Undergraduate research and graduate school so far

Jason Xu, Department of Statistics

ACMS Seminar
Jan 14, 2016

Why undergraduate research?

First of all, it's fun

- Classroom assignments can feel boring and contrived
- Work on new and relevant questions
- Low-stress way to decide whether research is for you
- Collaborate with excited and interested peers
- Travel to research programs, conferences, etc.

Some practical considerations

Second, it's good for you

- Closer relationship with professors
- Looks good on CV
- Attractive to employers
- Great preparation for grad school
- You often get paid

My undergraduate research experiences

My professor informed me about (and got me into) an NSF REU program in combinatorics and number theory

- Spent the summer in Oregon researching combinatorial game theory
- The following summer: REU in Claremont on Markov chain Monte Carlo techniques
- This one got me seriously interested in research, specifically in statistics and stochastic processes
  - I am now in grad school
  - I am in the statistics department
  - I research stochastic processes

Next steps

- Continued in that area as a research assistant at my home institution
- Had something to write about when applying to graduate school, fellowships, etc.
- Had stronger recommendation letters from research mentors
- Summer *after* graduation, participated in a summer program working on research problems posed by various companies (UCLA RIPS)

Some takeaways

- NSF REU programs are a great resource!
- UMN Duluth, Williams (SMALL), Cornell, Wisconsin-Madison (Emory?)
- Other unique programs sponsored by industry and other foundations
- Don't be afraid to apply, email professors, etc.

Some things I've worked on in grad school

Briefly, I've worked on inference for dynamic stochastic systems when computationally expensive techniques such as MCMC are intractable.

- Approximate Bayesian treatment of massive hidden Markov models
  - Variational inference, efficient optimization, parallel computing
  - "Machine learning", application to pattern discovery for chromatin segmentation in genomics
- Likelihood methods for partially observed Markov branching processes
  - PDEs, Fourier methods, generating functions
  - "Methodology", applications to molecular epidemiology

Some things I've worked on in grad school

- Accelerating generating function techniques via sparsity
  - "Computation", compressed sensing, proximal methods
- Smooth, minimally Lipschitz function interpolation
  - "Theory", analysis, empirical processes, convex optimization

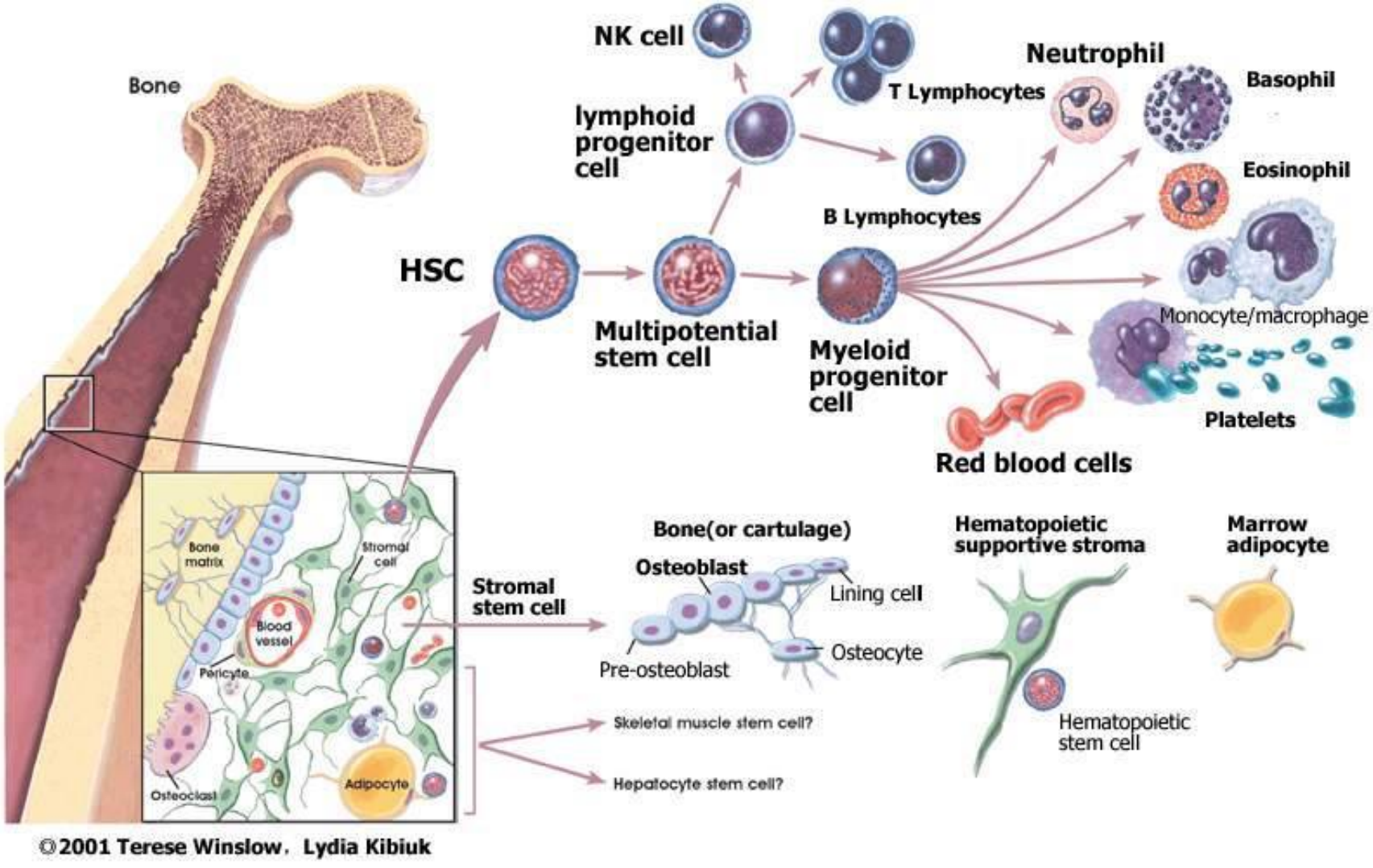
Currently, developing parameter estimation techniques for new models of cell differentiation applied to recent single-cell lineage tracking of *in vivo* hematopoiesis

What is hematopoiesis?

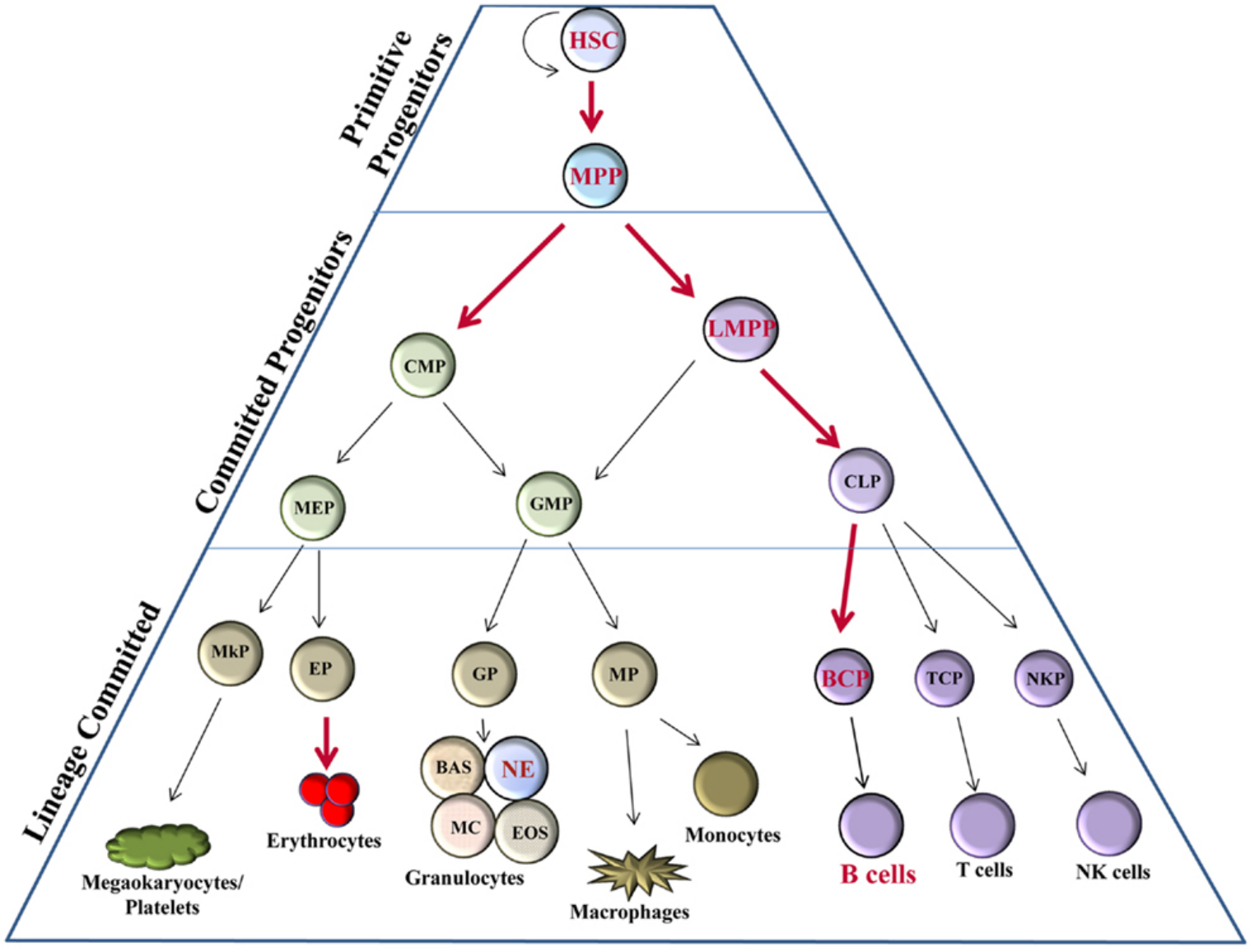
Complex system in which self-renewing hematopoietic stem cells (HSCs) differentiate via a series of intermediate progenitor cell stages to produce blood cells

- Dynamics and structure are largely unknown
- Clinically important: stem cell transplantation is a mainstay of cancer therapy; all blood cell diseases are caused by malfunctions in the hematopoietic process
- Modeled using stochastic compartmental models: cells replicate and differentiate as a continuous-time *Markov branching process*

Proposed models of hematopoiesis



Branching structure



Controversial tree structure

Science DOI: 10.1126/science.aab2116

Read Full Text to Comme

RESEARCH ARTICLE

Distinct routes of lineage development reshape the human blood hierarchy across ontogeny

Faiyaz Notta[1,2,*,†], Sasan Zandi[1,2,*], Naoya Takayama[1,2], Stephanie Dobson[1,2], Olga I. Gan[1], Gavin Wilson[2,4], Kerstin B. Kaufmann[1,2], Jessica McLeod[1], Elisa Laurenti[6], Cyrille F. Dunant[7], John D. McPherson[3,4], Lincoln D. Stein[2,4], Yigal Dror[5], John E. Dick[1,2,‡]

comments related other discussions (2) filter by field

Stem-cell scientists led by Dr. John Dick have discovered a completely new view of how human blood is made, upending dogma from the 1960s.

4938

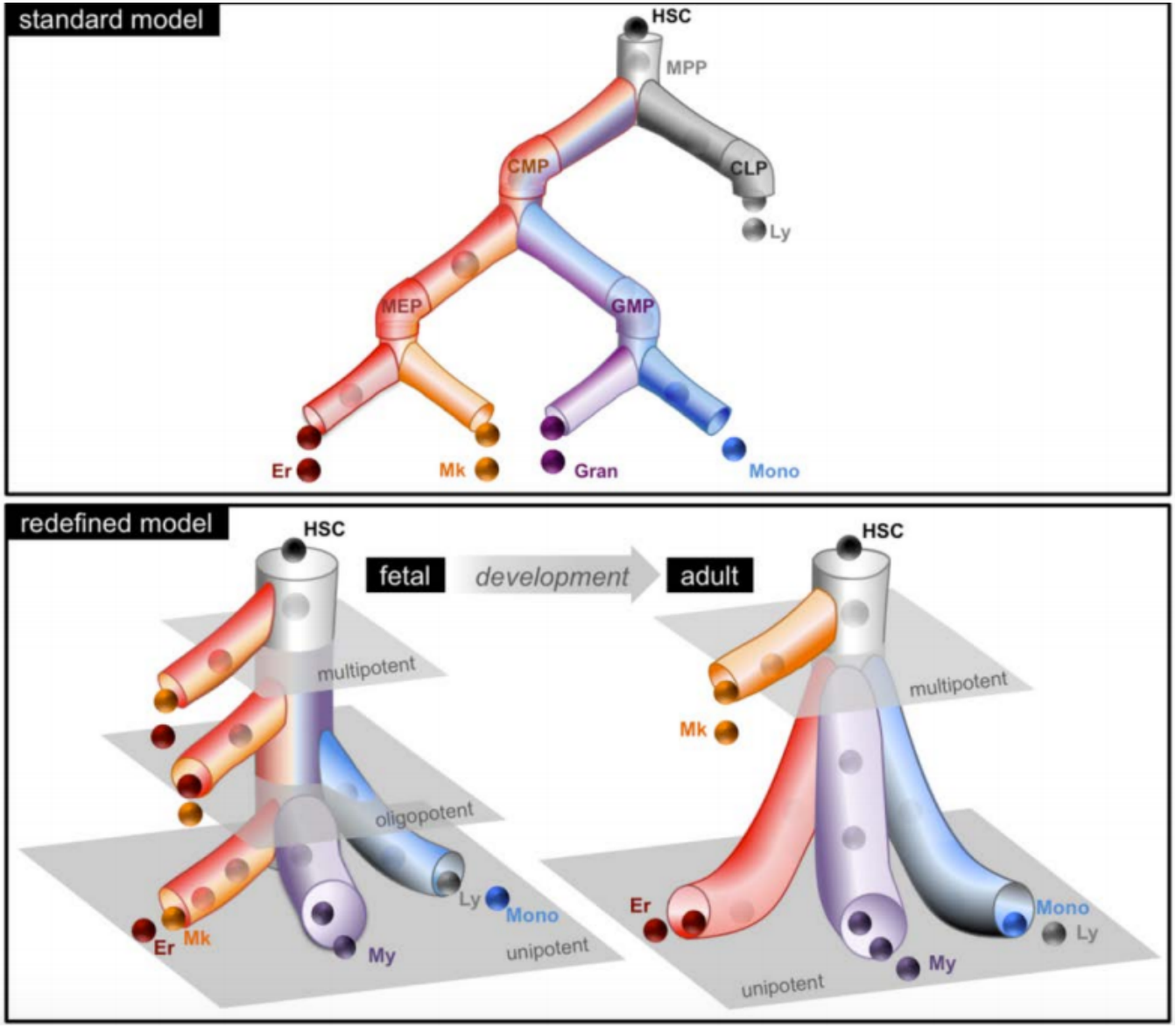
sciencedaily.com

4 days ago by jjgfun MS | Biology | Environmental Science

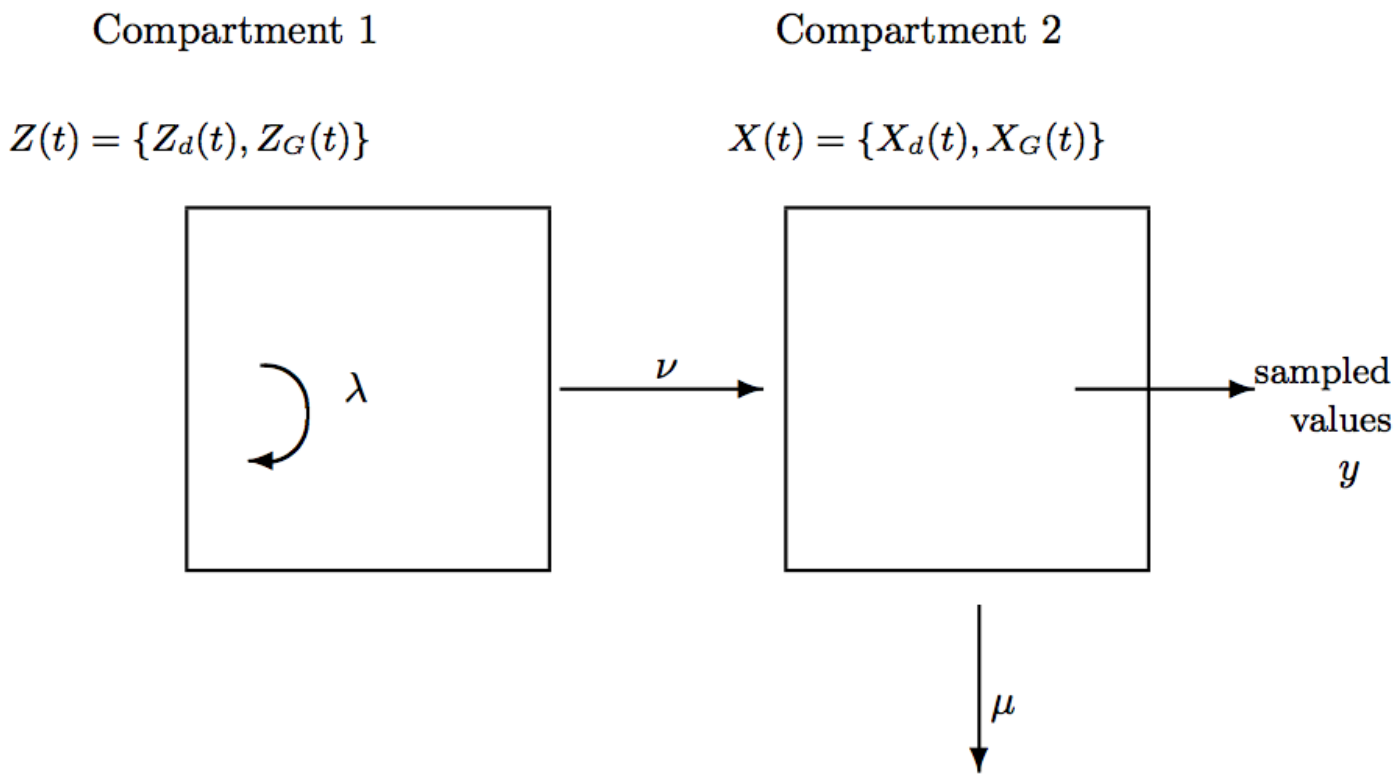
276 comments source share save hide give gold report hide all child comments

Top 200 Comments show all 276 subscribe

Controversial tree structure



Statistical simplifications



Limited to overly simple model due to statistical challenges and resolution of data, yet estimation is already difficult. Past attempts include simulation studies, normal approximation, reverse jump MCMC, method of moments.

Our dataset: rhesus macaques lineage tracking

Cell Stem Cell
Article

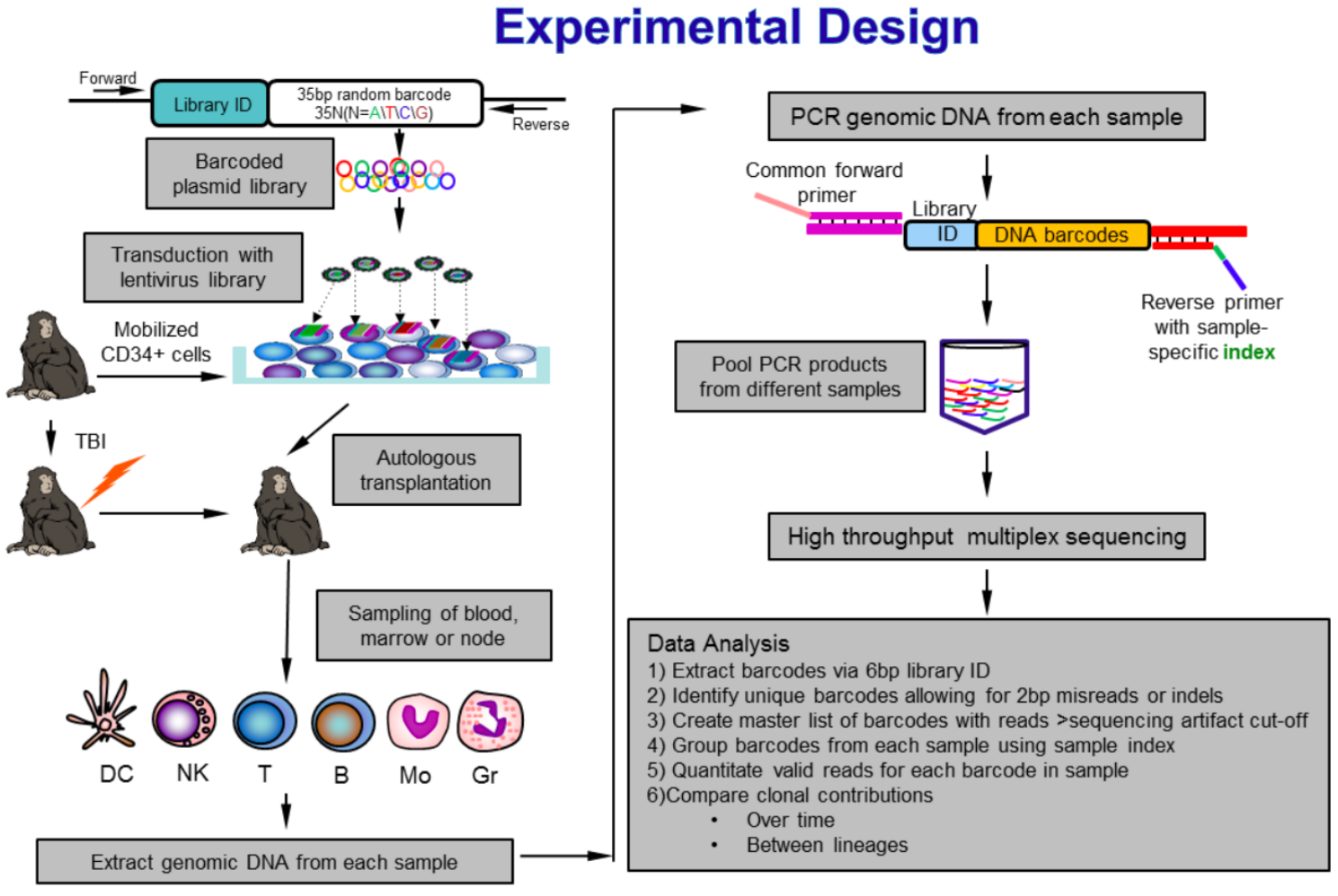
Clonal Tracking of Rhesus Macaque Hematopoiesis Highlights a Distinct Lineage Origin for Natural Killer Cells

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

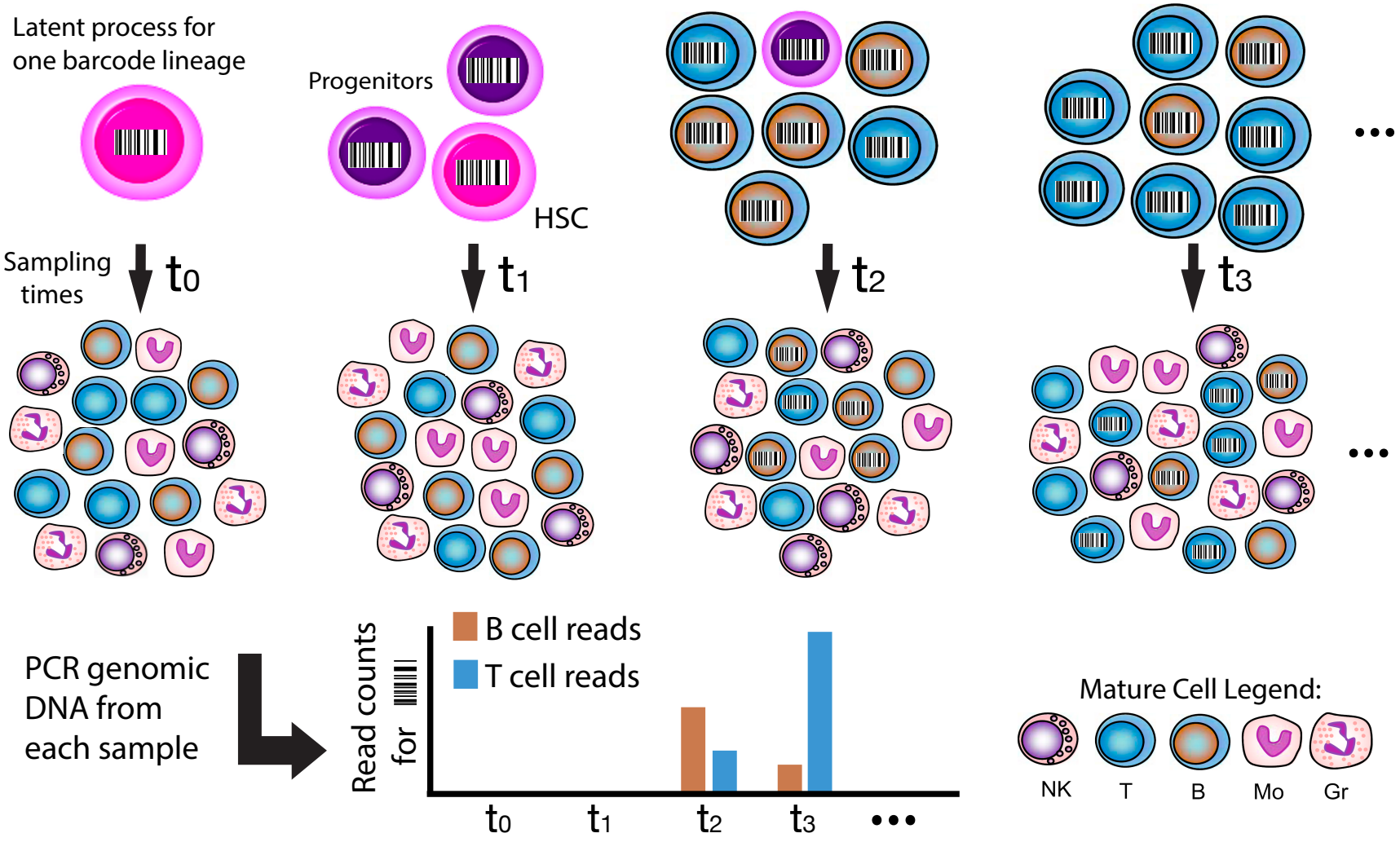
NIH group: cell lineage barcoding data in primates. Their in-house analysis has already produced interesting findings.

Our dataset: rhesus macaques lineage tracking



Goal: develop **model framework** and **estimation methods**

A hidden stochastic model



Data and challenges:

- At each observation time, the experimental protocol yields a read count corresponding to each barcode ID present in each cell type sample
- Data: independent, identically distributed time series of observations from the hematopoietic process
- $\Rightarrow$ fit more complex models than previous statistical studies

Challenges

- Hidden process, discrete observations, experimental noise
- Huge latent populations, many interacting types

Richer latent branching models

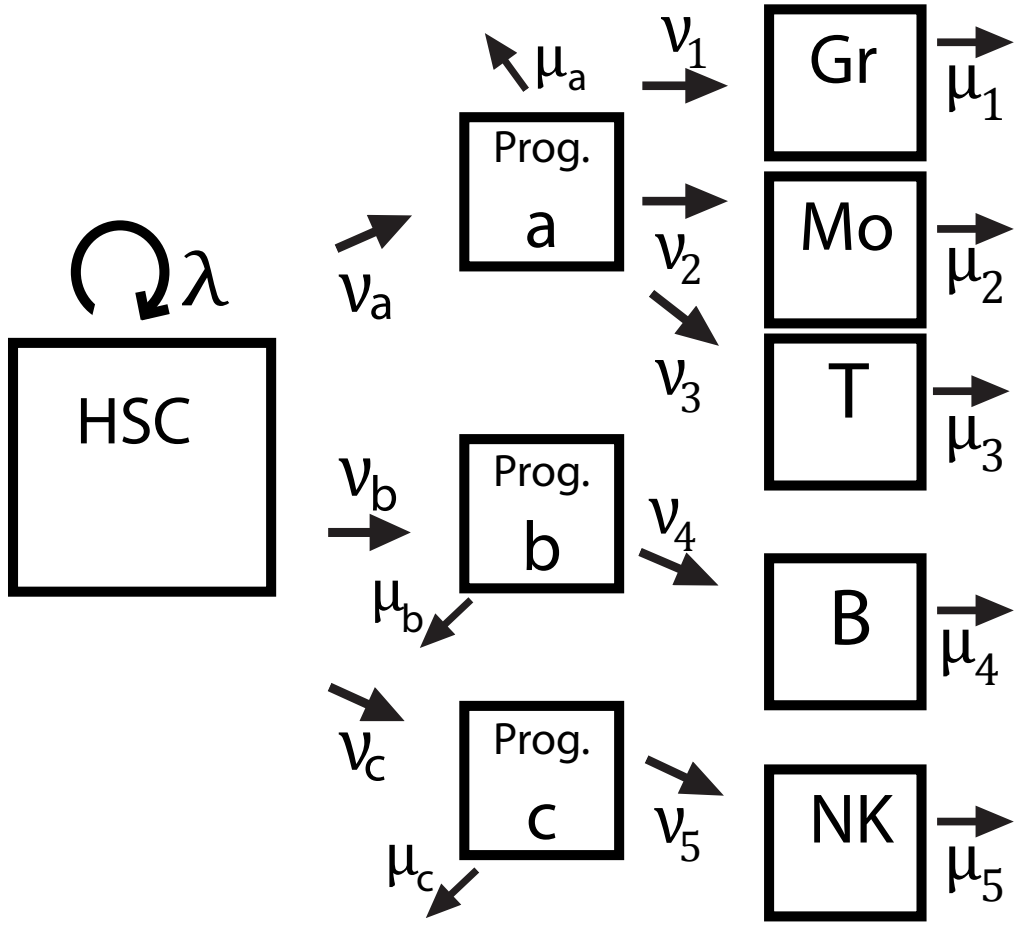


Figure: Our class of models allows an arbitrary number of progenitors and mature cell types

Inference: our method

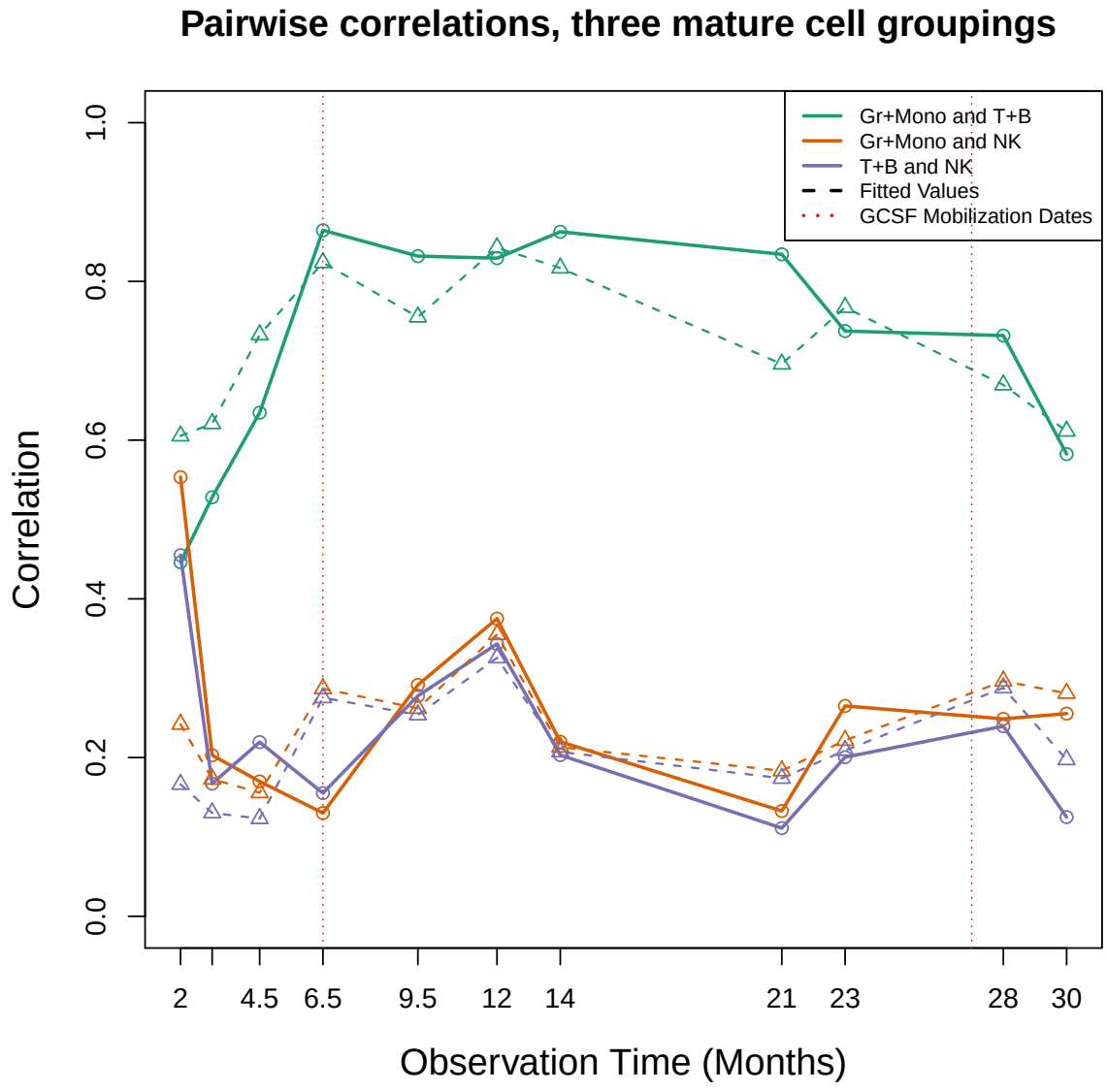
Loss function estimation: we match population correlations with empirical read count correlations across barcode lineages via minimizing the *loss function*

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{J} \left[\psi_j(\boldsymbol{\theta}; \mathbf{Y}) - \hat{\rho}_j(\mathbf{Y})\right]^2 ,$$

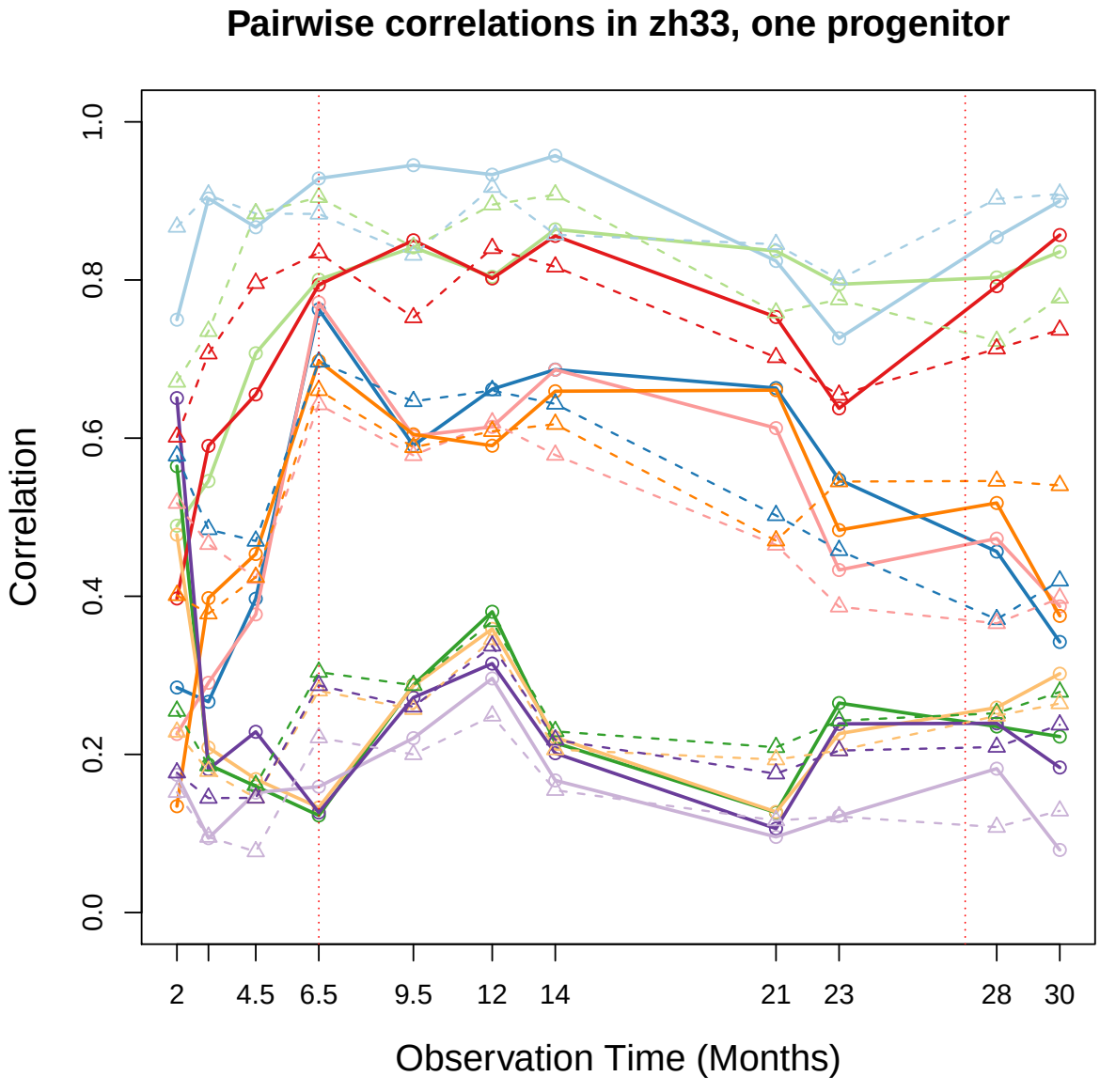
- $\hat{\rho}_j$ denotes empirical correlation across IDs p at time t_j
- $\psi_j(\boldsymbol{\theta}; \mathbf{Y})$ denotes model-based correlations that we calculate mathematically, given current parameters $\boldsymbol{\theta}$

Estimating best parameters reduces to **nonlinear least squares optimization**: $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})$.

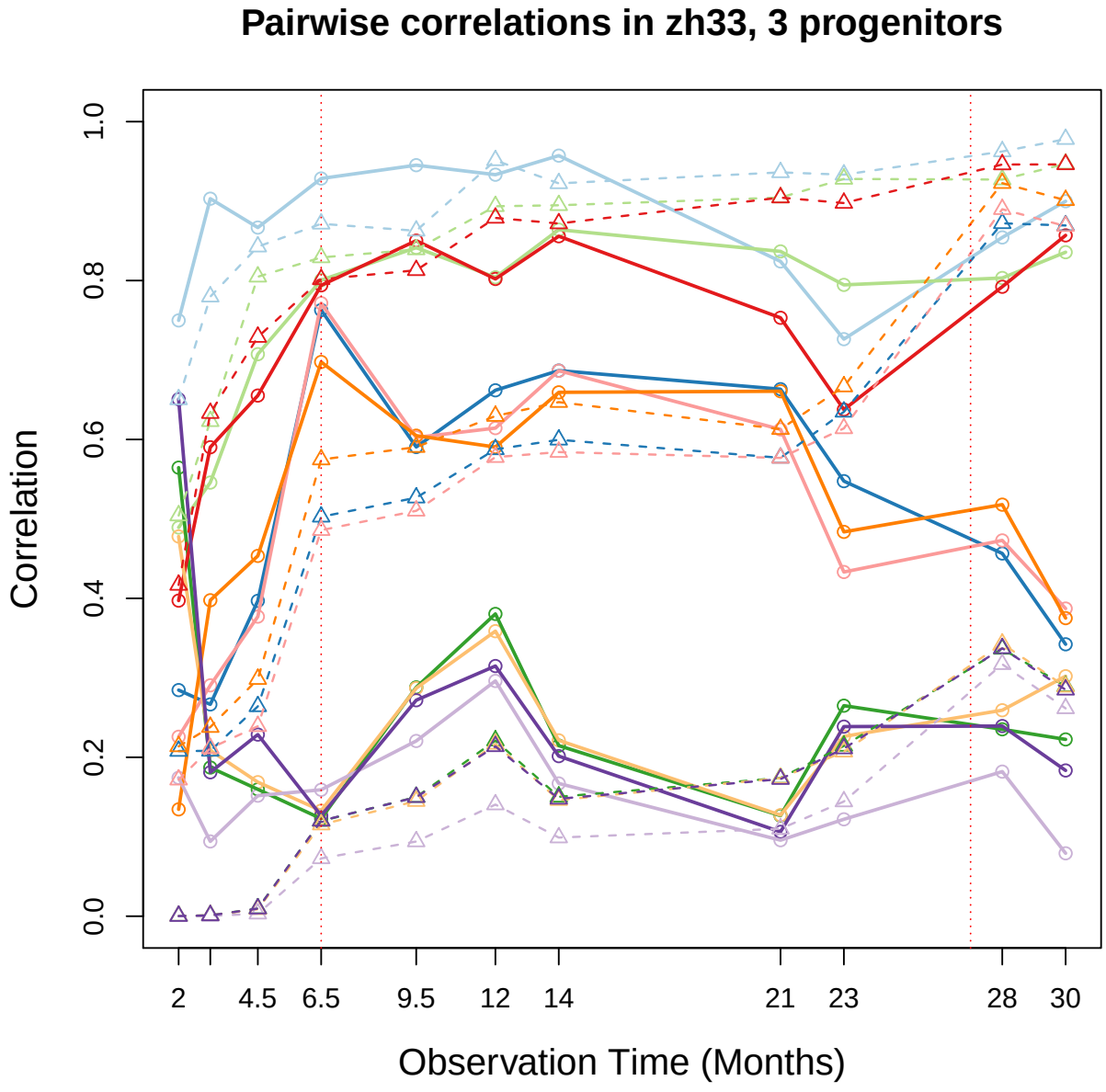
Results


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)

Results


Pairwise correlations in zh33, one progenitor
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)

Results


Pairwise correlations in zh33, 3 progenitors
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)

Some findings

- HSC self-renewal rate $\hat{\lambda}$ estimated to be about once every 31 weeks, consistent with previous studies focusing on the stem cell
- Estimate of initial marking levels indicates $\sim 15\%$ of cells are marked at the HSC level, while others are marked further down the line, and is consistent with side information from fluorescence data
- Newly estimated intermediate differentiation rates are reasonable given information about cell abundances and lineage relations from initial analyses
- Obtaining rate estimates in a generative model for the data is novel; initial distributions previously unidentifiable

Summary

- Richer, more flexible statistical models of hematopoiesis and experimental protocol than previous studies
- First parameter estimation methods for fitting time-series data from lineage barcoding experiments
- Efficient method applies to a general class of multi-compartmental models
- Future: likelihood-based approaches, model selection strategies

Thanks for listening! More questions: `jasonxu@uw.edu`