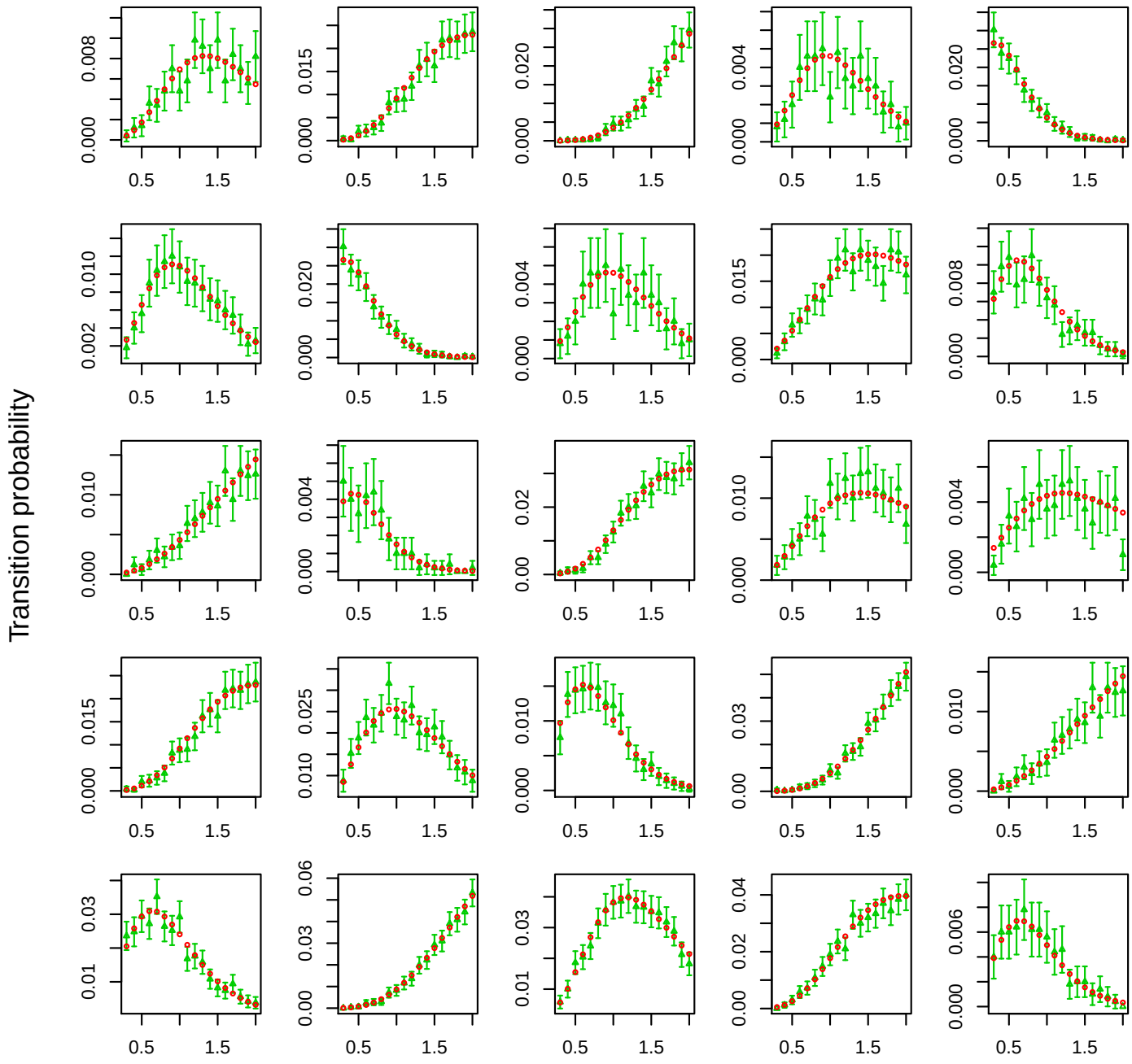
- Inverting the Fourier series representation recovers transition probabilities efficiently:

$$p_{(jk),(lm)}(t) = \int_0^1 \int_0^1 \phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) e^{-2\pi i l w_1} e^{-2\pi i m w_2} dw_1 dw_2$$

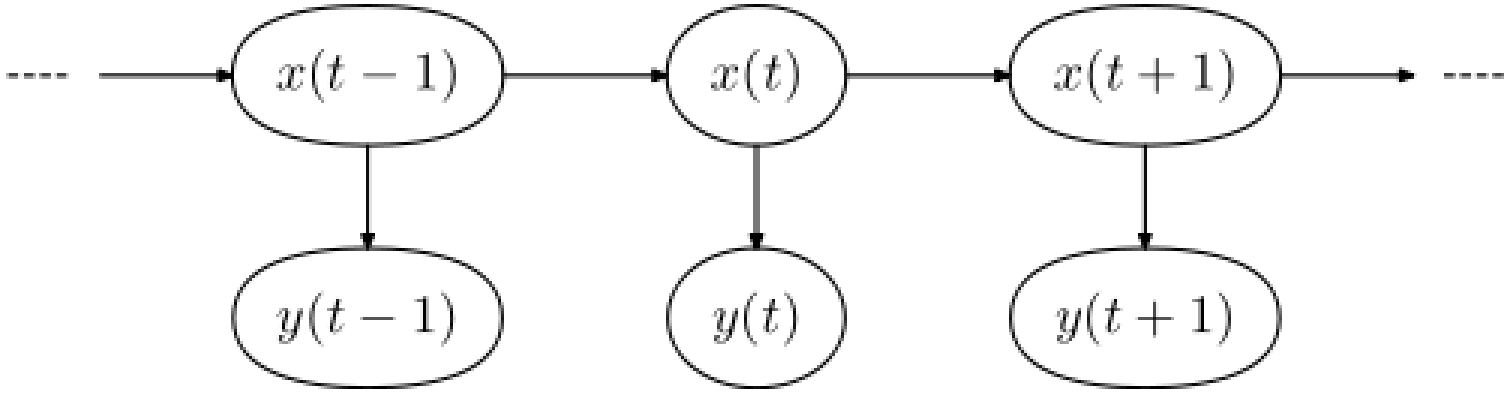
(applying a Riemann sum approximation)

$$\approx \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) e^{-2\pi i l u/N} e^{-2\pi i m v/N}.$$

- Can *simultaneously* compute probabilities $\{p_{(jk),(lm)}(t)\}$ for all $l, m = 0, \ldots, N$ via Fast Fourier Transform (FFT)
- Compute discrete-data likelihood practically; similar approach yields conditioned moments useful for EM
  [Xu, Guttorp, Kato-Maeda, Minin 2015]


Transition probability
Shift Rate ν
0.5
1.5
0.000
0.008
0.015
0.020
0.004
0.002
0.010
0.00
0.02
0.025
0.03
0.01
0.06
0.04
0.006

More missing data: partially observed processes



Process $\mathbf{X}(t)$ poses same challenges as before, but now we only glimpse partial information (sampling, measurement error)

- Direct marginalization impractical for large Ω
- Data augmented MCMC: slow mixing, need efficient proposals
- Simulation approaches (particle filtering, SMC) are flexible, but quickly become limited for large populations [Andrieu et al 2010]

Hematopoietic lineage barcoding data

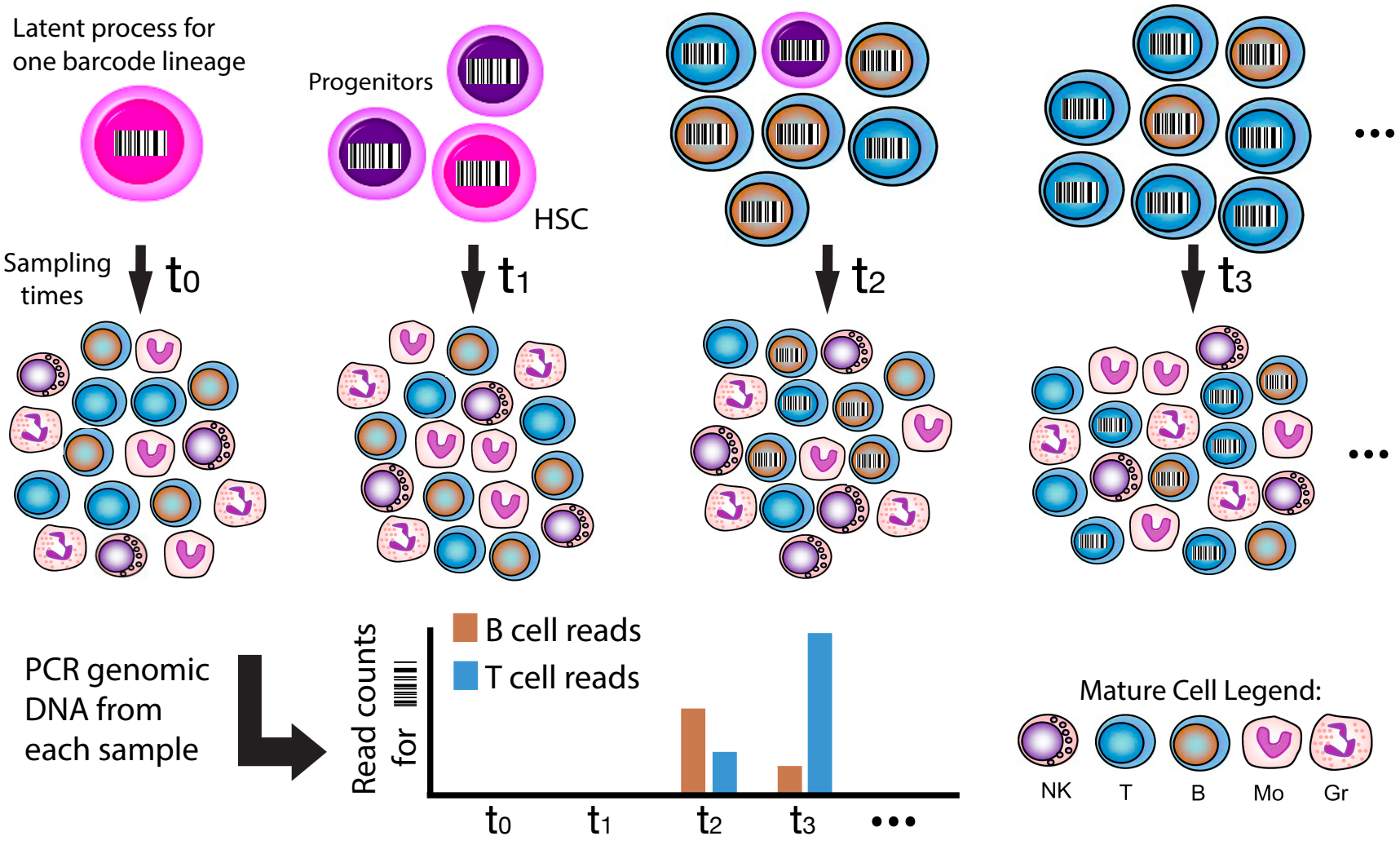
Cell Stem Cell
Article

Clonal Tracking of Rhesus Macaque Hematopoiesis Highlights a Distinct Lineage Origin for Natural Killer Cells

Chuanfeng Wu,[1,7] Brian Li,[1,7] Rong Lu,[2,7] Samson J. Koelle,[1] Yanqin Yang,[3] Alexander Jares,[1] Alan E. Krouse,[1] Mark Metzger,[1] Frank Liang,[4] Karin Loré,[4] Colin O. Wu,[5] Robert E. Donahue,[1] Irvin S.Y. Chen,[6] Irving Weissman,[2] and Cynthia E. Dunbar[1,*]
[1]Hematology Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA

- DNA barcoding experiments enable individual cell lineage tracking through time, *in vivo*
- **IID** time series data: DNA read counts partially inform the populations of each barcode ID present among each cell type
- Monitored at discrete times over 30 months, dataset contains 110 million read counts across 9635 unique barcode IDs
- Discrete hidden space of multiple very large, hidden populations

Illustration: experimental design



The hidden branching process model

- Latent process: each barcode lineage evolves as a continuous-time, multitype branching process $\mathbf{X}(t)$ whose components are counts of each cell type
- Observation process: flexible choice of emission distribution for sample counts: we use multivariate hypergeometric distribution $\widetilde{\mathbf{Y}} \sim \text{mvhypgeo}(\mathbf{X})$
- Read data: read counts $\mathbf{Y}$ are proportional to $\widetilde{\mathbf{Y}}$ with unknown amplification constant

Moment-based method of inference

Loss function estimation: match pairwise model-based and empirical correlations across barcode lineages [Xu et al 2017],

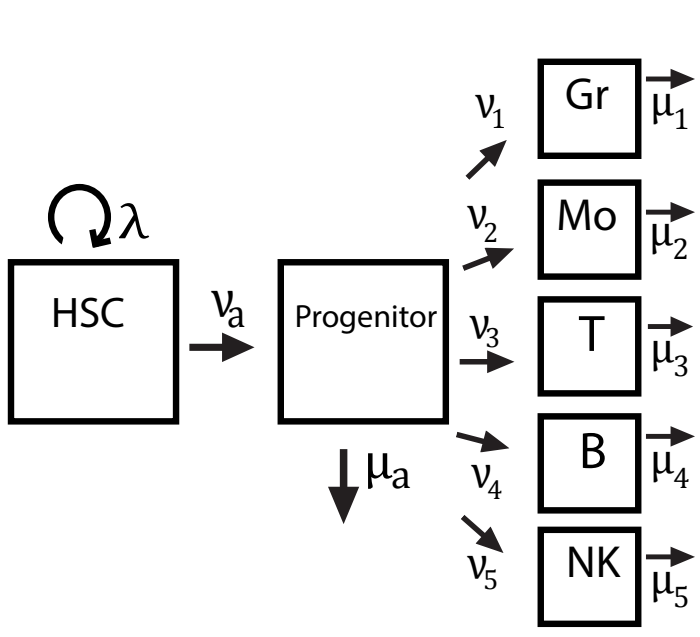
$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{t_j} \sum_{m} \sum_{n \neq m} \left[\psi^j_{mn}(\boldsymbol{\theta}) - \hat{\psi}^j_{mn}(\mathbf{Y}) \right]^2,$$

$$\psi^j_{mn}(\boldsymbol{\theta}) = \left[\rho(Y_m(t_j), Y_n(t_j)); \boldsymbol{\theta} \right], \qquad \text{and}$$

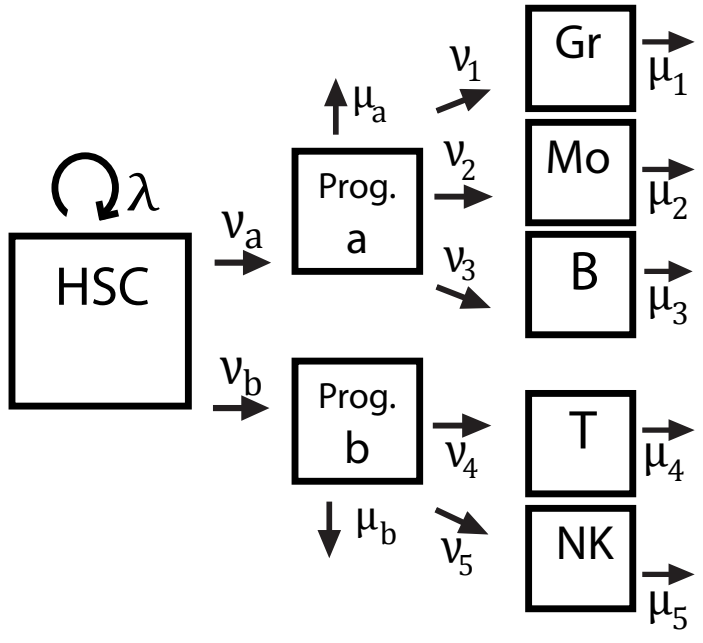
$\hat{\psi}^j_{mn}$ denotes the corresponding sample correlations at time t_j

- Estimating parameters θ reduces to **nonlinear least squares optimization**: $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})$
- Consistent under mild assumptions: $\{\hat{\boldsymbol{\theta}}_N\} \to \boldsymbol{\theta}_0$ in probability

A richer class of compartmental models



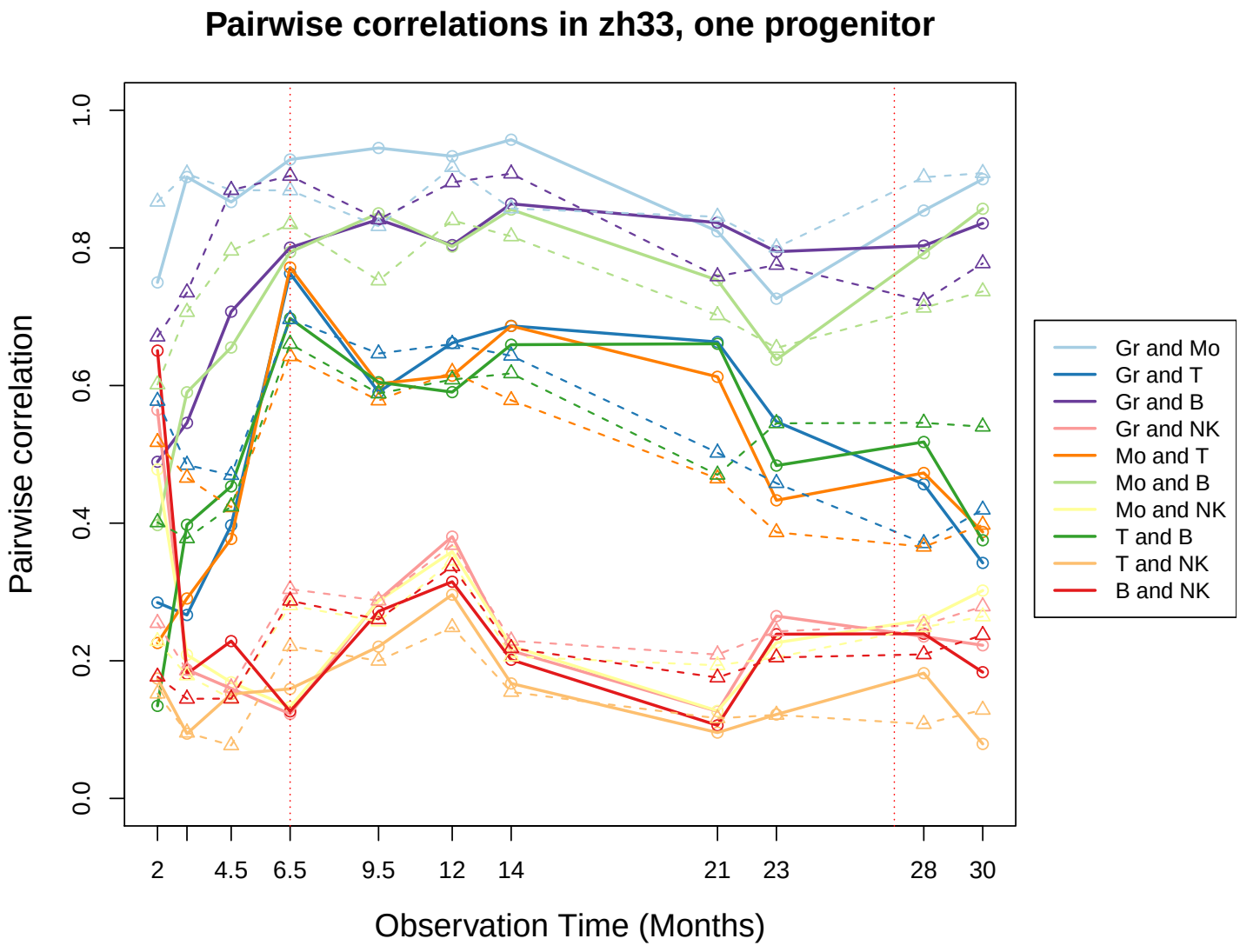
(a) Multipotent progenitor



(b) Oligopotent progenitors

Allows for an arbitrary number of intermediate progenitors and mature cell types, requiring that each mature type can be descended from only one possible progenitor type

Fitted correlations: macaque lineage tracking data

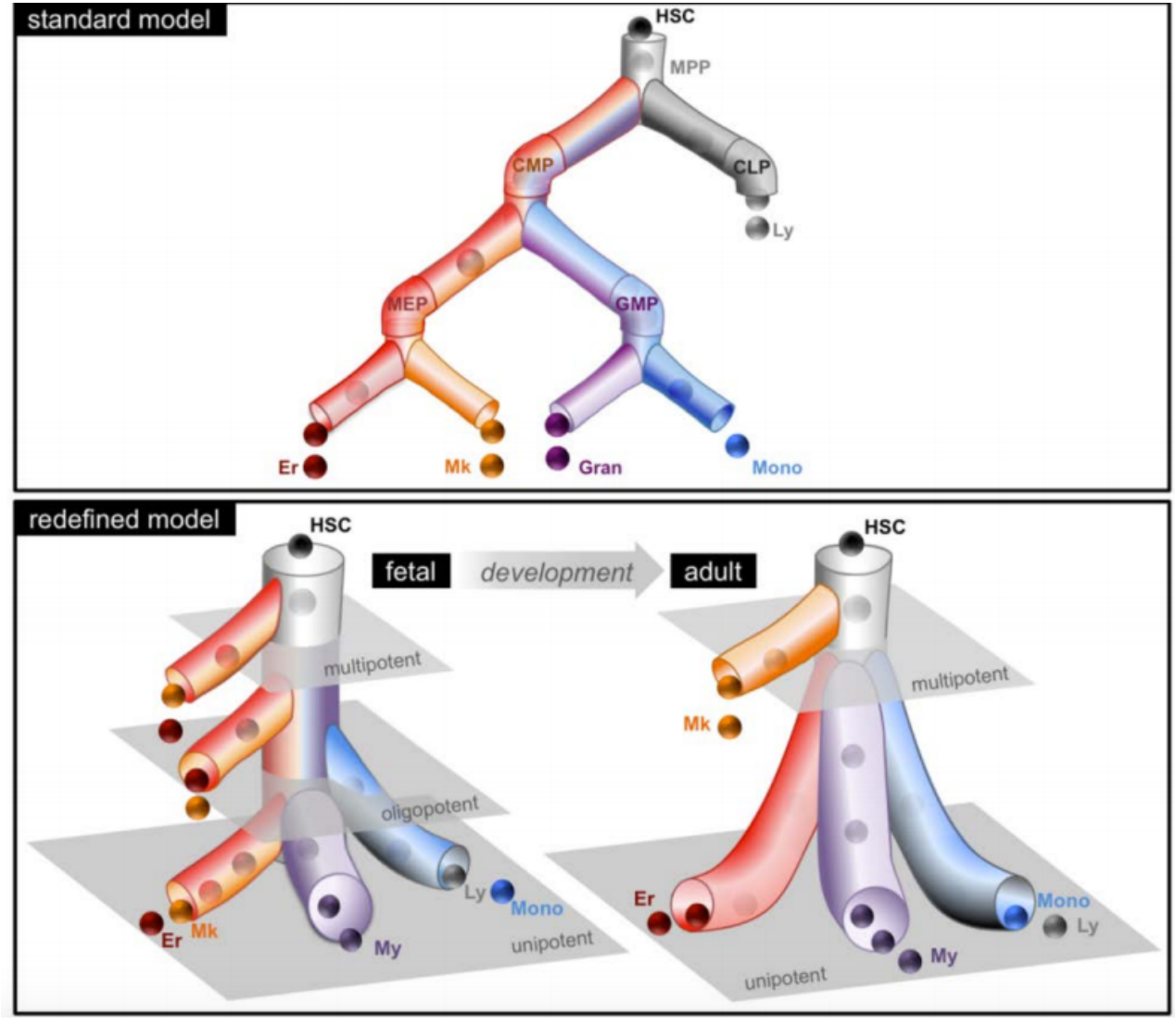


Solid lines denote empirical correlation profiles; dotted lines denote model-based correlations from best fitting estimates $\hat{\theta}$

Overview of results

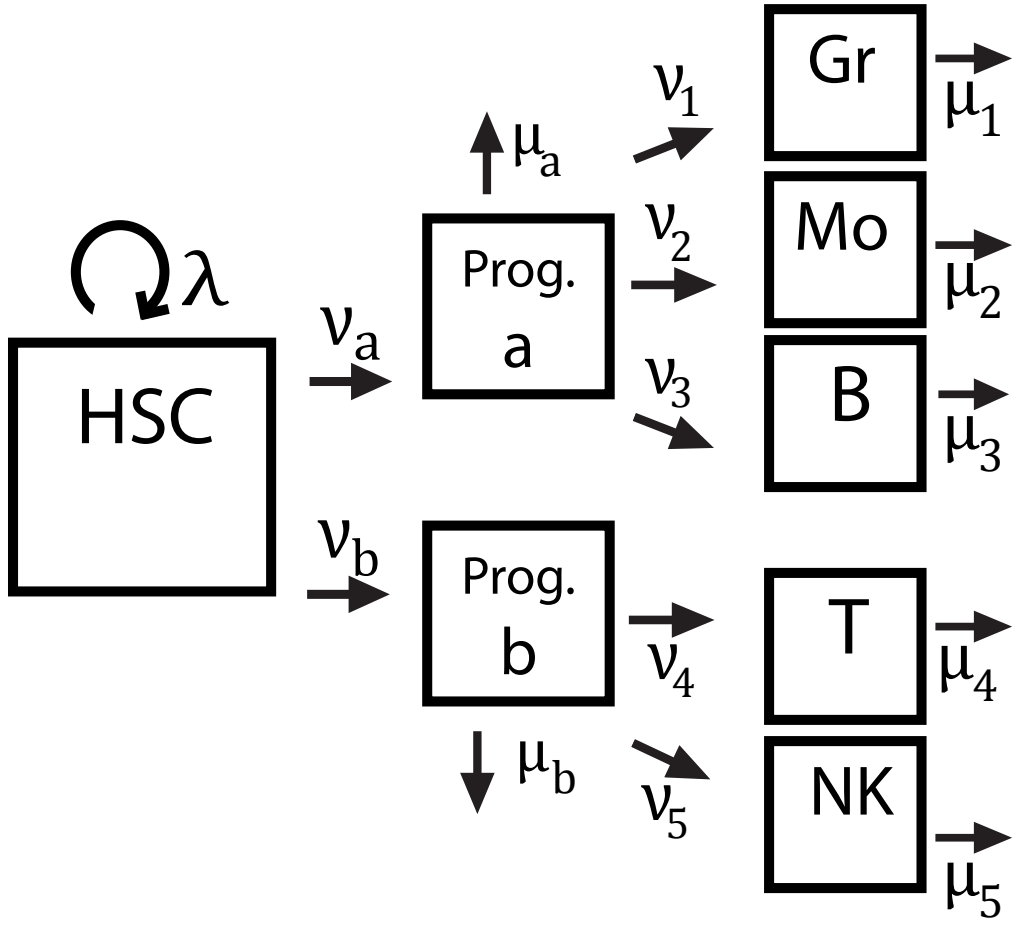
- HSC self-renewal rate $\hat{\lambda} = .0593$ (every 12 weeks) falls into the confidence interval $(0.0095, 0.0649)$ obtained in previous primate studies
- Initial distribution $\hat{\pi} = .139$ consistent with GFP marking levels stabilizing at 13%
- Intermediate rates ν_i suggest granulocytes and monocytes are produced much more rapidly than T, B and NK cells, and individual progenitors can each produce thousands of cells daily (not previously estimated)
- NK cells track distinctly from other mature blood cells
- Single-progenitor models fit best, affirming recent findings [Notta 2015] of *in vitro* human hematopoiesis that challenge the traditional oligopotency assumption

Evidence against oligopotency



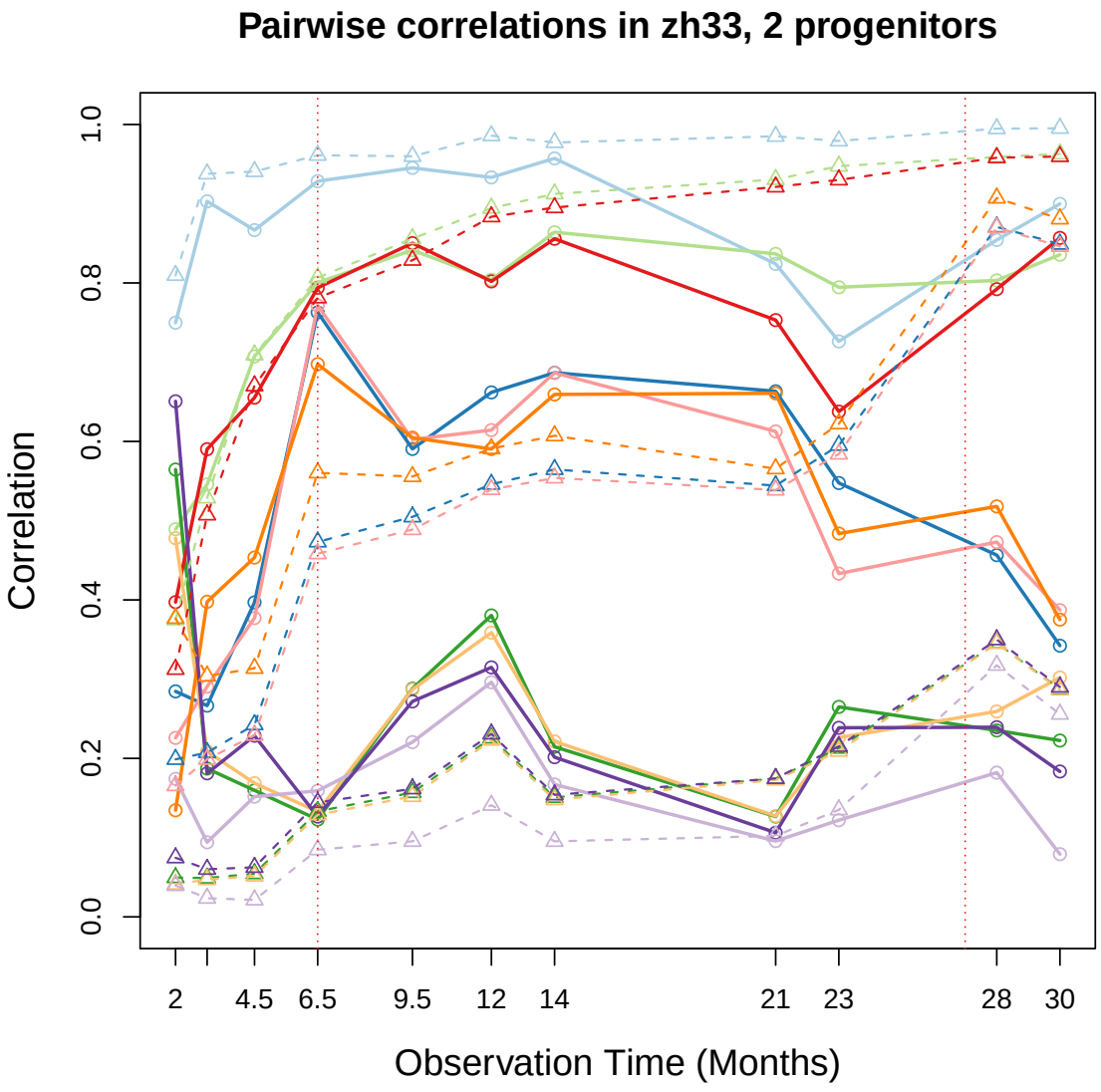
Notta et al. 2015

An open challenge: model selection



A model with two oligopotent progenitors instead of one common multipoint progenitor leads to poor model fit

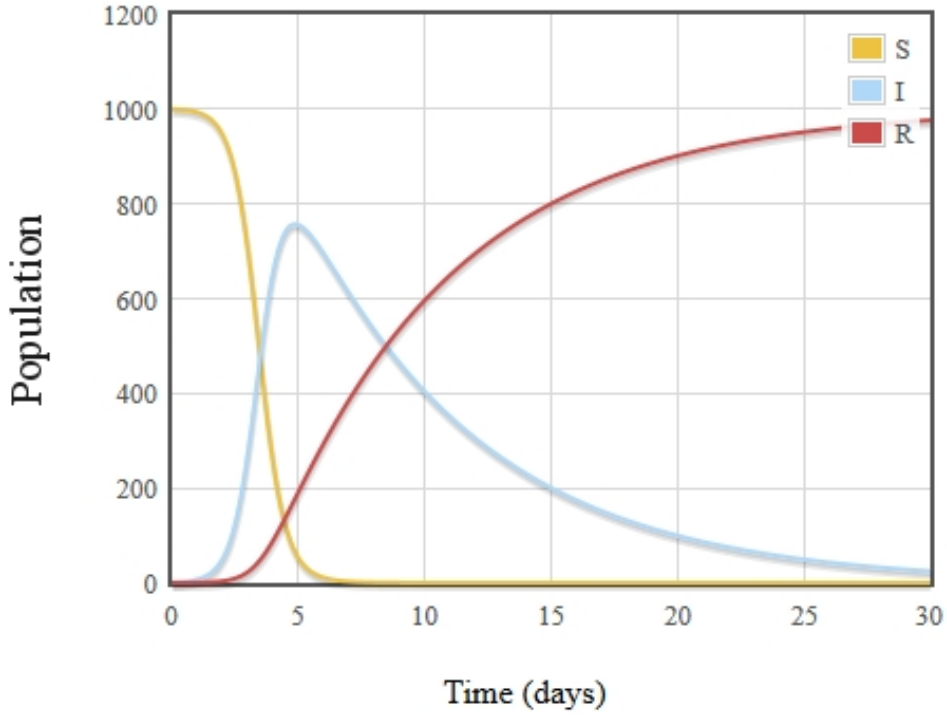
An open challenge: model selection



- Efficient parameter estimation for fitting cell lineage barcoding data to rich models of hematopoiesis
- Need model selection techniques for rigorous conclusions about pathway structure

Nonlinear compartmental models

Motivating example: stochastic SIR model of infection



"*The associated mathematical manipulations required to generate solutions can only be described as heroic.*"

— E. Renshaw, 2015,
Stochastic population processes: analysis, approximations, simulations.

SIR dynamics

$$\Pr(S(t+h) = x_h, I(t+h) = y_h | S(t) = x_t, I(t) = y_t)$$

$$= \begin{cases} \beta x_t y_t h + o(h) & \text{if } (x_h, y_h) = (x_t - 1, y_t + 1) \\ \gamma y_t h + o(h) & \text{if } (x_h, y_h) = (x_t, y_t - 1) \\ 1 - (\beta x_t y_t + \gamma y_t)h + o(h) & \text{if } (x_h, y_h) = (x_t, y_t) \end{cases}$$

- Parameters: infection rate β, recovery rate γ
- Nonlinearity arises from interactions: does not satisfy particle independence $\Rightarrow$ cannot analyze as branching process
- Finite-time behavior (transition probabilities) challenging

Transition probabilities: a different route

Very briefly, working in the Laplace domain,

$$\phi_{ab}(s) := \mathcal{L}[P_{ab}^{a_0 b_0}(t)](s) = \int_0^\infty e^{-st} P_{ab}^{a_0 b_0}(t) dt$$

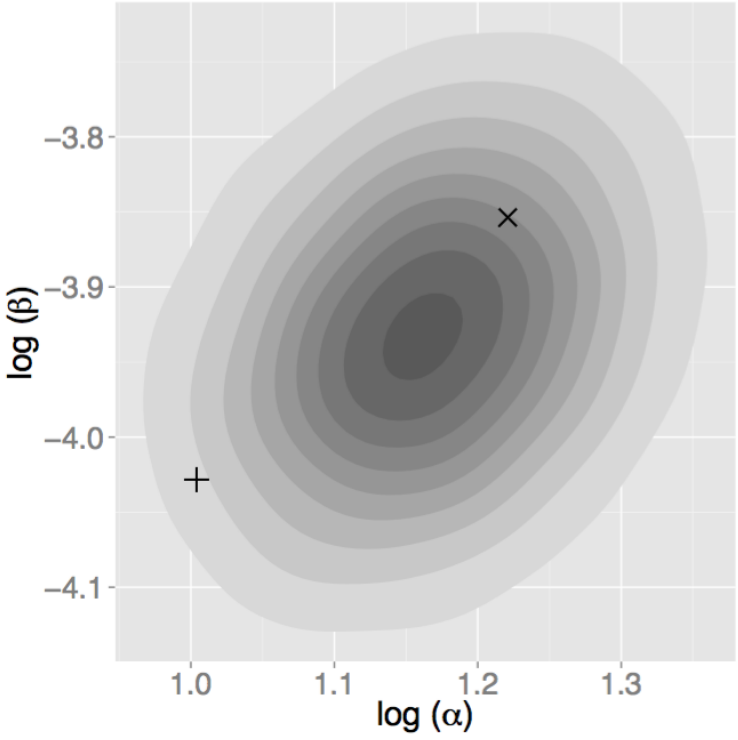
satisfies a recursion with continued fraction representation

$$\phi_{ab}^{(0)}(s) = \prod_{i=1}^{b} x_{ai} \cfrac{x_{a,b+1}}{Y_{a,b+1} + \cfrac{x_{a,b+2} Y_{ab}}{y_{a,b+2} + \cfrac{x_{a,b+3}}{y_{a,b+3} + \cfrac{x_{a,b+4}}{y_{a,b+4} + \cdots}}}}$$

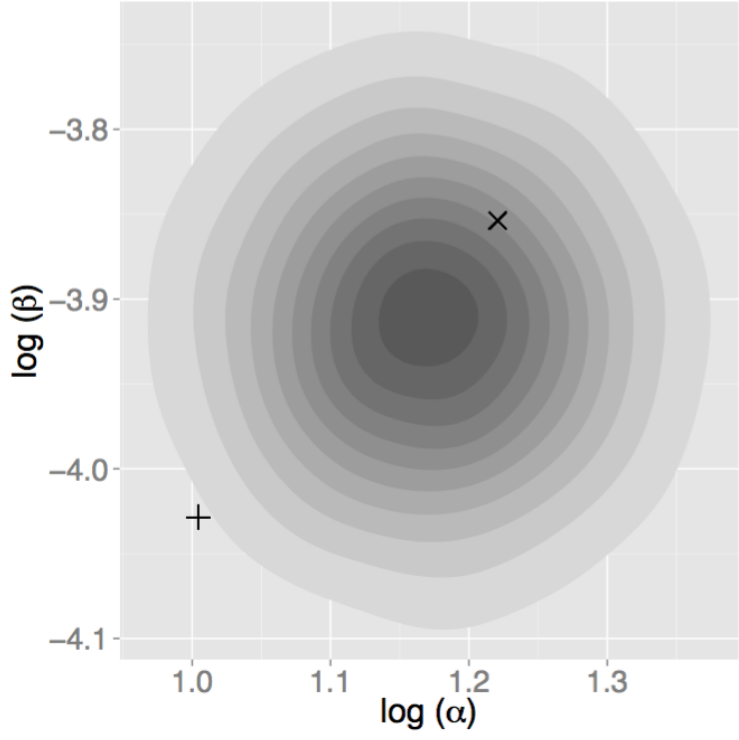
- Evaluate to finite depth, numerically invert Laplace transform [Ho, Xu, Crawford, Minin, Suchard 2017]

Back to branching processes

- Continued fraction method is limited to moderate outbreak sizes; derivation is delicate and hard to extend
- Two-type branching approximation yields analytic transition probabilities



(c) Continued fraction

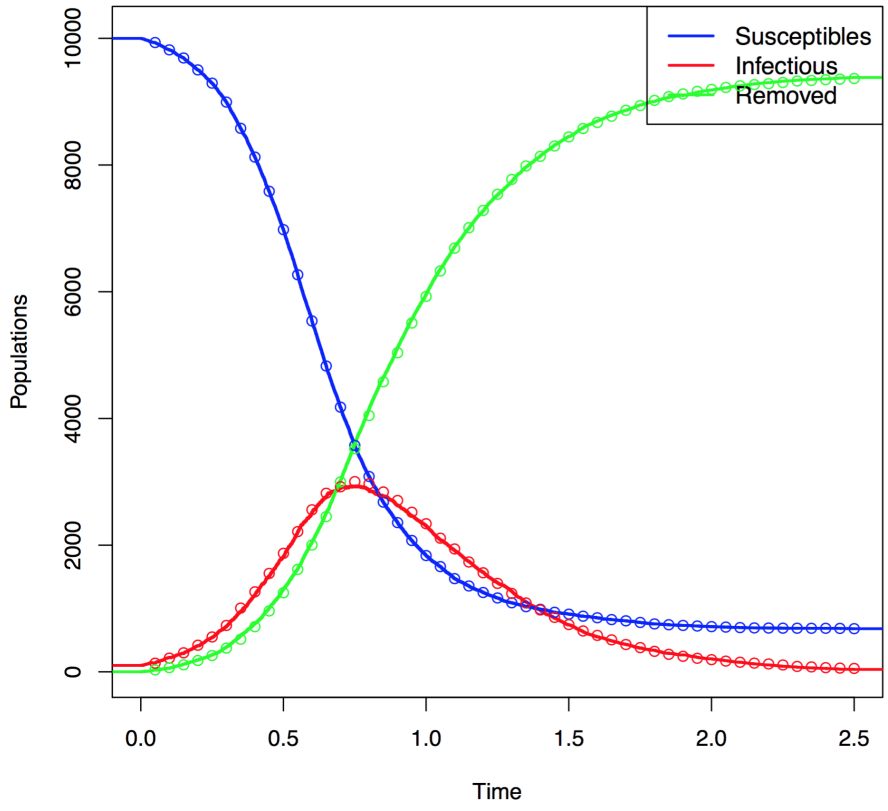


(d) Branching process

Current/future work: correcting the approximation

Branching process model as proposal density within MCMC

- Metropolis-Hastings step corrects approximation error



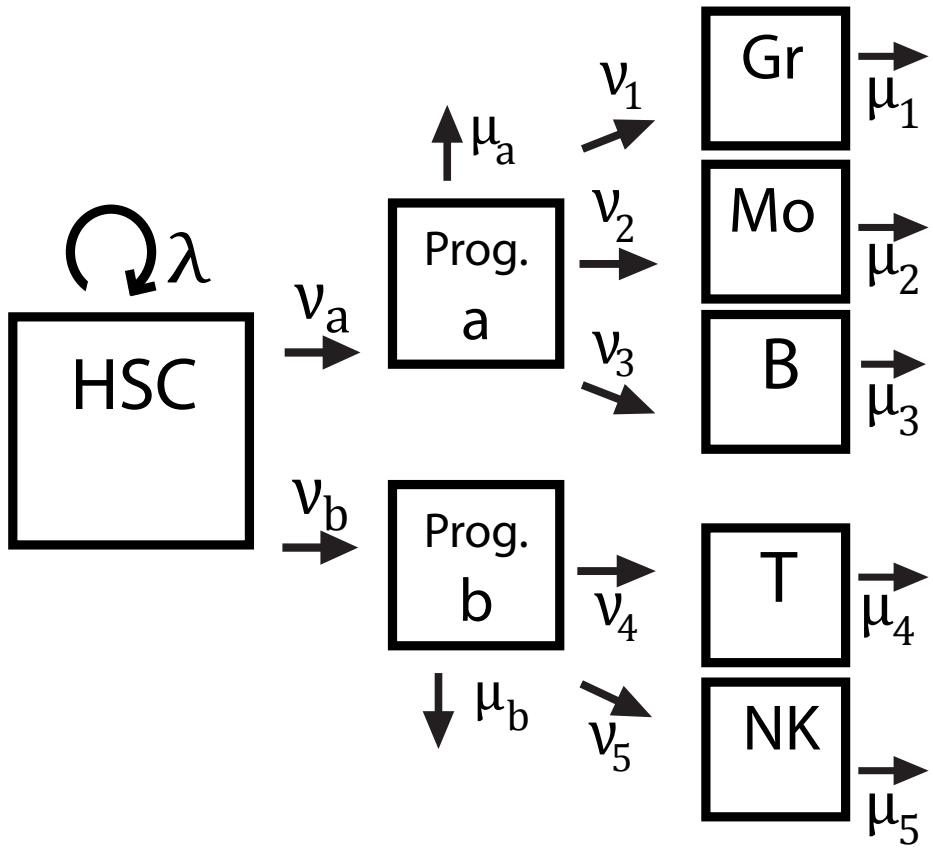
Circles represent true populations. I and R curves proposed from branching process, given true β, γ, and observed S population

References

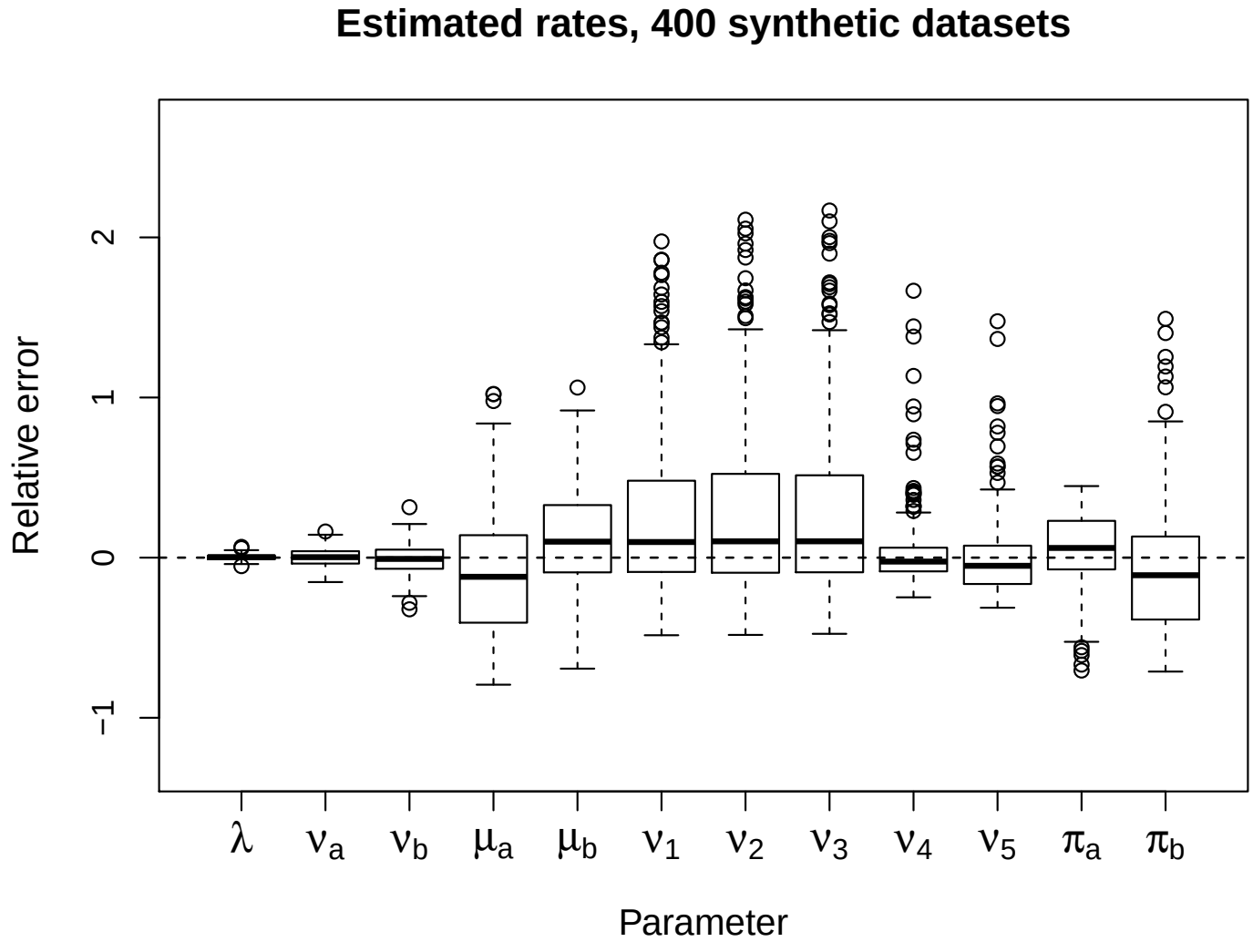
- JL Abkowitz, ML Linenberger, MA Newton, GH Shelton, RL Ott, and P Guttorp. Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. Proceedings of the National Academy of Sciences, 1990.
- SN Catlin, JL Abkowitz, and P Guttorp. Statistical inference in a two-compartment model for hematopoiesis. Biometrics, 2001.
- D Golinelli, P Guttorp, and JL Abkowitz. Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. Mathematical Medicine and Biology, 2006.
- 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, 2015.
- C Andrieu, A Doucet, R Holenstein. Particle Markov chain Monte Carlo methods. JRSS: Series B, 2010.
- J Xu, S Koelle, P Guttorp, C Wu, C Dunbar, JL Abkowitz, VN Minin. Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis. ArXiv preprint, 2016.
- L Ho, J Xu, FW Crawford, VN Minin, MA Suchard . Birth (death)/birth-death processes and their computable transition probabilities with statistical applications. Journal of Mathematical Biology, 2017 (to appear).

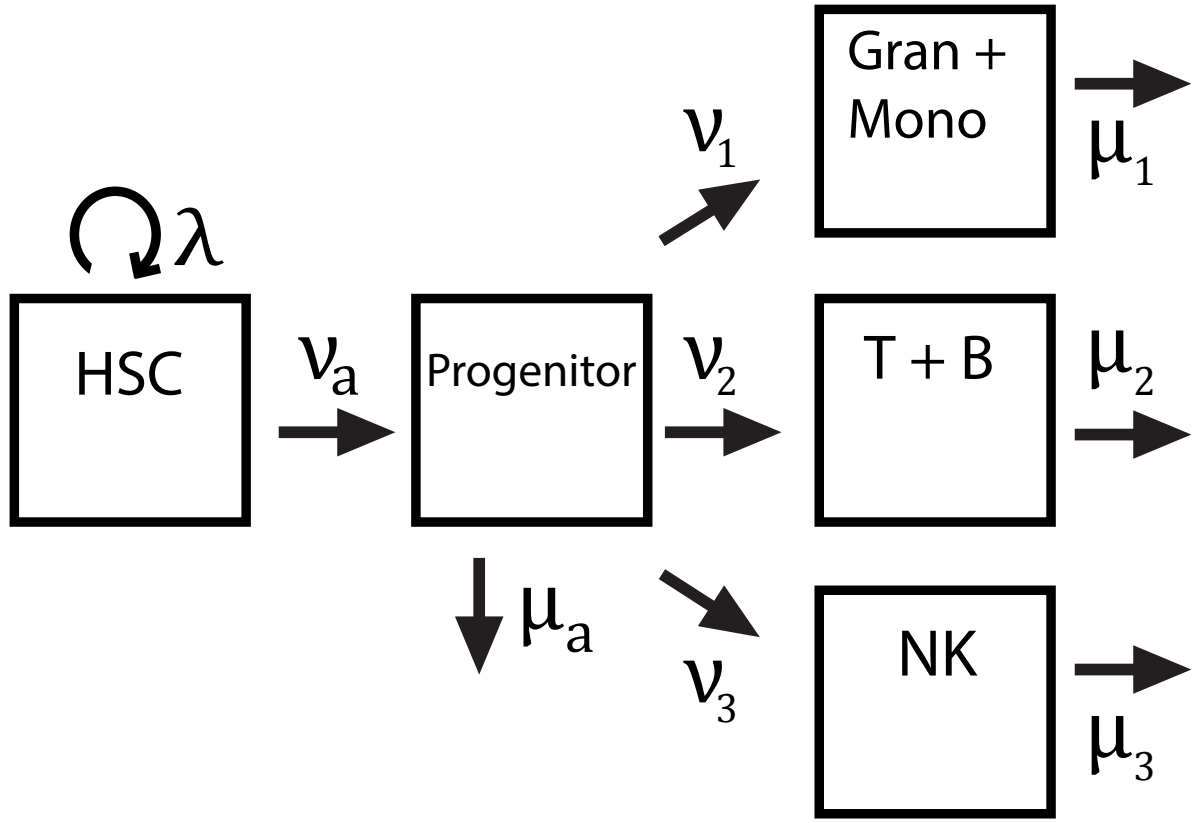
Simulation study

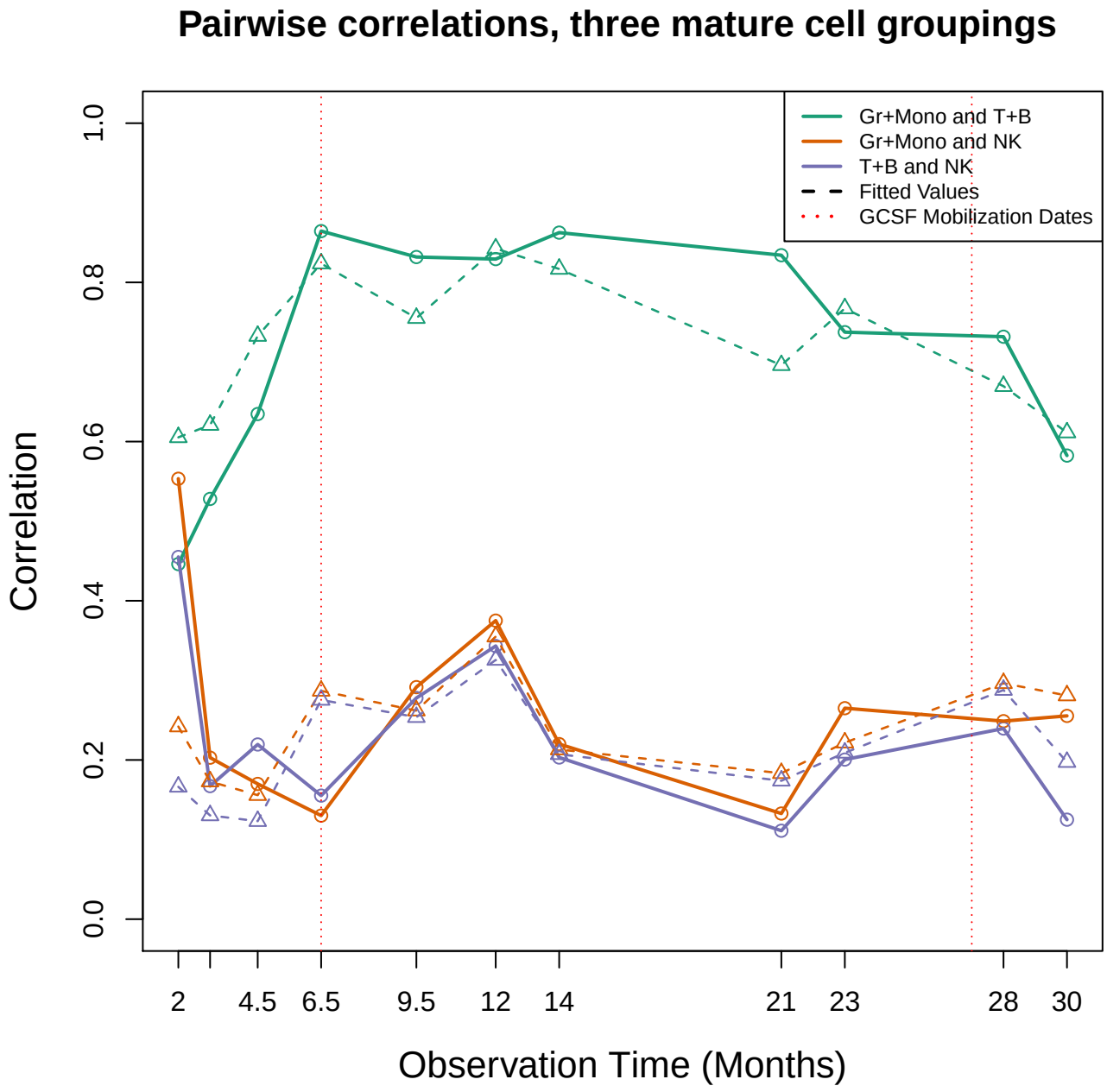
Infer parameters for partially observed datasets simulated from models in our class: we use the following as an example

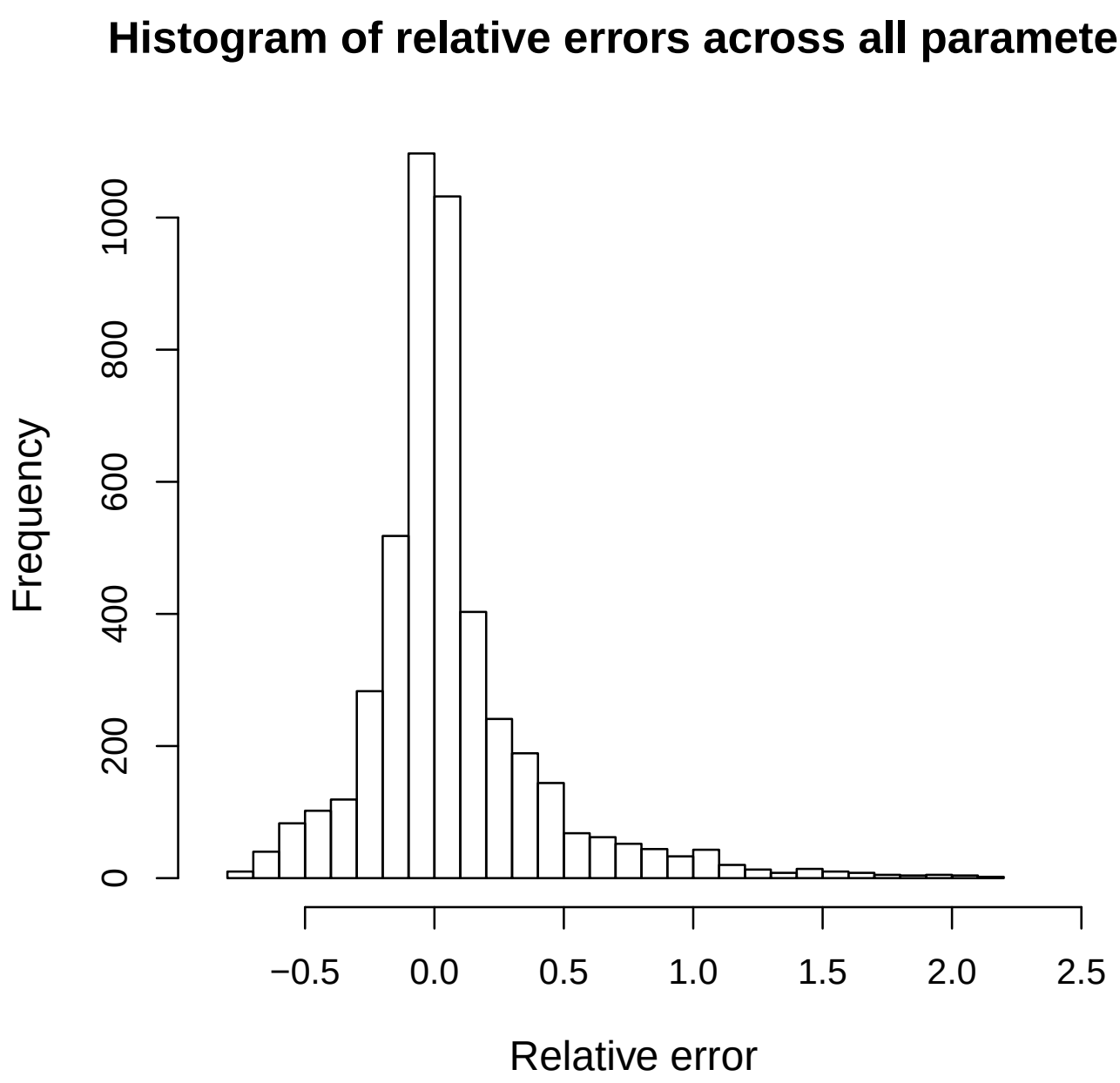


Results: simulation study




Gran +
Mono
ν_1
μ_1
λ
HSC
ν_a
Progenitor
ν_2
T + B
μ_2
μ_a
ν_3
NK
μ_3


Pairwise correlations, three mature cell groupings
Gr+Mono and T+B
Gr+Mono and NK
T+B and NK
Fitted Values
GCSF Mobilization Dates
Correlation
0.0
0.2
0.4
0.6
0.8
1.0
2
4.5
6.5
9.5
12
14
21
23
28
30
Observation Time (Months)


Histogram of relative errors across all parameters
Frequency
0
200
400
600
800
1000
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Relative error

Performance under model misspecification

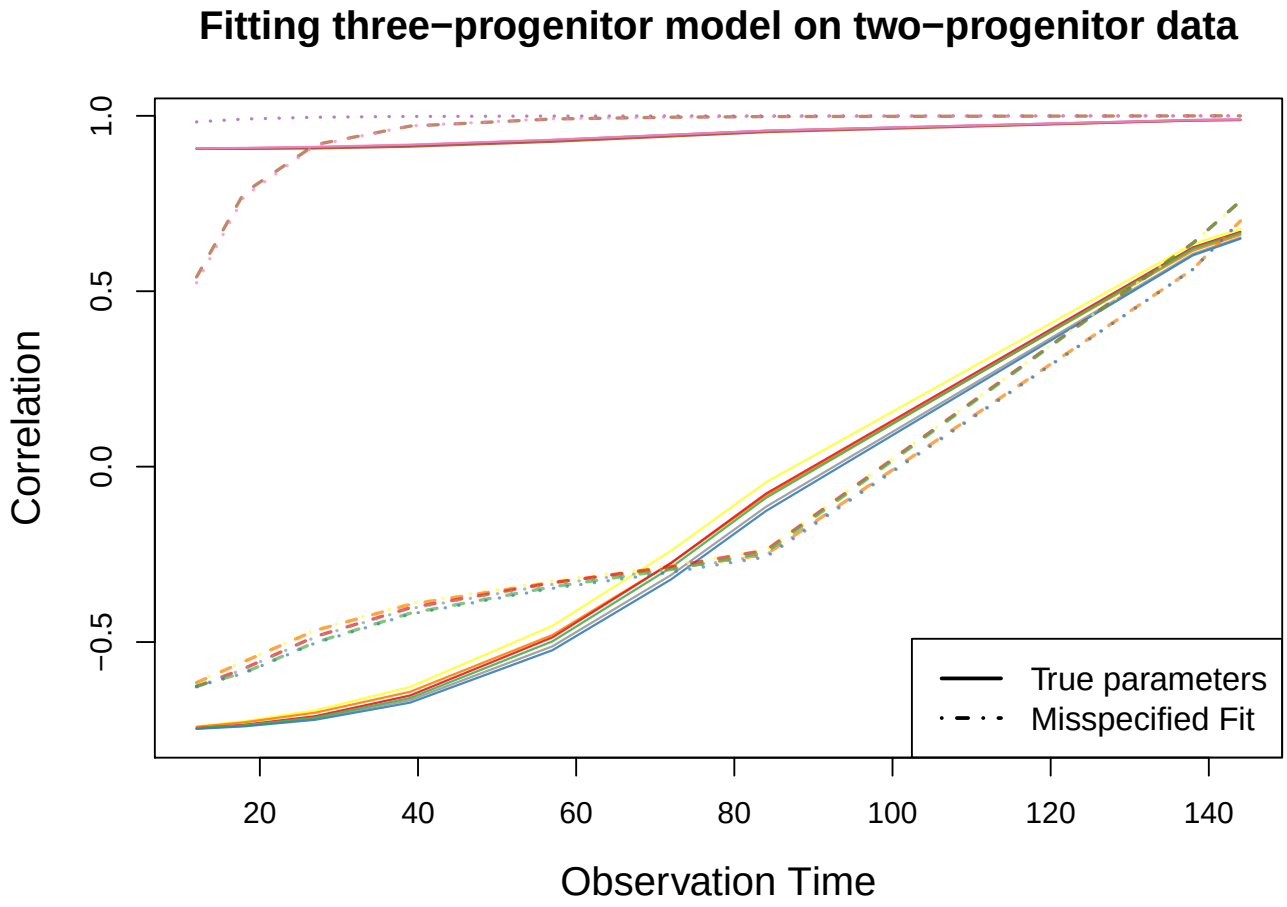


Figure: Inference on same synthetic data generated from the two-progenitor model, but misspecifying a three-progenitor model

Performance under model misspecification

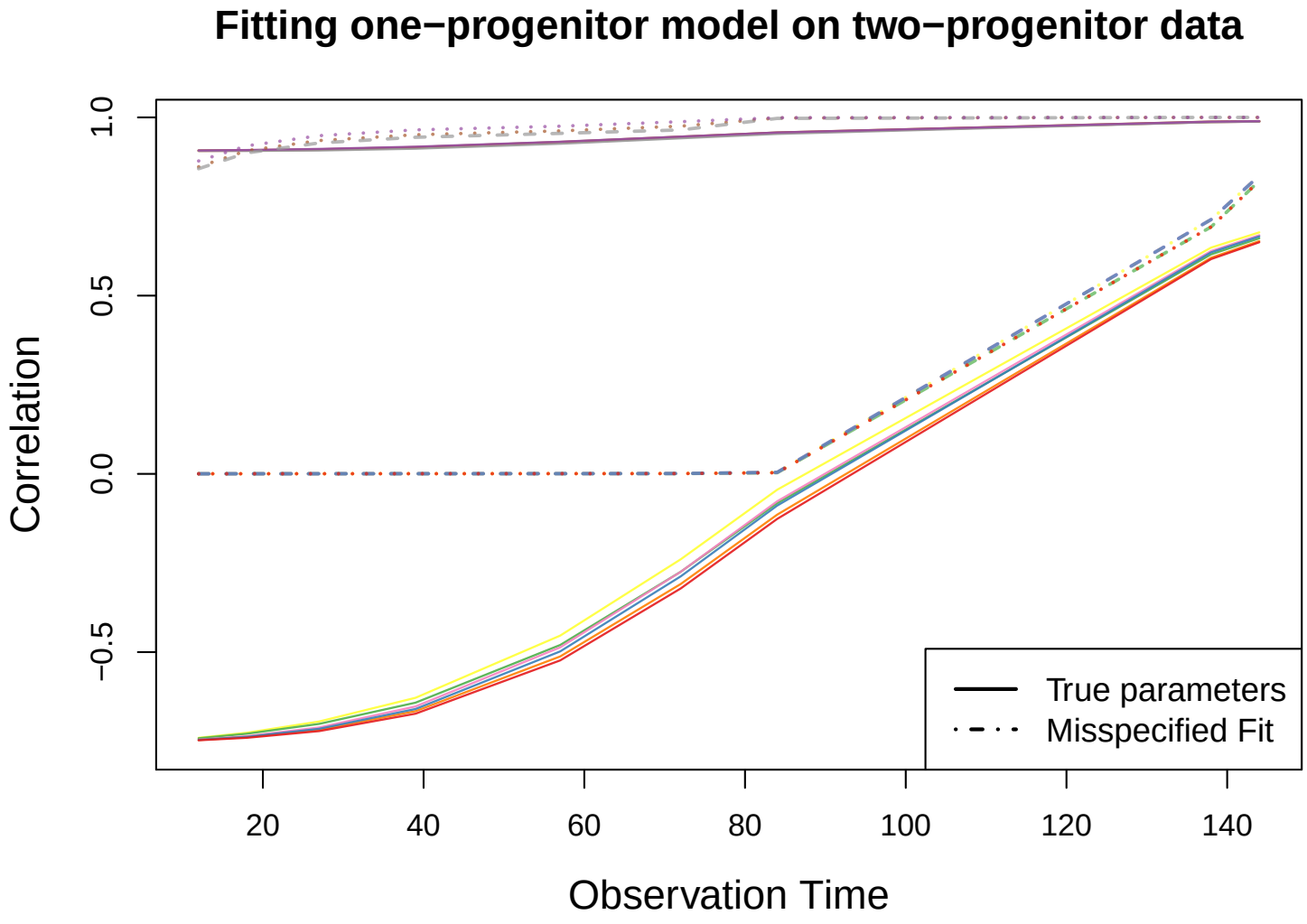


Figure: Here we wrongly assume there is one common progenitor

Performance under model misspecification

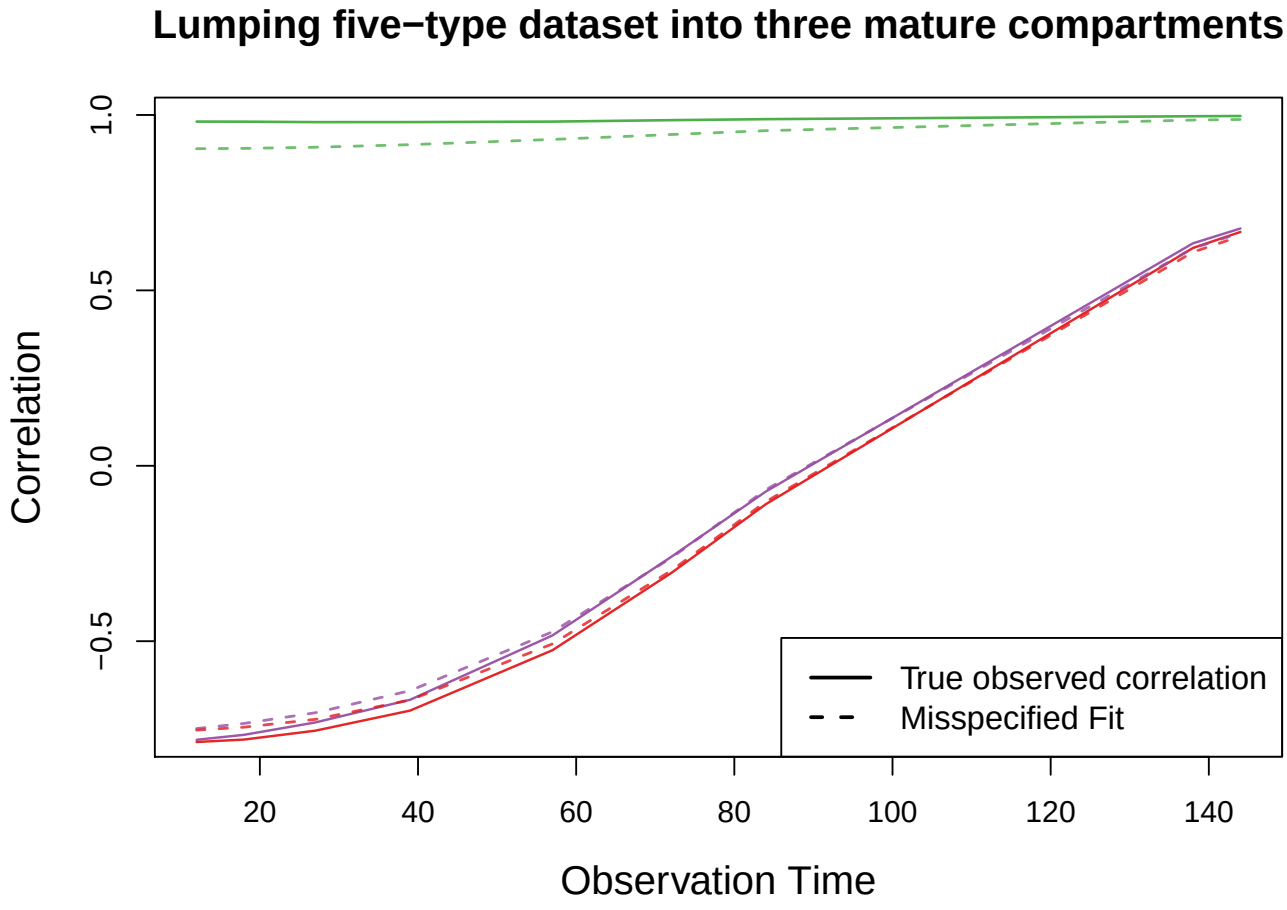


Figure: When we lump mature compartments together but otherwise correctly specify progenitors they are descended from, the fit is still good.

Parameter estimates

Estimated parameter	5-type model fit	3-type model fit
HSC renewal λ	0.0634	0.0497
HSC diff ν_0	2.80×10^{-6}	1.11×10^{-5}
Progen. death μ_0	0.000	0.000
Progen. diff. to Type 1 ν_1	1614.7	2635.7
ν_2	6093.6	283.3
ν_3	39.6	173.1
ν_4	126.1	NA
ν_5	64.4	NA
Mature death of Type 1 μ_1	0.5	0.7
μ_2	0.7	0.01
μ_3	0.01	0.40
μ_4	0.01	NA
μ_5	0.45	NA
Percentage barcoded at HSC	0.289	0.148

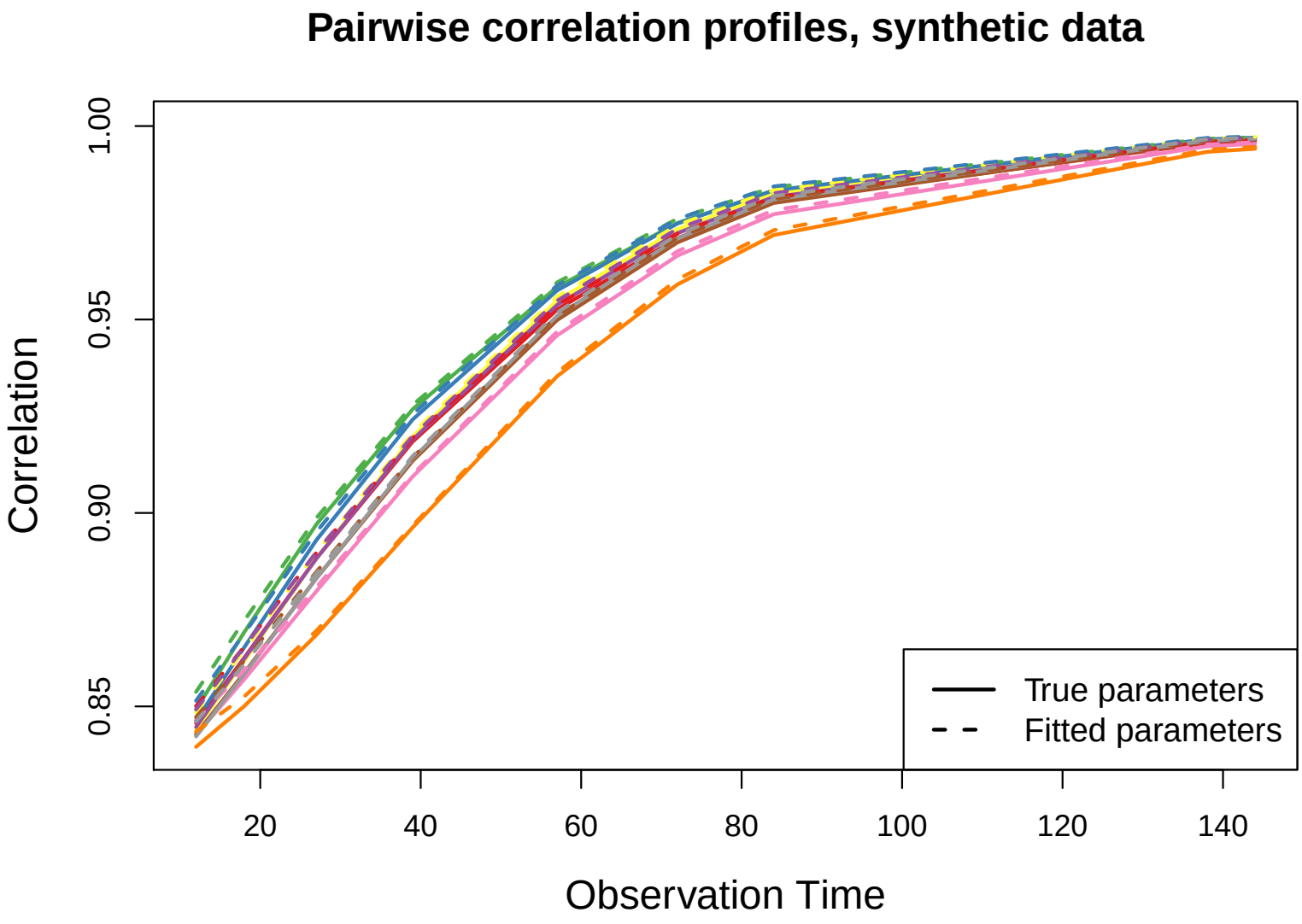
The emission distribution

Observation model: read count data $\mathbf{Y}^p(t)$ for barcode p at time t are distributed according to *multivariate hypergeometric distribution*:

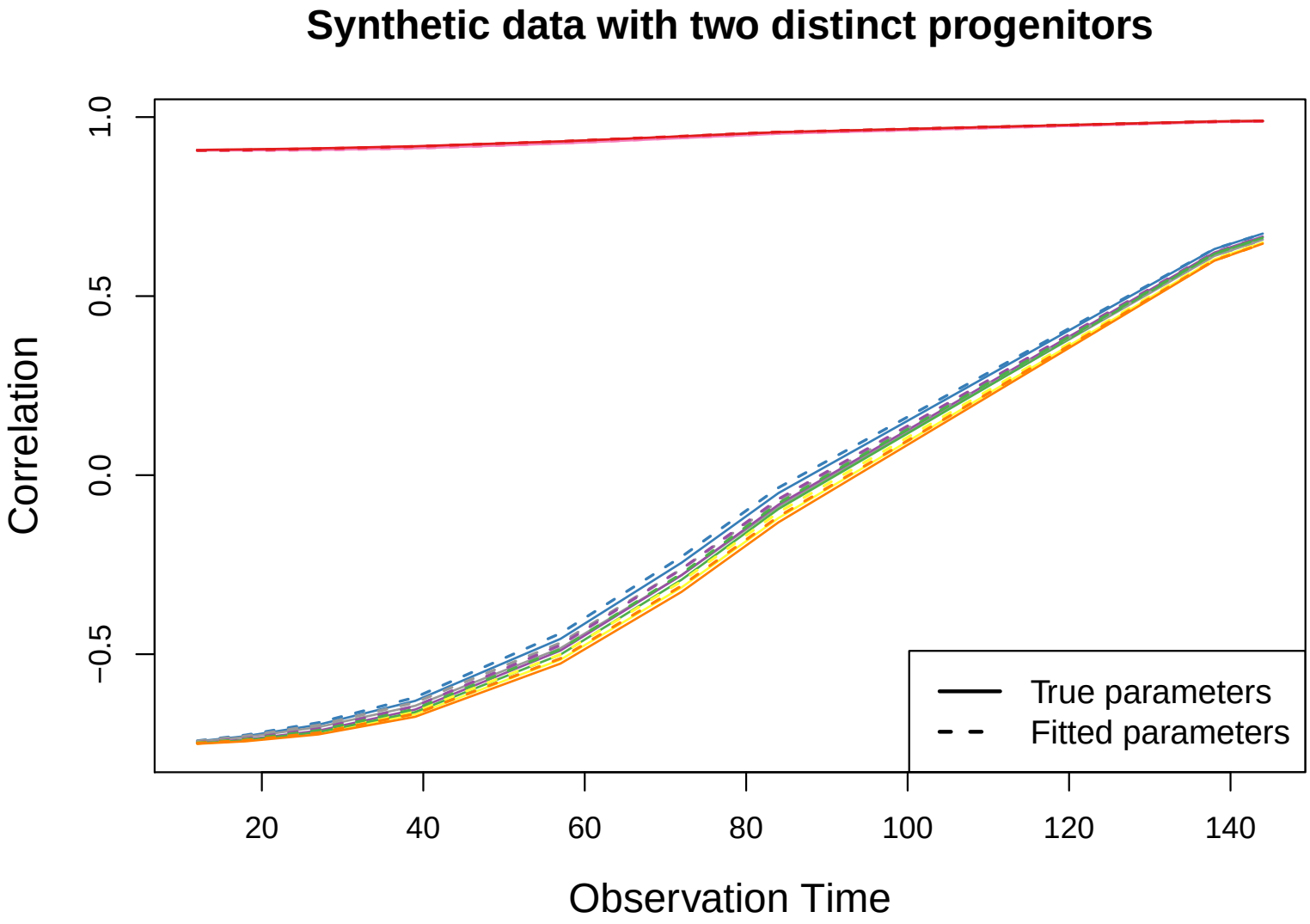
$$Y_1^p(t) \mid \mathbf{X}(t) \sim \text{hypergeom}(N_3, X_3^p(t), n_1),$$
$$Y_2^p(t) \mid \mathbf{X}(t) \sim \text{hypergeom}(N_4, X_4^p(t), n_2),$$

- n_1 and n_2 are known numbers of sampled cells of types 3 (Gr+Mono) and 4 (T+B+NK)
- N_3 and N_4 are known total numbers of barcoded type 3 and 4 cells in the animal
- $X_3^p(t)$ and $X_4^p(t)$ are unknown numbers of types 3 and 4 cells with barcode p

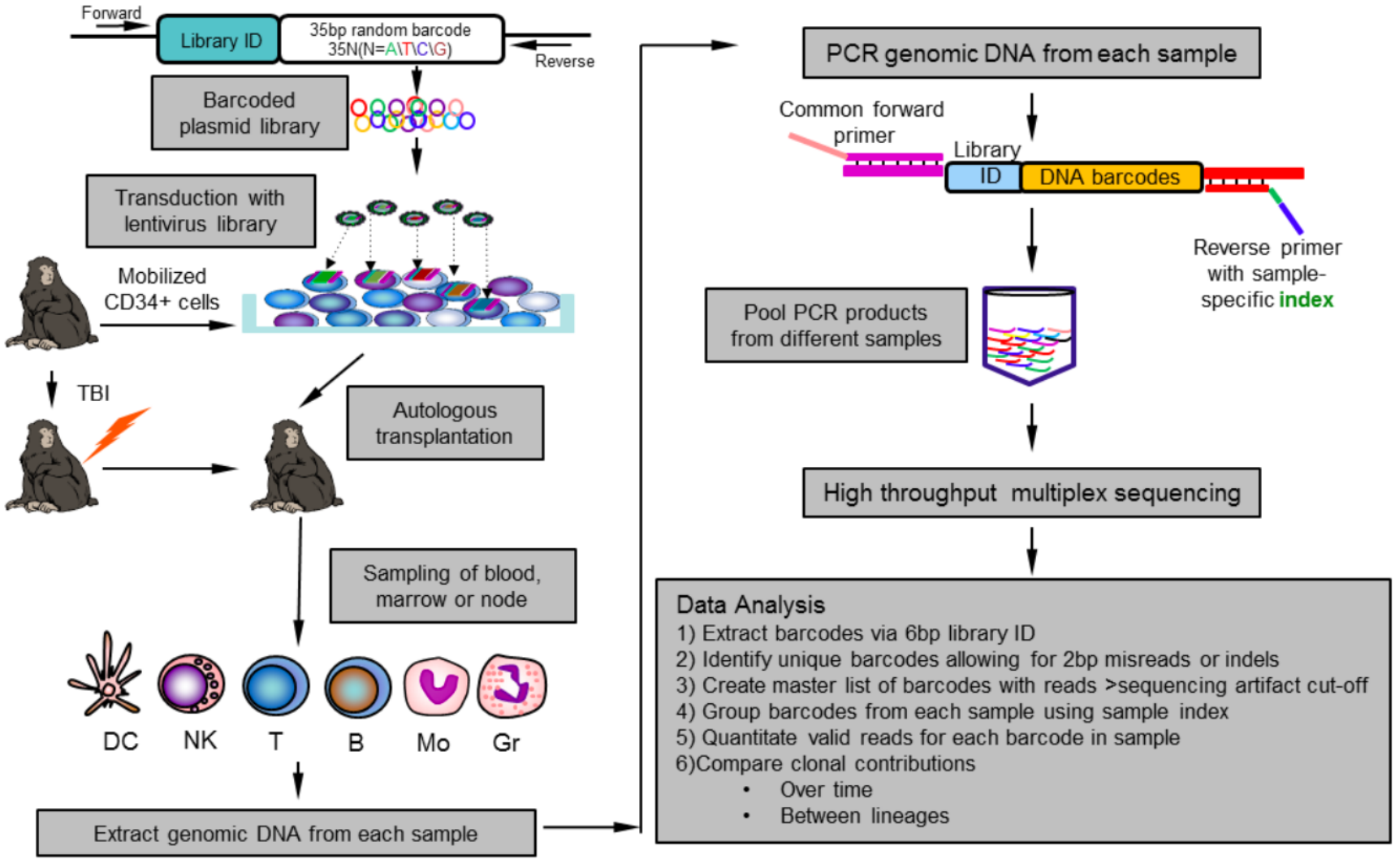
Fitted correlation plots



Fitted correlation plots



Experimental Design



Wu et al. 2014