UNIVERSITY OF CALIFORNIA

Los Angeles

Mathematical Modeling of Clonal Dynamics in Primate Hematopoiesis

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomathematics

by

Song Xu

2018

ABSTRACT OF THE DISSERTATION

Mathematical Modeling of Clonal Dynamics in Primate Hematopoiesis

by

Song Xu
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2018
Professor Tom Chou, Chair

Recent developments in cell labeling techniques allow studying activities of the massive cell population at single-cell or clone-level resolution. The generated data are usually featured by a large number of labels but small sizes of samples. However, the underlying clonal dynamics are usually stochastic and high-dimensional in nature. Thus, inferring the full upstream mechanisms from the downstream sample data is essentially an overfitting problem and poses new challenges in associated computational, statistical, and modeling methods. In this work, I study the clonal dynamics in the hematopoietic system of rhesus macaques based on a decade-long clonal-tracking experiment. I first develop a computational algorithm that tries to improve the quality of the sampled data by correcting DNA-sequencing errors. Then, I take the cell count (clone size) data and build a multi-compartment neutral model to study the dynamics of each labeled clone. To avoid overfitting, I simplify the mechanistic model and select robust statistical features of the data. Finally, when analyzing the birth-death-immigration (BDI) clonal dynamics under global carrying capacity, I find that the usually invoked mean-field approach fails to predict simulated distribution of clone sizes. I solve this problem by transforming the problem and further approximating the carrying-capacity effect by the fixed-total-size constraint in a Moran model. I hope this work not only solves the current technical problems, but also contributes to the ongoing efforts in understanding the long-term multi-clonal dynamics in complex systems.

The dissertation of Song Xu is approved.

Irvin Sy Chen

Alexander Jacob Levine

Alexander Hoffmann

Tom Chou, Committee Chair

University of California, Los Angeles

2018

dedicated to my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without the financial support of the China Scholarship Council (No. 201408020004), NSF DMS-1516675 (PI: Dr. Tom Chou), R01 AI110297 (PI: Dr. Irvin Chen), and R01 HL125030 (PI: Dr. Irvin Chen).

I am grateful to all of those with whom I have had the pleasure to work with during my graduate study. I would especially like to thank my advisor, Dr. Tom Chou, for providing me with guidance on critical thinking and technical skills (especially the emphasis on mathematical rigor), sharing with me his vision in science, and helping me improve my writing skills. I also want to thank my co-authors Dr. Sanggu Kim, who taught me about the domain knowledge and the technical details of the clone-tracking experiment, and Dr. Irvin Chen, who supervised the whole research project and gave me key suggestions on how to make the work biologically relevant. Chapter Three is a version of the submitted work "Modeling large fluctuations of thousands of clones during hematopoiesis: the role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque" to the journal of PLoS Computational Biology (revision received on March, 2018).

Each of the members of my Dissertation Committee has provided me insightful suggestions and comments on my research and presentation. I also want to thank my cohorts at UCLA Biomathematics, especially Bhaven Mistry and Lae Un Kim, who shared the pains and joys of graduate study and encouraged and helped me when I got stuck. I owe a debt of gratitude to my lab mates back in Shanghai Jiao Tong University, especially Ying Tang and Ruoshi Yuan, for making valuable comments on my projects from a different angle. Most importantly, nobody has been more important to me in the pursuit of my career goal than the members of my family. Words cannot describe how fortunate I am to have you always standing by me so that I can focus and make progress both in science and in life.

VITA

2014–2018 Ph.D. Candidate in Biomathematics, UCLA

2013–2014 M.S. in Biomathematics, UCLA

2010–2013 M.S. in Computer Science, Shanghai Jiao Tong University (Dual M.S. in Electrical & Computer Engineering from Georgia Tech)

2006-2010 B.S. in Computer Science, Shanghai Jiao Tong University

PUBLICATIONS

Xu, S. & Chou, T., *Density-dependence-induced phase transition in clone abundances*, In preparation, (2018).

Xu, S., Kim, S., Chen, I., & Chou, T., *Mapping clone extinction and resurrection to intermittent hematopoietic stem cell differentiation: Analysis of a decade-long clone tracking study in rhesus macaque*, Revision under review: PLoS Computational Biology, (2018).

Tang, Y.*, **Xu, S.***, & Ao, P., *Escape rate for nonequilibrium processes dominated by strong non-detailed balance force*, Journal of Chemical Physics, 148(6), 064102, (2018).

Suryawanshi, G. W.*, **Xu, S.***, Xie, Y., Chou, T., Kim, N., Chen, I. S., & Kim, S., *Bidirectional Retroviral Integration Site PCR Methodology and Quantitative Data Analysis Workflow*, Journal of Visualized Experiments, 124, e55812-e55812, (2017).

Xu, S., Jiao, S., Jiang, P., & Ao, P., *Two-time-scale population evolution on a singular landscape*, Physical Review E, 89(1), 012724, (2014).

Jiao, S., **Xu, S.**, & Ao, P., *Adaptive landscape with Singularity in Evolutionary Processes.* In Nonlinear Dynamics and Complexity (pp. 163-189), Springer International Publishing, (2014).

CHAPTER 1

Introduction

Hematopoiesis is a process by which hematopoietic stem cells (HSCs) produce all the mature blood in an animal through a series of proliferating and differentiating divisions [ACG96]. Maintenance of balanced hematopoietic output is critical for an organism's survival and determines its response to disease and clinical procedures such as bone marrow transplantation [MF14, SHM14, GKC15, BKB15]. How the relatively small HSC population generates more than 10^{11} cells of multiple types daily over an organism's lifetime has yet to be fully understood. HSCs are defined primarily by their function but are often quiescent [SW10b]. *In vivo*, it is difficult to track the dynamics of individual HSCs, while HSCs *in vitro* do not typically proliferate or differentiate as efficiently. Therefore, the dynamics of HSCs can be inferred only from analyses of populations of progenitors and differentiated blood cells [BVZ12] and it is useful to investigate HSC dynamics through mathematical modeling and simulations [SBM14, SM11, HR16].

While most studies model population-level HSC behavior [SKL07, ZLM12, BKB15], certain aspects of HSCs, such as individual-level heterogeneity in repopulation and differentiation dynamics, have to be studied on a single-cell or clonal level [KKP14]. Single HSC transplant mouse data [SRM11] and clonal tracking of HSCs [CBE12, SRC14] in mice have shed some light on repopulation dynamics under homeostasis and after bone marrow transplantation [MSB12, VBZ13, BKB15]. How each individual HSC contributes to the blood production process over long times in much larger human and non-human primates is less clear and more difficult to study [DNL12]. Human clinical data may not be ideal for the purpose of studying normal HSC repopulation as they are usually collected from diseased people.

Recently, results of a long-term clonal tracking of hematopoiesis in normal-state rhesus macaques has been made available [KKP10,KKP14]. The experiment extracted and uniquely "labeled" hematopoietic stem and progenitor cells (HSPCs) from four rhesus macaques with viral tags that also carry an enhanced green fluorescent protein gene. After autologous transplantation, if any of the tagged HSPCs divide and differentiate, its progeny will inherit their unique tags and ultimately appear in the peripheral blood. All cells that share the same tag form a "clone". Blood samples were drawn every few months over $4-14$ years (depending on the animal) and the sampled cells were counted and sequenced. Of the $\sim 10^6-10^7$ unique HSC tags transplanted, $\sim 10^2-10^3$ clones were detected in the sampled peripheral blood. In the original paper describing the clonal tracking experiment, Kim *et al.* [KKP14] observed "A small fraction $(4-10\%)$ of tagged clones predominately contribute to a large fraction $(25-71\%)$ of total blood repopulation." They described the fluctuations of tags that appeared in each sample as "waves of clones", but did not address why some clones can disappear at certain times and reappear in a latter sample.

To get a more quantitatively understanding of the data, it is necessary to have sufficient counts of cells that carry each label. In this experiment, blood cell numbers were counted by identifying and enumerating the number of DNAs that correspond to specific vector-cellular DNA junctions in the samples by PCR and next-generation-sequencing (NGS). The exact counts of DNAs reflect the cell numbers of specific clones and are significant for the subsequent quantitative analysis. However, the true counts of DNAs obtained from NGS technology are often underestimated because of certain sequencing errors (e.g. insertion, deletion, substitution). This results in failure of sequence alignment to the genome, which can nonetheless be partially recovered by computational tools. For the current NGS data generated by 454 pyrosequencing, the main source of reading errors is the homopolymer indel (insertion and deletion) error, a common type of error in 454 sequencing and the newer technology Ion Torrent [WHL14]. My first goal is to more accurately recover the true counts of sampled DNAs with each specific VIS tag by developing a computer algorithm and integrating it into a established data-processing pipeline. In Chapter 2, I will describe how the algorithm automatically handles various types of sequencing errors, with a special

consideration for the homopolymer errors, by comparing similarities of DNA sequences of different alignment qualities.

After obtaining enhanced data of cell counts for each clone, I seek to better understand the observed clone size distributions and the large temporal variability in clonal populations. To address these observations, I ask: Is heterogeneity in HSCs necessary for peripheral blood clone size heterogeneity, or can a neutral model explain most of the observed differences among clones? Are clones that disappear and reappear from sample to sample simply missed by random blood sampling, or do other mechanisms of temporal variability need to be invoked? Unlike previous models that describe the evolution of lineages of different cell types and their regulation [SBM14, SM11, HR16, HBK16], we will consider simpler neutral models that describe the birth-death-migration dynamics of specifically granulocyte populations carrying different tags. The model will be multi-stage (or multi-compartment) in nature, consisting of the stem cell stage, the progenitor stage (both in the bone marrow), and the mature stage (in the peripheral blood). By a "neutral" model, I assume the tagging operation does not affect cell dynamics and all cells in the same stage are identical. Of central interest is the competition among the thousands of clones under a neutral environment that gives rise to fluctuations, extinctions, and resurrections in individual clone populations. I will focus on the granulocyte lineage (mostly neutrophil) because they have relatively simple dynamics in contrast to other lineages. Mature granulocytes have a relatively short lifespan (hours to days) and do not further proliferate in peripheral blood. Thus, their production in the peripheral blood comes only from HSCs and their sampled counts are good indicators for recent activities of the upstream stem and progenitor cells. In Chapter 3, I will describe the multi-stage model and further apply asymptotic analysis to reduce the model complexity. I then propose a novel statistical feature for each individual clone that allows robust inference of key parameters from the data.

During the study of the multi-stage clonal dynamics, I consider a "clone count" statistics in a single stage. For example, this can be the stage of progenitor cells with birth, death, and immigration (from the stem cell pool). However, limited proliferation is not included here while the limitation of the bone marrow space is modeled by a global capacity capac-

ity on the progenitor cell population. Also, unlike the usual cell count statistics for each tagged clone, it counts the number of clones of a specific size. Thus the clone identity is not relevant in this description. Such description is particularly useful for describing how a distribution of clone sizes arises from the initial same-sized clones under stochastic neutral dynamics. It was widely applied to characterize systems of various scales, including gene-barcoded, virally tagged, or TCR-decorated [JHH13, ZES13, QLC14, GKC15, DMW16] cellular clones, microbial populations [HWH03, HBJ06], and ecological species [Hub01, MEG07, GT05]. Most previous theoretical studies of the birth-death process (even with arbitrarily time-dependent birth and death rates [Ken48a]) explicitly or implicitly assumed mean-field models [MEG07] which predicts a hyperbolical power-law-like distribution: Many small clones, a few intermediate-sized clones, and one or several large clones. However, when simulating the clonal dynamics under zero or very low immigrations, there always emerges one clone that acquires an unexpectedly large size in each simulation. As immigration rate increases, this phenomenon disappears and the clone sizes return to the usual hyperbolical distribution. After carefully analyzing the simulated clonal dynamics, I find that this large clone is induced by the genetic-drift-like effect of random birth/death and global carrying capacity (or density dependence) which creates an additional local stable state of the system to have a large-size clone. This latter regulatory effect is weak when initially most clones are small, but gradually becomes very strong for growing clones. It also disappears as immigration rate increases which brings a phase transition to the system. In Chapter 4, by approximating this regulatory effect with a Moran-type fixed-total-population constraint and employing the analytical tool of energy landscape, I am able to find a good approximation for the clone-count distribution and unravel the mechanistic reason for the failure of the mean-field assumption.

In sum, I hope this computational algorithm for the data quality improvement, mathematical modeling of the multi-stage clonal dynamics, and theoretical analysis of the phase transition induced by global carrying capacity (density dependence) will advance researches of high-dimensional stochastic systems in clonal tracking, stem cell biology, immune diversity, ecology, and other related fields.

CHAPTER 2

Clone Size Data Augmentation

2.1 Background

Integration Site (IS) assays, in combination with next-generation sequencing, were used as a cell-tracking tool to characterize progenies of stem cells that share the same IS. For the accurate comparison of repopulating stem cell clones within and across different samples, the detection sensitivity, data reproducibility, and high-throughput capacity of the assay are among the most important assay qualities [SXX17]. Aiming at reducing the sequencing errors, especially those homopolymer-induced ones, in a previously established data pipeline, I will develop an algorithm that tries to improve the quality of the output sequence data.

A homopolymer segment is a string of identical nucleotides appearing in a row, e.g. 'AAAA' (homopolymer of 'A's of length 4). In 454 sequencing, each homopolymer segment is called in a single flow indicated by a light signal. The brightness of the light indicates the length of the homopolymer. When the same nucleotide appears several times in a row, it may be hard to distinguish the exact brightness of the signal, resulting in erroneous measurement of the homopolymer length. Such homopolymer errors lead to extra indels (insertion and deletion) in the sequence reads and affect the alignments of sequences to the genome data. The true counts of sequences identified for specific clones will be affected. In a previous study, Huse *et al.* showed that insertion, deletion, and substitution errors caused by homopolymers have error proportions of 20%, 9% and 10%, respectively [HHM07].

In the literature, homopolymer errors were commonly handled by two types of approaches. One type is merges sequences that are "similar enough" to each other according to a certain similarity (distance) measure. A straightforward measure is the Levenshtein

distance [Lev66], which calculates the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one sequence into the other. A similar measure is the Needleman-Wunsch distance [NW70], which maximizes the similarity between two sequences instead of minimizing the edit distance between sequences. It was shown that the two approaches are equivalent [Sel74]. The frequently used Smith-Waterman algorithm [SW81] for a local sequence alignment, however, looks for the region of highest similarity between two proteins without aligning the entire lengths (especially the ends) of two sequences. For the current purpose of finding sequences that carry the same VIS information, matching sequence ends is essential because sequences with the same VIS are supposed to start from the same gene site.

Based on the Levenshtein distance, an intuitive way to handle the more frequent homopolymer errors is to assign a smaller penalty score (e.g. a positive value between 0 and 1) to the homopolymer-induced indel errors than that of the regular indels (with penalty 1). However, it is unclear what penalty score should be chosen. Some researches [HHV12] chose to treat it as normal indel error (assign penalty 1) or completely ignore the homopolymer-length information rather than use it. For example, some algorithms filter CATAAAG as CATAG [Led12], where the homopolymer length information is lost. This can lead to overestimation of sequence difference when there is a substitution in the homopolymer segment, for example TTTC (filtered as TC) and TCTC (filtered as TCTC). Other options include coding CATAAAG as C1A1T1A3G1 (nucleotide + homopolymer lengths) or as $C^1A^1T^1A^3G^1$ (A^1 and A^3 are considered as different bases) [Led12]. This treatment does not necessarily decrease the computational complexity of the distance algorithm and the algorithm will have to be modified substantially for the extra set of characters.

A different type of approach corrects homopolymer errors by extracting statistical information from the sequence data itself [QLD11, BSI12, WHL14]. This operation is often performed before mapping the raw sequences to the genome database. The approach establishes a spectrum of trusted k-mers (k consecutive bases) from the input dataset and then modifies each sequence so that it only contains k-mers from the spectrum. However, there is no guarantee that more frequent k-mers represent true ones for all DNA segments.

DNA segments of rare clones may be incorrectly modified which can lead to failure of its alignment to the genome. For the current purpose of identifying clones that carry certain gene segments of the host genome, finding authentic DNA segments is more important. Such frequency-based error correction approach may not be an ideal choice.

In this Chapter, I will describe a Levenshtein distance-based algorithm with a length-dependent penalty for homopolymer-induced errors to calculate the similarities among sequences that were successfully pre-aligned to the reference genome. I use multiple rules to elect a best candidate sequence and merge the counts of similar sequences (whether they are successfully pre-aligned or not) to its count. This way, the total valid data are augmented.

2.2 Materials and Methods

2.2.1 Raw data and the original pipeline

The experimental protocol to generate data is described in [KKP10]. Directly collected data from the experiment include: sequence data, search motifs for vector and linker sequences, and restriction enzyme information [SXX17]. The raw data contain around 10^7 DNA segments, which has undergone the following existing computational pipeline: (1) VIS authentication. The sequences were first demultiplexed and trimmed of the vector, linker, and primer sequences and then mapped onto the reference genome using a BLAST-like alignment tool. This way, information of whether there is a single, multiple, or no 'hit' is obtained. (2) Sequence enumeration of unique VIS. Sequences were clustered into groups based on their similarities (95%) of the processed sequences. It was roughly estimated that homopolymer errors fail this step by about 26%. A three-step treatment was applied: using a lower stringency condition such as 90%, electing a "most likely" genome sequences in a each group re-clustering the sequences based on a 95% threshold. The above process was then repeated by using an even lower threshold 80%. However, this existing protocol is not able to treat homopolymer error separately and may include quite a few non-homopolymer 'fake' sequences in the VIS counts while still missing many homopolymer-related 'genuine'

sequences. For this reason, manual efforts were then paid to achieve reliable final results.

2.2.2 Multi-rule score

Instead of using only sequence frequency information, we use the following information of a sequence as criteria to evaluate the quality of the read: mapping type, initial counts (frequency), and read length. There are three mapping types of sequences obtained by using BLAT and GMAP algorithm [KKP10]: Single, which means that the sequence was mapped to a single site on the genome; MultiHits, which means that the sequence was mapped to multiple genome sites; NoHits, which means that the sequence was not mapped to any genome site. The initial counts were obtained by simply counting the number of mapped sequences and those similar to the mapped ones by treating homopolymer indels as regular indels. The read length of a sequence is between 25 bp (below which the sequence read is considered unreliable) and 500 bp (above which the sequencing efficiency dropped significantly [KKP10]). Generally, the longer the read is, the more reliable the sequence is considered to be.

2.2.3 Length-based homopolymer penalty

To calculate the distances between any two sequences, the basic idea is to use a modified Levenshtein algorithm with an adaptive penalty for the homopolymer indels. One concern about the Levenshtein-based method is the time complexity; based on the current need of merging $10^3 - 10^4$ unmapped sequences to $10^2 - 10^3$ mapped sequences (representing several hundreds of distinguishable clones) and some obvious optimization techniques, it is a realistic approach. We collect indel mismatches among 95% similar sequences from the post-alignment data. This way, we get a rough sense of the occurrence rates for different types of errors. Although it is likely that some errors are not counted when some genuine copies are below 95% similarity from the reference segment, relative ratios among different types of errors should not be affected much. In Figure 2.1(a), the total homopolymer indel rate is about three times that of the normal indel, we thus set the penalty score for homopolymer

indel as $\frac{1}{3}$. By only focusing on the homopolymer segments of these sequences, we plot Figure 2.1(b) to show how the error rate increases with the homopolymer length. Intuitively it is more likely to have homopolymer errors in a segment 'AAAAA' than in 'AA'. When the homopolymer length is larger than 7, there is more than 50% probability that more than 1 indel errors will occur, so we set the penalty in this region to be 0.

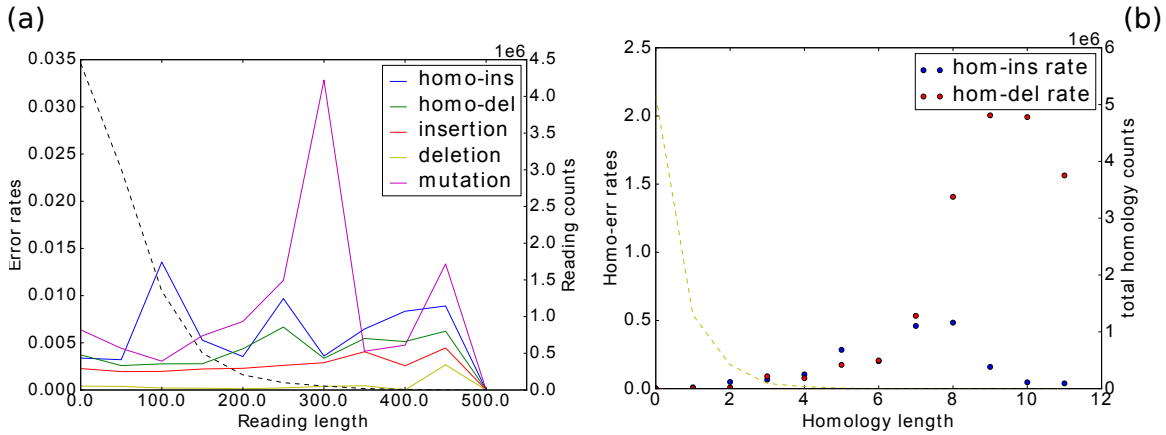

Figure 2.1: Homopolymer error rates

(a) Rates of different error types as a function of reading length. The average rates are 0.0049 for homopolymer insertion, 0.0027 for homopolymer deletion, 0.0036 for regular insertion, 0.0002 for regular deletion, and 0.0054 for mutation. (b) Homopolymer error rates as a function of homopolymer length.

2.2.4 Similarity threshold

We assume two situations that sequences might be similar to each other: First, a sequence is a 'genuine' copy of a genome segment with a small number of mutations, insertions, or deletions. Second, a sequence is a 'fake' copy of a genome segment with a substantial number of overlapping nucleotides as a result of the functional redundancy of primate genome. Ideally, 'fake' sequences from the second group is farther away than the 'genuine' ones from the first group. Our aim is to find a proper threshold that can distinguish the two groups of similar sequences. For this purpose, we sampled a few genome sequences that were successfully mapped and plot their distances with all raw sequences in Figure 2.2. It seems that as long

as we choose a threshold between $92\% - 96\%$ (currently we choose 95%), 'genuine' and 'fake' copies of the genome segments can be well separated.

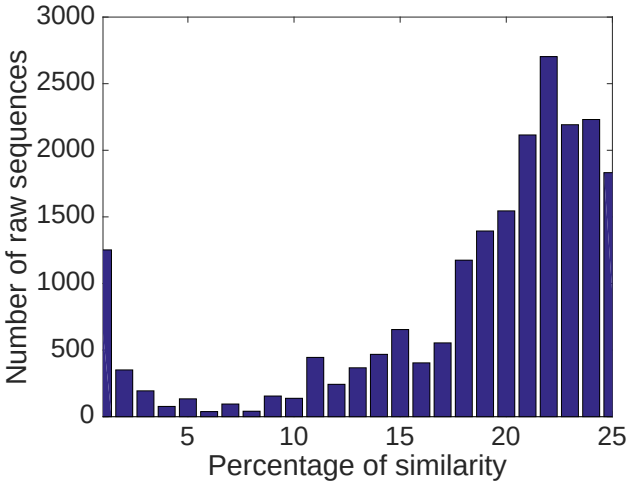

Figure 2.2: Similarity distribution among sequences

Pairwise sequence distance distribution between 150 randomly selected genome sequences and 16384 raw sequences from the experiment.

2.2.5 Pseudo-code and examples

(0) Input filename

(1) Pre-processing e.g. eliminate length < 25 bp sequences, delete reads of 'N', cut the piece with linker sequences, etc.

(2) If want to achieve time-efficiency and care only about recovering the counts of mapped sequences, go to (2.1); otherwise (also recover the counts of unmapped sequences), go to (2.2).

(2.1) Unsupervised clustering

(2.1.1) Calculate the pair-wise similarities of all sequences

(2.1.2) Group all sequences (e.g. if A=B, B=C, but A$\neq$C, still group {A, B, C}), output grouping results

(2.1.3) Rank sequences and elect the highest rank one as the representative

(2.1.4) If want to increase sensitivity/decrease specificity, go to (2.1.4.1), otherwise go to

(2.1.4.2)

(2.1.4.1) Inclusive merging Example: If INCLUSIVE is True, we have:

...	...	...	...	...
10	Single	-	98	CTAGGAAAACGA−TTATAGCTGCACAAAC−A−CTTT−GT−CTCG...
10	R|NoHits775		149	CTAG−AAA−−GA−TTATAGCTGCACAAACGA−CTTT−GT−CTCG...
10	R|NoHits795		41	CTAGGAAA−−GAATTATAGCTGCACAAAC−AACTTTTGTTCTCG
10	R|NoHits786		154	CTAGGAAAA−GAATTATAGCTGCACAAAC−A−CTTTTGTTCTCG...
...	...	...	...	...

(2.1.4.1.1) Merge all other sequences in a group into the representative one

(2.1.4.1.2) Output merged sequences

(2.1.4.2) exclusive merging Example: If INCLUSIVE is False, we have (NoHits775 got separated out):

...	...	...	...	...
10	Single	-	98	CTAGGAAAACGA−TTATAGCTGCACAAAC−A−CTTT−GT−CTCG...
-10	R|NoHits775		149	CTAG−AAA−−GA−TTATAGCTGCACAAACGA−CTTT−GT−CTCG...
10	R|NoHits795		41	CTAGGAAA−−GAATTATAGCTGCACAAAC−AACTTTTGTTCTCG
10	R|NoHits786		154	CTAGGAAAA−GAATTATAGCTGCACAAAC−A−CTTTTGTTCTCG...
...	...	...	...	...

(2.1.4.2.1) In each group, for those 95% similar to the representative, merge to the representative; otherwise, separate and mark the unqualified sequences

(2.1.4.2.2) Output sequences

(2.2) Supervised clustering

(2.2.1) Separate master (MultiHits/Single) sequences and slave (NoHits) sequences

(2.2.2) Calculate the pair-wise similarities between masters and slaves

(2.2.3) for those slaves that is similar to a unique master, merge and output

(2.2.4) for those that is similar to multiple masters, if randomly assign to a sequence, go to (2.2.4.1); if output information, go to (2.2.4.2)

(2.2.4.2) Mark and output

Example: When RANDOM ASSIGN is False, the suspicious NoHits2743 is separated and

marked, merging to neither Single 197 nor Single 238:

...	...	...	...	...
197	Single	+	411	GTTAGTAGAGACAGGGTTTCACCATGTTGGCCAGGCTGG...
...	...	...	...	...
238	Single	-	112	GTTAGTAGAGACAGGGTTTCACCATGTTGGCCAGGCTGG...
...	...	...	...	...
197 238	R|NoHits2743		39	GTTAGTAGAGACAGGGTTTCACCATGTTGGCCAGGCTCG
...	...	...	...	...

(2.2.4.1) Randomly merge to a similar master, output Example: When RANDOM ASSIGN is True: NoHits2743 is merged to Single 197.

(2.2.5) for those slaves that is similar to no master, directly output

2.2.6 Detailed settings

Homopolymer penalty (use penalty 1/6 as an example). The penalty value for homopolymer errors

Example: If HOMO PENALTY is 1/6, the distance of the following two Seqs is $1/6+1/6+1/6+1=1.5$; if HOMO PENALTY is 1 (same as normal indel), the distance is 4.

Seq 1: TATACTTGGGGCTATTT−CACAAT−GGAAATAATTAGCCCGT

Seq 2: TATACTTGGG−CTATTTTCACAATTGGAAATAATT−GCCCGT

End gap (default: False). Whether to consider the end gap when comparing two sequences

Example: If END GAP is False, the following two Seqs are considered the same. Otherwise, they are not considered the same:

Seq 1: GCCTGCCCCCCGAGCTCTCCCGTGTGGATCCCGCA

Seq 2: GCCTGCCCCCTGAGCTCTCCCGTGT

Similarity Threshold (default: 0.95, range: [0, 1]). The similarity threshold for two sequences to be considered the same.

Example: (With HOMO PENALTY 1/6, END GAP is False) The distance of the following

two Seqs is $1/6 + 1 + 1 + 1/6 = 2.33$, the similarity is $1 - 2.33/40 = 0.942 < 0.95$; they are not considered the same:

Seq 1: GTAG−AGCTTTTAC−ATACTTACAGGCATATGCACAG−CAA−TC

Seq 2: GTAGGAGCTTTTACTATACTTACAGGCATATGCACAGACAAAGT

Score Weights (default: [0.5, 0.3, 0.2], range: $0 \leq p_i \leq 1$, $\sum p_i = 1$). The importance weights among different rules, including hit type, post-alignment count, and sequence length.

Hit-type Weights (default: [0.70, 0.25, 0.05], range: $0 \leq p_i \leq 1$, $\sum p_i = 1$), The importance weights among various hit types, including MultiHits, Single, and NoHits.

Clustering strategy (default: exclusive). After the first step of calculating all pairwise distances between mapped and unmapped sequences, we consider two sequences equivalent if they are of $\geq 95\%$ similarity with each other. However, this may bring inconsistency when extending the criterion to group more than two sequences, since essentially this definition of equivalence is not transitive. Specifically, if we use the equal sign A = B to denote that two sequences A and B are equivalent, we have a problem when A = B, B = C but A $\neq$ C. The sources of this inconsistency include the intrinsic redundancy nature of different DNA segments, the shortness of the reading length of sequences, and the impossibility to perfectly capture the true variations of different sequences with the same set of parameters (e.g. similarity threshold) across all sequences. We give the option of considering A=C (inclusive strategy) and considering A≠C (exclusive strategy) when grouping sequences. These two options give the upper and lower bounds for the estimate of counts for sequence A.

2.3 Results

Performance We checked how many valid VISs were recovered by performing our homopolymer algorithm. We also checked our results with the eye-balling results obtained by three lab technicians using three weeks. In comparison, our computation tool used less than

20 minutes on a Thinkpad laptop (with 4GB memory and Intel i5@1.6GHZ CPU).

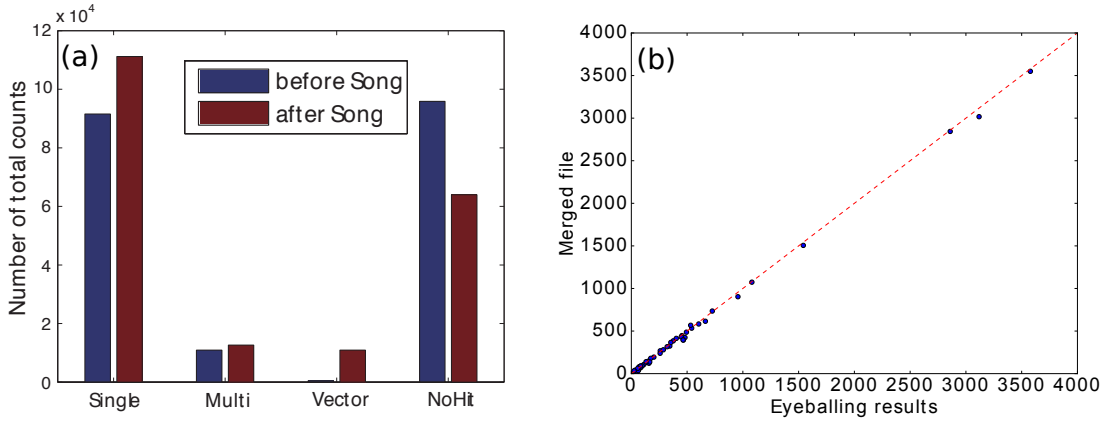

Figure 2.3: Homopolymer-error correction results

(a) Counts recovery by applying the homopolymer algorithm to the original pipeline. Single count increases by 23%. NoHits count decreases by 30%. (b) Sequence-by-sequence count comparison between the homopolymer algorithm and the eye-balling results by lab technicians.

Runtime We plot the runtimes of the algorithm for three datasets on three computing systems in Figure 2.4 for both the supervised (only merge unmapped sequences to mapped sequences) and unsupervised (merge among all sequences) strategies: Thinkpad (Memory: 4GB, CPU: Intel i5@1.6GHZ), mac (Memory: 16GB, CPU: Intel i7@2.5GHZ), and lab server (Memory: 64GB, CPU: Intel x86-64@1.2GHZ) for 454 (454 pyrosequencing data that contains 800 NoHits, 312 Single/MultiHits), illu16 (Illumina data that contains 1454 NoHits, 566 Single/MultiHits), and illu17 (Illumina data that contains 726 NoHits, 482 Single/MultiHits).

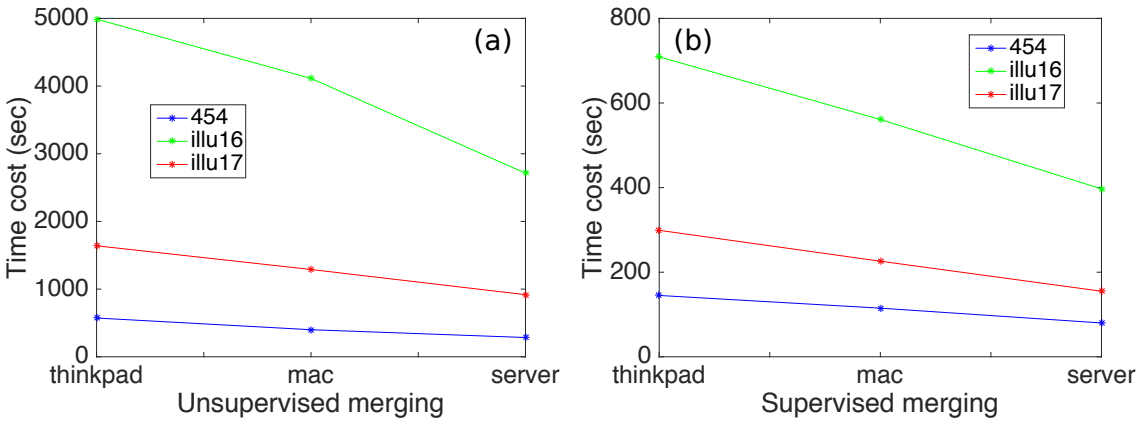

Figure 2.4: Time efficiency of the algorithm

Time cost comparisons of the unsupervised merging strategy (a) and the supervised merging strategy (b) on three datasets in three computing systems.

2.4 Future work

As is pointed out by [KKP14], there exists intrinsic bias in the cell counts of each tag in such VIS-based sequencing. The reason is that PCR show different amplification efficiencies among sequences of different lengths. The next step to improve the quality of the clonal data is to integrate statistical analysis of the PCR bias into the algorithm to compensate the cell counts. Another way of optimization is to apply machine learning techniques to automatically find optimal parameters instead of using the manually assigned ones. It will also be valuable to compare the performance of the algorithm with existing sequence denoising tools on various datasets.

CHAPTER 3

Modeling Clonal Dynamics in Granulopoiesis

3.1 Background

After obtaining enhanced-quality data, we try to infer the upstream stem cell dynamics in the bone marrow from the downstream data sampled from the peripheral blood. We will first focus on the clonal dynamics in the granulocyte lineage. Even when considering only one cell type, realistic mathematical models may need to include complex multilevel biochemical feedback mechanisms of regulation [CPG08, AC09, SBM14, HMJ15, ORK03, MSH09, MTB13]. Many mechanisms may contribute to temporal fluctuations, including extrinsic noise and heterogeneity of HSCs, progenitors, or mature granulocytes. Large time gaps between samplings and small sample sizes also add to the uncertainty of the underlying dynamics. In order to feasibly compare with experimental data, our modeling philosophy will be to recapitulate these complexities into simple, effective models and infer parameters that subsume some of these regulatory effects. This approach and level of modeling are similar to those taken by *e.g.*, Yang, Sun, and Komarova [SK12, YSK15].

Trying to infer all possible mechanisms and associated parameters from the experimental data would essentially be an overfitting problem. After careful consideration of a number of key physiological mechanisms, we hypothesize that stochastic HSC self-renewal, generation-limited progenitor cell proliferation, and sampling frequency statistics provide the simplest reasonable explanation for the observed clonal size variability and large temporal fluctuations. HSCs that are generated from self-renewal of the founder population share the same tag as their founder HSC. Thus, during intense self-renewal after myeloablative treatment and HSPC transplantation, each originally transplanted HSCs begets a clonal HSC subpop-

ulation. Subsequently, heterogeneous clone sizes are stochastically generated even though each tag was initially represented by only a single cell. These expanded HSC clones then go on to repopulate the clones in the progenitor and mature blood population, which are also distinguishable by their corresponding tags.

Relative to HSCs, progenitor cells have limited proliferative potential that can explain the apparent extinctions of clones. This limited proliferation potential can be thought of as an "aging" process. Different types of aging, including organism aging [AC09, GC16, CG16], replicative senescence of stem cells [MSW09], and generation-dependent birth and death rates, have been summarized by Edelstein *et al.* [EIL01]. Here, the clonal "aging" mechanism we invoke imposes a limit to the number of generations that can descend from each newly created (from HSC differentiation) "zeroth generation" progenitor cell. Possible sources of such a limit include differentiation-induced loss of division potential [BBM03] and telomere shortening (as in the Hayflick limit) [RBK99, Hod99, Mil00]. Mathematically, genealogical aging can be described by tracking cell populations within each generation. After a certain number of generations, progenitor cells of the final generation stop proliferating and can only differentiate into circulating mature cells or die.

In the following sections, we first present the mathematical equations and corresponding solutions (whenever possible) of a model that incorporates the above processes. We then develop a new statistical measure that tracks the numbers of absences of clones across the samples. Measured clone abundances of animal RQ5427 are statistically analyzed within our mechanistic model to infer estimates for key model parameters. The data and corresponding statistical analyses for animals 2RC003 and RQ3570 are also provided in the Results section.

3.2 Materials and Methods

Below, we describe available clonal abundance data, mechanistic models, and a statistical model we will use for parameter inference.

3.2.1 Clone abundance data

In the experiments of Kim *et al.* [KKP14], cells in samples of peripheral blood were counted to extract $\hat{S}^+(t_j)$, the total number of EGFP+ tagged cells in sample $1 \leq j \leq J$ taken at time t_j. After PCR amplification and sequencing, $\hat{f}_i(t_j)$, the relative abundance of the i^{th} tag among all sampled, tagged cells is also quantified. The "$\hat{}$" notation will henceforth indicate experimentally measured quantities.

Within mature peripheral blood, lymphocytes such as T cells and B cells proliferate or transform in response to unpredictable but clone-specific immune signals [DP13]. They also vary greatly in their lifespans, ranging from days in the case of regular T and B cells to years in the case of memory B cells. On the other hand, mature granulocytes do not proliferate in peripheral blood and have relatively shorter life spans [BVZ12]. Granulocyte dynamics can thus be analyzed with fewer confounding factors [SKL07]. Thus, in this work, we restrict our analysis to granulocyte repopulation and extract all variables, including $\hat{S}^+(t_j)$ and $\hat{f}_i(t_j)$ described above, that are associated exclusively with granulocyte populations.

In Figure 3.1(a), we plot the total numbers of sampled granulocytes from one of the macaques, RQ5427. The subpopulation of EGFP+ granulocytes and the subset of EGFP+ granulocytes that were extracted for PCR amplification and analysis are also plotted. Data for two other animals, 2RC003 and RQ3570, are qualitatively similar, while those for the fourth animal, 95E132, did not separate granulocytes from peripheral blood mononuclear cells. As shown in Figure 3.1(b), not only are the clone abundances $\hat{f}_i(t_j)$ heterogeneous, but individual clone abundances vary across samples taken at different times. The variation is so large that many clones can go extinct and reappear from one sample to another, as shown in Figure 3.1(c). Since large numbers of progenitor and mature cells are involved in blood production, the observed clone size fluctuations cannot arise from intrinsic demographic stochasticity of progenitor- and mature-cell birth and death. Moreover, we will show later in the Results section that random sampling alone cannot explain the observed clonal variances and mechanisms that involve other sources of variation are required.

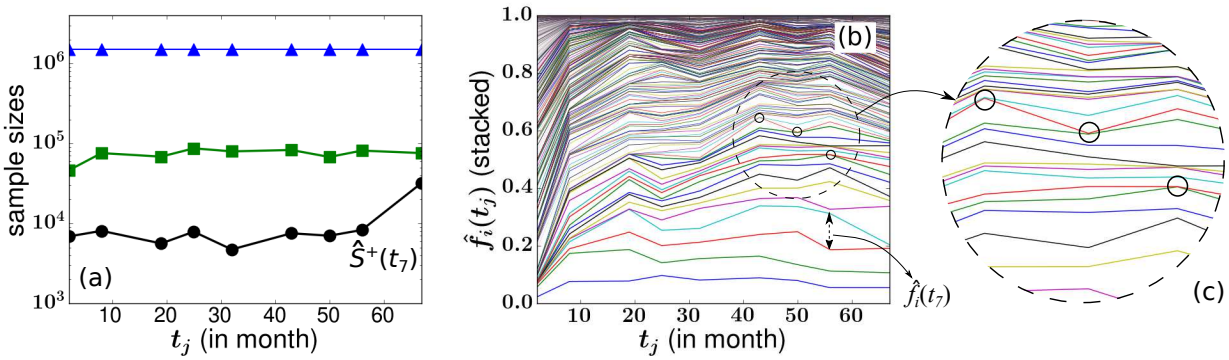

Figure 3.1: Blood sample data from animal RQ5427

(a) The total numbers of sampled granulocytes (blue triangles), EGFP+ granulocytes (green squares), and the subset of EGFP+ granulocytes that were properly tagged and quantifiable were extracted for PCR amplification and analysis (black circles) [KKP14]. This last population defined by $\hat{S}^+(t_j)$ is used to normalize clone cell counts. We excluded the first sample at month 2 in our subsequent analysis so, for example, the sample at month 56 is labeled the 7th sample. There were 536 clones detected at least once across the eight samples taken over 67 months comprising an average fraction 0.052 of all granulocytes. The abundances of granulocyte clones are shown in (b). The relative abundance $\hat{f}_i(t_j)$ of granulocytes from the i^{th} clone measured at month t_j is indicated by the vertical distances between two adjacent curves. The relative abundances of individual clones feature large fluctuations over time. "Extinctions" followed by subsequent "resurrections," were constantly seen in certain clones as indicated by the black circles in (b) and in the inset (c).

3.2.2 Nomenclature and lumped mechanistic model

Figure 3.2 depicts our neutral model of hematopoiesis which is composed of five successive stages, or compartments, describing the initial single-cell tagged HSC clonal populations immediately after transplantation (Compartment **0**), the heterogeneous HSC clonal populations after a short period of intense self-renewal (Compartment **1**), the transit-amplifying progenitor cell compartment (Compartment **2**), the peripheral blood pool (Compartment **3**), and the sampled peripheral blood (Compartment **4**), respectively. Each distinct color or shape in Figure 3.2 represents a distinct clone of cells with the same tag.

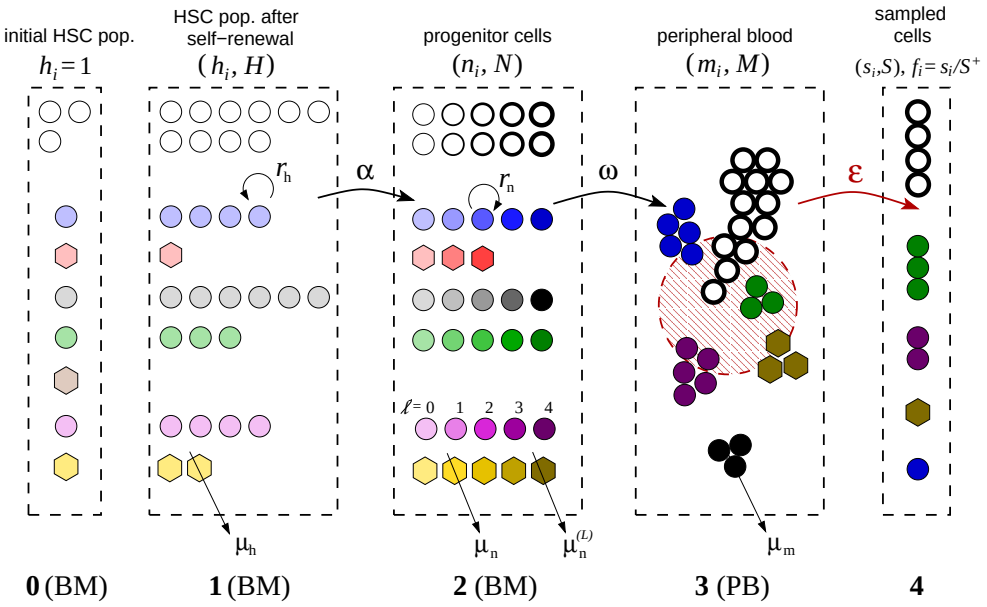

Figure 3.2: Multi-stage model of hematopoietic clones

Schematic of a neutral multi-stage or multi-compartment hematopoiesis model. BM and PB refer to bone marrow and peripheral blood, respectively. Cells of the same clone have the same color. White circles represent untagged cells which were not counted in the analysis. Stages **0**, **1**, and **2** describe cell dynamics that occur mainly in the bone marrow. Stage **1** describes HSC clones ($C_\mathrm{h} = 6$ in this example) after self-renewal that starts shortly after transplantation with rate r_h. After self-renewal, the relatively stable HSC population ($H^+ = 20$ in this example) shifts its emphasis to differentiation (with per-cell differentiation rate α). Larger clones in Stage **1** (*e.g.*, the circular blue clone, $h_\mathrm{blue} = 4$) will have a larger total differentiation rate αh_blue while smaller clones (*e.g.*, the red hexagonal clone, $h_\mathrm{red} = 1$) will have smaller αh_red. The processes of progenitor-cell proliferation (with rate r_n) and maturation (with rate ω) in Compartments **2** and **3** are considered deterministic because of the large numbers of cells involved. The darker-colored symbols correspond to cells of later generations. For illustration, the maximum number of progenitor-cell generations allowed is taken to be $L = 4$. Compartment **4** represents a small sampled fraction ($\varepsilon(t_j) \approx 2.8 \times 10^{-5} - 2 \times 10^{-4}$) of Compartment **3**, the entire peripheral blood of the animal. In the example pictured above, $C_\mathrm{s} = 4$. Such small samples can lead to considerable sampling noise but is not the key driver of sample-to-sample variability.

In each compartment, relevant parameters include (using Compartment **1** as example): the total cell count $H(t)$, the untagged cell count $H^-(t)$, the tagged cell count $H^+(t)$, the total number of tagged clones $C_{\rm h}(t)$, and the number $h_i(t)$ of HSCs carrying the $i^{\rm th}$ tag. These quantities are related through $\sum_{i=1}^{C_{\rm h}} h_i(t) = H^+(t) \equiv H(t) - H^-(t)$.

In the progenitor pool, the total number of cells and the number with tag i are denoted $N(t)$ and $n_i(t)$, respectively. Further resolving these progenitor populations into those of the $\ell^{\rm th}$ generation, we define $N^{(\ell)}(t)$ and $n_i^{(\ell)}(t)$. In the mature granulocyte pool, the total granulocyte population and that with tag i are labeled $M(t)$ and $m_i(t)$. In the sampled blood compartment, we use $S(t_j)$, $S^+(t_j)$, $s_i(t_j)$, and $C_{\rm s}(t_j)$ to denote, at time t_j, the total number of sampled cells, the number of tagged sampled cells, the total number of tagged cells of clone i, and the total number of clones in the sample, respectively. In Compartment **4**, we further define $f_i(t_j) = s_i(t_j)/S^+(t_j)$ to denote the relative abundance of the $i^{\rm th}$ clone among all tagged clones.

By lumping together all clones (tagged and untagged) in each compartment, we can readily model the dynamics of total populations in each pool. After myeloablative treatment, the number of BM cells, including HSCs, is severely reduced. Repopulation of autogolously transplanted HSCs occurs quickly via self-renewal until their total number $H(t)$ reaches a steady-state. The repopulation of the *entire* HSC population and the subsequent entire progenitor and mature cell populations may be described via simple deterministic mass-action growth laws

$$\frac{{\rm d}H(t)}{{\rm d}t} = (r_{\rm h}(H(t)) - \mu_{\rm h})H(t), \tag{3.1}$$

$$\frac{{\rm d}N^{(\ell)}(t)}{{\rm d}t} = \begin{cases} \alpha H(t) - (r_{\rm n}^{(0)} + \mu_{\rm n}^{(0)})N^{(0)}(t), & \ell = 0, \\ 2r_{\rm n}^{(\ell-1)}N^{(\ell-1)}(t) - (r_{\rm n}^{(\ell)} + \mu_{\rm n}^{(\ell)})N^{(\ell)}(t), & 1 \le \ell \le L-1, \\ 2r_{\rm n}^{(L-1)}N^{(L-1)}(t) - (\omega + \mu_{\rm n}^{(L)})N^{(L)}(t), & \ell = L, \end{cases} \tag{3.2}$$

$$\frac{{\rm d}M(t)}{{\rm d}t} = \omega N^{(L)}(t) - \mu_{\rm m}M(t). \tag{3.3}$$

HSC self-renewal is a regulated process involving signaling and feedback [MCT02, SW10a, AC09, CPG08, HMJ15] and $r_{\rm h}$ may be a complicated function of many factors; however,

we will subsume this complexity into a simple population-dependent logistic law $r_{\rm h}(H(t))$ and assume a constant death rate $\mu_{\rm h}$. Alternatively, other studies have employed Hill-type growth functions [ZLM12, SK12].

We assume the per cell HSC differentiation rate α is independent of the tag and that differentiation is predominantly an asymmetric process by which an HSC divides into one identical HSC and one progenitor cell that commits to differentiation into granulocytes. An initial generation-zero progenitor cell further proliferates with rate $r_{\rm n}^{(0)}$, contributing to the overall progenitor-cell population. Subsequent generation-ℓ progenitors, with population $N^{(\ell)}$, proliferate with rate $r_{\rm n}^{(\ell)}$ until a maximum number of generations L is reached. By keeping track of the generation index ℓ of any progenitor cell, we limit the proliferation potential associated with an HSC differentiation event by requiring that any progenitor cell of the final $L^{\rm th}$ generation to terminally differentiate into peripheral blood cells with rate ω or to die with rate $\mu_{\rm n}^{(L)}$. For simplicity, we neglect any other source of regulation and assume α, $\mu_{\rm n}^{(\ell<L)} = \mu_{\rm n}$, $r_{\rm n}^{(\ell)} = r_{\rm n}$ and ω are all unregulated constants.

Our model analysis and data fitting will be performed using clone abundances sampled a few months after transplantation under the assumption that granulopoiesis in the animals has reached steady-state [GKC15] after initial intensive HSC self-renewal. Steady-state solutions of Eqs. (3.1), (3.2) and (3.3) are defined by $H_{\rm ss}$, $N_{\rm ss}^{(\ell)}$, and $M_{\rm ss}$. The first constraint our model provides relates these steady-state populations through

$$M_{\rm ss} = \frac{\omega}{\mu_{\rm m}} N_{\rm ss}^{(L)} = \frac{\omega}{\mu_{\rm m}} \left[\frac{\alpha H_{\rm ss}}{(\omega + \mu_{\rm n}^{(L)})} \left(\frac{2r_{\rm n}}{r_{\rm n} + \mu_{\rm n}} \right)^L \right] \equiv \frac{A_{\rm ss}\beta}{\mu_{\rm m}}, \tag{3.4}$$

where we have defined

$$A_{\rm ss} \equiv \alpha H_{\rm ss}, \quad \text{and} \quad \beta \equiv \frac{\omega}{\omega + \mu_{\rm n}^{(L)}} \left(\frac{2r_{\rm n}}{r_{\rm n} + \mu_{\rm n}} \right)^L \tag{3.5}$$

as the total rate of HSC differentiation and the average number of granulocytes generated per HSC differentiation, respectively. These constraints also hold for the EGFP+ subset (about $5\% - 10\%$) of cells, *e.g.*, $M_{\rm ss}^+ = A_{\rm ss}^+\beta/\mu_{\rm m}$ and $A_{\rm ss}^+ = \alpha H_{\rm ss}^+$. Since $M_{\rm ss}^+$ is inferred from the experiment, Eq. (3.4) places a constraint between $A_{\rm ss}^+ = \alpha H_{\rm ss}^+$ and β. This steady-state constraint will eventually be combined with statistics of the fluctuating clone abundances data to infer estimates for the underlying model parameters.

3.2.3 Clone-resolved mechanistic model

Although the lumped model above provides important constraints among the steady-state populations within each compartment, the clone-tracking experiment keeps track of the populations of sampled granulocytes that arise from "founder" HSCs that carry the same tag. Thus, we need to resolve the lumped model into the clonal subpopulations described by h_i, $n_i^{(\ell)}$, and m_i.

Even though the total HSC populations $H(t)$ and $H^{\pm}(t)$ are large, the total number of clones $C_{\rm h} \gg 1$ in compartment **1** is also large, and the number of cells with any tag (the size of any clone) can be small. The population of cells with any specific tag i is thus subject to large demographic fluctuations. Thus, we model the stochastic population of HSCs of any tag using a master equation for $P(h,t)$, the probability that at time t the number of HSCs of any clone is h:

$$\frac{{\rm d}P(h,t)}{{\rm d}t} = \mu_{\rm h}(h+1)P(h+1,t) + (h-1)r_{\rm h}(H(t))P(h-1,t) - [\mu_{\rm h} + r_{\rm h}(H(t))]hP(h,t). \quad (3.6)$$

Recall that immediately after transplantation, each HSC carries a distinct tag before self-renewal ($h_i(0) = 1$) leading to the initial condition $P(h,0) = \mathbb{1}(h,1)$, where the indicator function $\mathbb{1}(x,y) = 1$ if and only if $x = y$. Because $h = 0$ is an absorbing boundary, clones start to disappear at long times resulting in a decrease in the total number $C_{\rm h}(t)$ of HSC clones. Before this "coarsening" process significantly depletes the entire population, each clone constitutes a small subpopulation among all EGFP+ cells, $h(t) \ll H(t)$, and the stochastic dynamics of the population h of any clone can be approximated by the solution to Eq. (3.6) with $r_{\rm h}(H(t))$ replaced by $r_{\rm h}(t)$. Hence, evolution of each HSC clone follows a generalized birth-death process with time-dependent birth rate and constant death rate. We show below that for $H \gg 1$ the solution to Eq. (3.6) can be written in the form [Ken48b]

$$P(h,t) = (1 - P(0,t))(1 - \lambda(t))\lambda(t)^{h-1}, \quad (3.7)$$

where $0 \leq \lambda(t) < 1$ depends on $r_{\rm h}(t)$ and $\mu_{\rm h}$. Here, $\lambda(t)$ determines "broadness" (level of clone size heterogeneity) of the clone size distribution. For the relevant initial condition of unique tags at $t = 0$, $\lambda(0) = 0$ and $\lambda(t \to \infty) \to 1$. When $\lambda(t)$ is small, the distribution is

weighted towards small h. For $\lambda(t) = 0$, $P(h,t) = \mathbb{1}(h,1)$ which was the limit used in Goyal *et al.* [GKC15] to assume no HSC self-renewal after transplantation. In the limit $\lambda(t) \to 1$, the distribution becomes flat and a clone is equally likely to be of any size $1 \leq h \leq H$.

To further resolve the progenitor population into cells with distinct tags, we define $n^{(\ell)}(t)$ as the number of generation-ℓ progenitor cells carrying any one of the viral tags. The total number of progenitor cells with a specific tag is $n(t) \equiv \sum_{\ell=0}^{L} n^{(\ell)}(t)$. Since the sizes h_i of individual clones may be small, differentiation of HSCs within each clone may be rare. However, since the size of each tagged progenitor clone quickly becomes large ($n(t) \gg 1$), we model the dynamics of $n^{(\ell)}(t)$ using deterministic mass-action growth laws:

$$\frac{\mathrm{d}n^{(\ell)}(t)}{\mathrm{d}t} = \begin{cases} \text{Poisson}(\alpha h(t)) - (r_\mathrm{n} + \mu_\mathrm{n}) n^{(0)}(t), & \ell = 0, \\ 2r_\mathrm{n} n^{(\ell-1)}(t) - (r_\mathrm{n} + \mu_\mathrm{n}) n^{(\ell)}(t), & 1 \leq \ell \leq L-1, \\ 2r_\mathrm{n} n^{(L-1)}(t) - (\omega + \mu_\mathrm{n}^{(L)}) n^{(L)}(t), & \ell = L. \end{cases} \tag{3.8}$$

Our model is neutral (all clones have the same birth, death, and maturation rates), so these equations are identical to Eqs. (3.2). However, since creation of the zeroth-generation subpopulation $n^{(0)}(t)$ derives only from differentiation of HSCs of the corresponding clone, which has a relatively small population $h(t)$, we invoke a Poisson process with rate $\alpha h(t)$ to describe stochastic "injection" events associated with asymmetric differentiation of HSCs of said clone. Each discrete differentiation event leads to a temporal burst in $n^{(\ell)}(t)$.

Finally, the dynamics of the population $m(t)$ of any granulocyte clone in the peripheral blood are described by an equation analogous to Eq. (3.3):

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \omega n^{(L)}(t) - \mu_\mathrm{m} m(t), \tag{3.9}$$

where we have assumed that only the generation-L progenitor cells undergo terminal differentiation with rate ω. An alternative model allows progenitor cells of earlier generations ($\ell < L$) to also differentiate and circulate but does not give rise to qualitatively different results.

To study the dynamics of the burst in $n_\mathrm{b}^{(0)}(t)$ immediately following a *single, isolated* asymmetric HSC differentiation event at $t = 0$, we set the initial condition $n_\mathrm{b}^{(0)}(0) =$

1, $n_{\rm b}^{(\ell)}(0) = 0$ $(1 \le \ell \le L)$, remove the Poisson$(\alpha h(t))$ term in Eq. (3.8) and find,

$$n_{\rm b}^{(\ell)}(t) = \begin{cases} \dfrac{(2r_{\rm n}t)^{\ell}}{\ell!} e^{-(r_{\rm n}+\mu_{\rm n})t}, & 0 \le \ell \le L-1, \\ 2r_{\rm n} \displaystyle\int_0^t n_{\rm b}^{(L-1)}(\tau) e^{-\omega(t-\tau)} d\tau, & \ell = L. \end{cases} \tag{3.10}$$

Bounded analytic solutions to $n_{\rm b}^{(L)}(t)$ involving the lower incomplete gamma function can be found. Upon using the solution $n_{\rm b}^{(L)}(t)$ in Eq. (3.9) the mature blood population within a clone associated with a single HSC clone differentiation even is described by

$$m_{\rm b}(t) = \omega \int_0^t n_{\rm b}^{(L)}(\tau) e^{-\mu_{\rm m}(t-\tau)} {\rm d}\tau. \tag{3.11}$$

The populations associated with a single HSC differentiation event, $n_{\rm b}^{(\ell)}(t)$ and $m_{\rm b}(t)$, are plotted below in Figure 3.3 of the Results section. Then $m_i(t)$, the total number of mature cells with the $i^{\rm th}$ tag at time t, is obtained by summing up all $m_{\rm b}(t-\tau_k)$ bursts initiated by HSC differentiations at separate times $\tau_k \le t$ with the $i^{\rm th}$ tag.

Besides the burst dynamics described above, the data shown in Figure 3.1(a) are subject to the effects of small sampling size, uncertainty, and bias induced by experimental processing such as PCR amplification, and data filtering. In this experimental system, PCR generates a smaller uncertainty than blood sampling so we focus on the statistics of random sampling. Each blood sample drawn from rhesus macaque RQ5427 contains about 10μg of genomic DNA [KKP14]. After PCR amplification, deep sequencing, and data filtering, the total number $\hat{S}^+(t_j)$ of quantifiable tags corresponds to $\sim 5 \times 10^3 - 3 \times 10^4$ tagged cells. The sample ratio is defined by $\varepsilon(t_j) \equiv \hat{S}^+(t_j)/\hat{M}_{\rm ss}^+ = 3 \times 10^{-5} \sim 2 \times 10^{-4}$ where $\hat{M}_{\rm ss}^+ \approx 1.6 \times 10^8$ is the estimated total number of tagged cells in the peripheral blood. The number of sampled cells with the $i^{\rm th}$ tag from the $j^{\rm th}$ sample then approximately follows a Binomial distribution $B(S^+(t_j), \frac{m_i(t_j)}{M_{\rm ss}^+}) \approx B(m_i(t_j), \varepsilon(t_j))$ in our model. To quantitatively explore the feature of apparent extinctions of clones from a sample, we calculate the probability that no peripheral blood cell from clone i is found in a sample of size $S^+(t_j) \ll M_{\rm ss}^+$: $P(f_i(t_j) = 0 | m_i(t_j)) = \binom{M_{\rm ss}^+ - m_i(t_j)}{S^+(t_j)} \Big/ \binom{M_{\rm ss}^+}{S^+(t_j)} \approx \exp\left(-\frac{m_i(t_j) S^+(t_j)}{M_{\rm ss}^+}\right)$. Thus, if $m_i(t_j) < \varepsilon^{-1} = \hat{M}_{\rm ss}^+/\hat{S}^+(t_j) \sim 2 \times 10^4$ the $i^{\rm th}$ clone is likely to be missed in the sample. The value ε^{-1} is also used to threshold the population $m_{\rm b}(t)$ to define the measurable duration $\Delta\tau_{\rm b}$ of a

burst (as indicated in Figure 3.3(a)).

3.2.4 Parameter values

Parameters determined by the experimental procedure or estimated directly from the experiments include the weight of the animal, the sampling times t_j, the EGFP+ ratio, and the total number of tagged cells detected in each sample $\hat{S}^+(t_j)$. Since $\hat{M}^+(t_j)$ does not fluctuate much, we use its average for $\hat{M}^+_{\rm ss}$ and the relevant experimental parameters for each animal become $\theta_{\rm exp} = \{\hat{M}^+_{\rm ss}, \hat{S}^+_i(t_j), t_j\}$. These will also be used as inputs to our models.

Our multi-stage model also contains many other intrinsic parameters, including $\theta_{\rm model} = \{\lambda, C_{\rm h}, \alpha, r_{\rm n}, \mu_{\rm n}, \mu_{\rm n}^{(L)}, L, \omega, \mu_{\rm m}\}$. We first found parameter values that have been reliably independently measured. Some parameters were measured in human clinical studies rather than in rhesus macaques, but can nonetheless serve as reasonable approximations for non-human primates due to multiple physiological similarities [CQD09]. These estimates can certainly be improved once direct measurements on rhesus macaques become available. Model parameters, their estimates, and the associated references are given in Table 1 below.

3.2.5 Model properties and implementation

Using parameter estimates, we summarize the dynamical properties of our model and describe how the key model ingredients including stability of HSC clone distributions and subsequent "bursty" clone dynamics that follow differentiation can qualitatively generate the observed clone-size variances.

Slow homeostatic birth-death of HSCs - The first important feature to note is the slow homeostatic birth-death of HSCs. After the bone marrow is quickly repopulated, $r_{\rm h}(H(t)) - \mu_{\rm h} \approx 0$, and stochastic self-renewal slows down. Because $h = 0$ is an absorbing state, the size distribution of the clones may still slowly evolve and coarsen due to stochastic dynamics leading to the slow successive extinction of smaller clones. The typical timescale for overall changes in h can be estimated by approximating $r_{\rm h}(H_{\rm ss}) \approx \mu_{\rm h}$ [PQP08] and considering

Parameter	Interpretation	Values & References
HSC pool (Compartments **1**)		
H_{ss}	total number of HSCs at steady state	$1.1 \times (10^4 - 10^6)$ [SKL07, ZLM12, GKC15]
α	per-cell HSC differentiation rate	$5.6 \times 10^{-4} - 0.02$ [SKL07, GKC15, ZLM12]
μ_{h}	HSC death rate	$10^{-3} - 0.1$ [BBM03, ZLM12]
Transit Amplifying Progenitor pool (Compartment **2**)		
r_{n}	growth rate of progenitor cell	$2 - 3$ [ZLM12]
μ_{n}	death rate of progenitor cell (generation $\ell < L$)	0 [BBM03, ZLM12]
$\mu_{\text{n}}^{(L)}$	death rate of progenitor cell (generation $\ell = L$)	$0 - 0.27$ [BBM03, ZLM12]
ω	maturation rate of generation-L cells	$0.15 - 0.17$ [DDH76, LEZ16]
L	maximum generation of progenitor cells	$15 - 21$ [BBM03, ZLM12]
Peripheral Blood pool (Compartment **3**)		
M_{ss}	total number of mature granulocytes at steady state	$(2.5 - 5) \times 10^9$ [CQD09, KKP14]
μ_{m}	death rate of mature granulocytes	$0.2 - 2$ [BBM03, PBV10, LEZ16]

Table 3.1: Summary of parameters reported from the literature

Summary of parameters, including their biological interpretation, ranges of values, and references. All rate parameters are quoted in units of per day. Other parameters are chosen to be within their corresponding reported ranges from the referenced literature. How variations in parameter values of affect our analysis will be described in the subsequent sections.

the mean time $T(h)$ of extinction of a clone initially at size $h \ll H_{\text{ss}}$. The standard result given in Gardiner [Gar85] and also derived Text is $T(h) \approx \frac{h}{\mu_{\text{h}}} \left(1 + \ln \frac{H_{\text{ss}}}{h}\right) \gtrsim 10^2$ months (for $\mu_{\text{h}} = 10^{-2},\ H_{\text{ss}} = 10^4,\ h = 10^1$ see Table 1 for applicable values).

Since this timescale is larger than the time of the experiment (67 months for rhesus macaque RQ5427), mean HSC clone sizes do not change dramatically during the experiment, consistent with the stable number of clones observed in the samples (for rhesus macaque RQ5427, the number of detected clones at month {2, 8, 19, 25, 32, 43, 50, 56, 67} are $C_{\text{s}}(t_j) = \{184, 145, 186, 193, 152, 189, 155, 286\}$) shown in Figure 3.1(b). Thus, as a first

approximation, we will use a static configuration $\{h_i\}$ drawn from $P(h)$ to describe how, through differentiation, HSC clones feed the progenitor pool.

Fast clonal aging of progenitors - In contrast to slow HSC coarsening, progenitor cells proliferate "transiently." We plot a single burst of progenitor and mature granulocytes, Eqs. (3.10) and (3.11), in Figure 3.3(a) using the parameter values listed in Table 1. Associated with each temporal burst of cells, we define the characteristic duration, or "width" $\Delta\tau_{\rm b}$ as the length of time during which the number $m_{\rm b}(t)$ is above the detection threshold within a sample of peripheral blood: $\varepsilon^{-1} = \hat{M}^{+}_{\rm ss}/\hat{S}^{+} \approx 2 \times 10^4$.

According to Eq. (3.11), the burst width and height depend nonlinearly on the parameters L, $r_{\rm n}$, $\mu_{\rm n}$, $\mu_{\rm m}$, and ω in their physiological ranges (see Table 1). The characteristic "width" of a burst scales as $\Delta\tau_{\rm b} \sim L/r_{\rm n} + 1/\omega + 1/\mu_{\rm m}$. This estimate is derived by considering the L rounds of progenitor cell division, each of which takes time $\sim 1/r_{\rm n}$. Terminal-generation progenitors then require time $\sim 1/\omega$ to mature, after which mature granulocytes live for time $\sim 1/\mu_{\rm m}$. In total, the expected life span of $\sim L/r_{\rm n} + 1/\omega + 1/\mu_{\rm m}$, which approximates the timescale of a HSC-differentiation-induced burst of cells fated to be granulocytes. Using realistic parameter values, the typical detectable burst duration $\Delta\tau_{\rm b} \sim 1-2$ months is much shorter than the typical sampling gaps $\Delta t_j = 5 - 11$ months.

With this "burst" picture in mind, we now show how fluctuations of sampled clone sizes can be explained. Small-h ($\alpha h_i \ll \frac{1}{\Delta\tau_{\rm b}}$) clones never or rarely appear in blood samples. Their appearance also depends on whether sampling is frequent and sensitive enough to catch the burst of cells after rare HSC differentiation events. On the other hand, large-h ($\alpha h_i \gg \frac{1}{\Delta\tau_{\rm b}}$) clones differentiate frequently and consistently appear in the peripheral blood. Their populations in blood samples are less sensitive to the frequency of taking samples. Figure 3.3(b) shows two multi-burst realizations of $m_i(t)$ corresponding to two values of h_i. The 2000-day trajectories were simulated by fixing h_i and stochastically initiating the progenitor proliferation process. Population bursts described by Eq. (3.11) were added after each differentiation event distributed according to Poisson(αh_i).

Thus, the statistics of clone extinctions and resurrections should be more sensitive to the

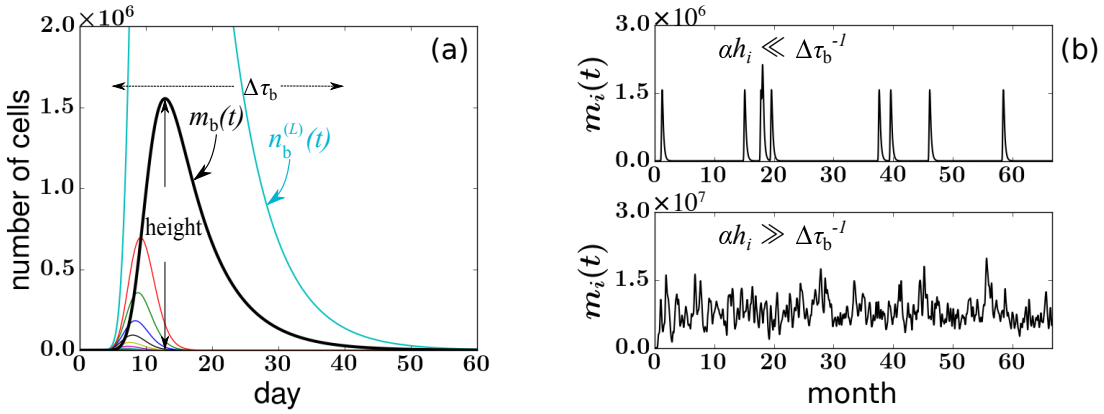

Figure 3.3: Bursty clonal dynamics

(a) A burst of cells is triggered by a single HSC differentiation event at time $t = 0$. A plot of representative solutions to Eqs. (3.10) and (3.11) for $r_{\rm n} = 2.5$, $L = 24$, $\mu_{\rm n} = \mu_{\rm n}^{(L)} = 0$, $\mu_{\rm m} = 1$, $A_{\rm ss}^{+} = 14.7$, and $\omega = 0.16$. Curves of different colors represent $n_{\rm b}^{(\ell)}(t)$, the progenitor cell population within each generation $\ell = 0, 1, 2, \ldots, L$, and $m_{\rm b}(t)$, the number of mature granulocytes associated with the differentiation burst. All populations rise and fall. (b) Realizations of PB numbers of a single clone arising from multiple successive differentiation events. The fluctuating populations are generated by adding together $m_{\rm b}(t)$ associated with each differentiation event. Time series resulting from small ($h_i/H^{+} = 0.0003$) and large ($h_i/H^{+} = 0.03$) HSC clones are shown. Small clones are characterized by separated bursts of cells, after which the clone vanishes for a relatively long period of time. The number of mature peripheral blood cells of large clones reaches a relatively constant level and almost never vanishes.

overall clonal differentiation rate αh_i than to the precise shape of a mature cell burst. This is confirmed by further simulation studies and analysis and motivates reducing the number of effective parameters (see Discussion).

We can further pare down the number of remaining parameters by finding common dependences in the model and defining an effective maximum generation number. We can rewrite Eq. (3.5) as $\beta \equiv 2^{L_{\rm e}}$, where

$$L_{\rm e} = L - L\log_2\left[\frac{r_{\rm n} + \mu_{\rm n}}{r_{\rm n}}\right] - \log_2\left[\frac{\omega + \mu_{\rm n}^{(L)}}{\omega}\right] \tag{3.12}$$

is an *effective* (and noninteger) maximum generation parameter. Later in Discussion, we show that uncertainties of the model structure, alternative mechanisms, and parameter values can be subsumed into $L_{\rm e}$. Henceforth, in our quantitative data analysis, we set the unmeasurable parameters $\mu_{\rm n} = \mu_{\rm n}^{(L)} = 0$ and subsume their uncertainties into an effective maximum generation $L_{\rm e}$. Finally, we invoke Eq. (3.4) to find the constraint

$$A_{\rm ss}^{+}\beta = A_{\rm ss}^{+}2^{L_{\rm e}} = M_{\rm ss}^{+}\mu_{\rm m}. \tag{3.13}$$

Since we can use the experimental value of $\hat{M}_{\rm ss}^{+}$ and $\mu_{\rm m}$ has been reliably measured in the literature, Eq. (3.13) constrains $A_{\rm ss}^{+}$ to $L_{\rm e}$.

After assigning values to parameters using Table 1 (setting $\mu_{\rm n} = 0$, $\omega = 0.16$ and $\mu_{\rm m} = 1$), subsuming parameters into $L_{\rm e}$ (setting $\mu_{\rm n}^{(L)} = 0$), describing the configuration $\{h_i\}$ through λ and $C_{\rm h}$ (setting $\mu_{\rm h} = 0$), and applying the constraint $A_{\rm ss}^{+}2^{L_{\rm e}} = \hat{M}_{\rm ss}^{+}\mu_{\rm m}$, we are left with four effective model parameters $\theta_{\rm model} = \{\lambda, C_{\rm h}, r_{\rm n}, L_{\rm e}\}$. Here we have included $r_{\rm n}$ in the key model parameters since it is not reliably measured and the cell burst width is sensitive to $r_{\rm n}$. Once $L_{\rm e}$ is inferred, Eq. (3.13) can be used to find $A_{\rm ss}^{+} = 2^{-L_{\rm e}}\hat{M}_{\rm ss}^{+}\mu_{\rm m}$.

3.2.6 Statistical model

The total number of tags observed across all samples (obtained by summing up the observed numbers of *unique* tags over J samples) can be used as a lower bound on $C_{\rm h}$. Even though estimates for animal RQ5427 give $C_{\rm h} \sim 550-1100$, the uncertainties $p_{\rm h}, K_{\rm h}$, and $H(0)$ makes λ and $P(h,t)$ difficult to quantify. Even if $P(h,t)$ were known, it is unlikely that the drawn $\{h_i\}$ will accurately represent those in the rhesus macaque, especially when $\lambda \approx 1$ and $P(h)$ becomes extremely broad (the variance of $P(h)$ approaches infinity). Thus we are motivated to find a statistical measure of the data that is insensitive to the exact configuration of $\{h_i\}$. The goal is to study the statistical correlations between various features of *only* the outputs, which should be insensitive to the input configuration $\{h_i\}$ but still encode information about the differentiation dynamics.

Two such features commonly used to fit simulated $f_i(t_j)$ to measured $\hat{f}_i(t_j)$ are the mean $y_i = \frac{1}{J}\sum_{j=1}^{J} f_i(t_j)$ and the variance $\sigma_i^2 = \frac{1}{J}\sum_{j=1}^{J}(f_i(t_j) - y_i)^2$. However, the small number

of measurement time points J and the frequent disappearance of clones motivated us to propose an even more convenient statistic that is based on

$$z_i = \sum_j \mathbb{1}(f_i(t_j), 0), \tag{3.14}$$

the number of absences across all samples of a clone rather than on σ_i. Here, the indicator function $\mathbb{1}(x, x') = 1$ when $x = x'$ and $\mathbb{1}(x, x') = 0$ otherwise. We illustrate in Subsection 3.5.7 alternatives such as data fitting based on σ_i and on an autocorrelation function but also describe the statistical insights gained from using statistics of z_i.

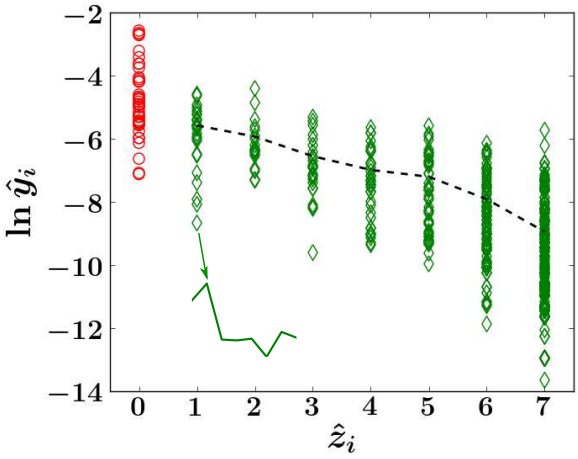

Figure 3.4: Scatterplot of clones in the feature space

Scatterplot of clone trajectories of animal RQ5427 displayed in terms of $\ln \hat{y}_i$, the log mean abundance of clone i, and $\hat{z}_i$, the number of samples in which clone i is undetected. The trajectory of each clone i is represented by a symbol located at a coordinate determined by its value of $\ln \hat{y}_i$ and $\hat{z}_i$. A trajectory of a clone that exhibits one absence within months $8 - 67$ is shown in the inset. The first sample at month 2 is excluded because only long-term repopulating clones are considered. Clones that are absent in all eight samples are also excluded, so the largest number of absences considered for animal RQ5427 is 7. The dashed black line denotes $\ln \hat{Y}_z$, where $\hat{Y}_z$ is the average of $\hat{y}_i$ calculated over i within each bin of z as shown in Eq. (3.15). When later analyzing $\hat{Y}_z$, $\hat{Y}_0$ (red circles) is not included.

The level of correlation between $\hat{z}_i$ and $\hat{y}_i$ is measured by the average of $\hat{y}_i$ conditioned

by their number of absences $\hat{z}_i$ (dashed curve) in Figure 3.4, where the distribution of the values of $\hat{y}_i$ at each $\hat{z}_i$ is clearly shown. To combine the correlated stochastic quantities z_i and y_i into a useful objective function, we take the expectation of y_i over clones that have the same z_i:

$$Y_z = \frac{\sum_i y_i \mathbb{1}(z_i, z)}{\sum_i \mathbb{1}(z_i, z)}. \tag{3.15}$$

In case no simulated or data-derived trajectories $f_i(t_j)$ exhibit exactly z absences, we set $Y_z = 0$. We then determine $Y_z(\theta_{\text{model}})$ from simulating our model and $\hat{Y}_z$ from experiment and use the mean squared error (MSE) between the two as the objective function:

$$\text{MSE}(\theta_{\text{model}}) = \sum_{z=1}^{J-1} \left[Y_z(\theta_{\text{model}}) - \hat{Y}_z \right]^2, \tag{3.16}$$

where $\theta_{\text{model}} = \{\lambda, C_{\text{h}}, r_{\text{n}}, L_{\text{e}}\}$ and the sum is taken only over those z for which both data and simulations produce at least one clone (in practice, when searching for the best fitting θ_{model}, we ensure at least 30 clones in each bin of z). Here Y_0 is excluded from the MSE calculation because the y_i values of clones that have $z_i = 0$ are not constrained by the burstiness of the model and Y_0 can be sensitive to the underlying configuration $\{h_i\}$.

We are now in a position to compare results of our model with experimental data. The general approach will be to choose a set of parameters, simulate the forward model (including sampling) to generate clone abundances $\{f_i(t_j)\}$, number of absences z_i, and ultimately $Y_z(\theta_{\text{model}})$, which is then compared to data-derived $\hat{Y}_z$. By minimizing Eq. (3.16) with respect to θ_{model}, we obtain their least square estimates (LSE). A schematic of our workflow is shown in Figure 3.5. We describe the details of the simulation of our model in Subsection 3.5.5.

3.3 Results

By implementing the protocol outlined in Figure 3.5, we find a number of results including the shape of the MSE, least-squares-estimates (LSE) of the parameters, validation of the mechanistic model, and sensitivity analysis.

Shape of the MSE function. For the range of values $r_{\text{n}} \in [0.01, 10]$ and $L_{\text{e}} \in [19, 28]$, the MSEs are fairly insensitive to $\lambda \geq 0.5$ and $500 \leq C_{\text{h}} \leq 1000$, but typically has lower

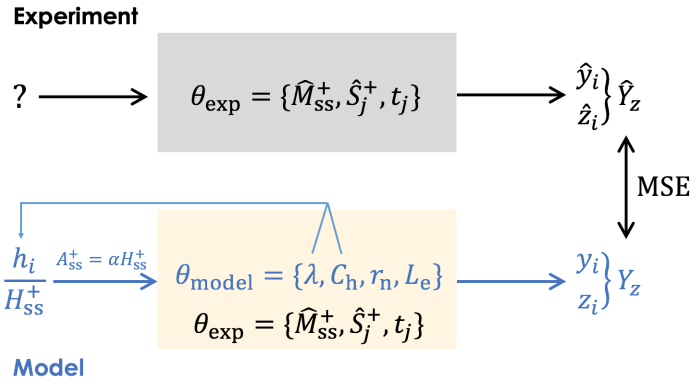

Figure 3.5: Workflow of the model

Workflow for comparing parameter-dependent simulated data with measured clone abundances. The initial input is the HSC clone distribution $P(h)$, which is unknown and experimentally unmeasurable. Using known experimental parameters θ_{exp} and choosing model parameters θ_{model} the theoretical quantities y_i and z_i are computed by simulating the mechanistic model and the sampling. The corresponding $\hat{y}_i$ and $\hat{z}_i$ are extracted from data and the theoretical $Y_z(\theta_{\text{model}})$ and the experimental $\hat{Y}_z$ compared through the MSE defined in Eq. (3.16). The MSE is then minimized to find least squares estimates for θ_{model}.

values near $\lambda \approx 0.99$ and $C_{\text{h}} \approx 500$. Note that $C_{\text{h}} \approx 500$ is close to the experimental estimate for animal RQ5427. Therefore, we fix $\lambda = 0.99$, $C_{\text{h}} = 500$ and minimize the MSE with respect to r_{n} and L_{e}. For each $\{r_{\text{n}}, L_{\text{e}}\}$ pair, simulation of the full model is repeated 200 times to generate 200 values of all (y_i, z_i) pairs, $Y_z(\lambda = 0.99, C_{\text{h}} = 500, r_{\text{n}}, L_{\text{e}})$, and $\text{MSE}(\lambda = 0.99, C_{\text{h}} = 500, r_{\text{n}}, L_{\text{e}})$. The average values of these MSEs are plotted as a function of r_{n} and L_{e} in Figure 3.6.

We find that the minimum of the MSE is relatively insensitive to L_{e} for $r_{\text{n}} \gtrsim 1$. To interpret this result, note that r_{n} does not affect the absolute value of β according to Eq. (3.13), but it affects the typical time $\sim L/r_{\text{n}} + 1/\omega$ it takes for a generation-0 progenitor cell to form a mature granulocyte. When $r_{\text{n}} < \mu_{\text{m}}$, the proliferation of progenitors cannot "catch up" with the loss of granulocytes, resulting in a quickly vanishing burst in m_{b}. A larger L_{e} would be required to compensate. When $r_{\text{n}} \gg \mu_{\text{m}}$, the accumulation of $m_i(t)$ is much quicker than

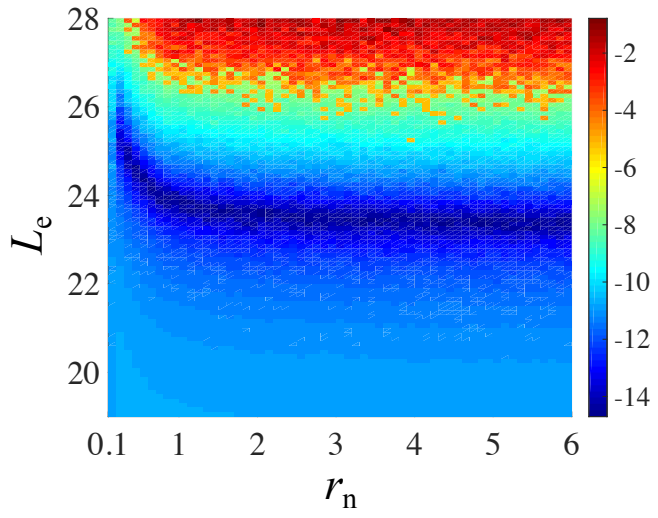

Figure 3.6: The objective function MSE(r_n, L_e)

Dependence of the mean MSE defined in Eq. (3.16) on r_n and L_e. For visualization purposes, we took the natural logarithms of MSE values and plotted them as a function of L_e and r_n. Blue area denotes smaller MSE values, thus better fitting. This energy surface was generated by averaging over 200 simulations using $C_h = 500$ and $\lambda = 0.99$.

its loss so the burst size is relatively stable and L_e^* is not very sensitive to r_n. Thus, the MSE objective function is fairly insensitive to r_n in its biologically meaningful value range.

Least-squares estimates of L_e and A_{ss}^+ for animal RQ5427. To obtain an explicit best-fit value for L_e we fix $r_n = 2.5$ [ZLM12] (and $\lambda = 0.99$, $C_h = 500$) and varied $L_e \in [19, 26]$. The MSE objective function as L_e is varied is shown in Figure 3.7(a). For *one* simulation at each chosen value of L_e, we can construct the MSE and find an LSE for L_e. Over 200 sets of simulations (for each chosen L_e), we find the expected LSE value $L_e^* = 23.4$, with a standard deviation of ± 0.12. So in practice, the randomness across simulations is negligible. Upon applying the constraint in Eq. (3.13), the corresponding $(A_{ss}^+)^* = 14.7$. Substituting the LSE results into Eq. (3.11) yields a burst width of $\Delta\tau_b \approx 32$ days, which is consistent to our assumption $\Delta\tau_b \ll \Delta t_j = 5 - 11$ months. Figure 3.7(b) shows how $Y_z(L_e^* = 23.4)$ fits $\hat{Y}_z$. We also plotted Figure 3.8 which shows systematic bias with larger or smaller values of L_e.

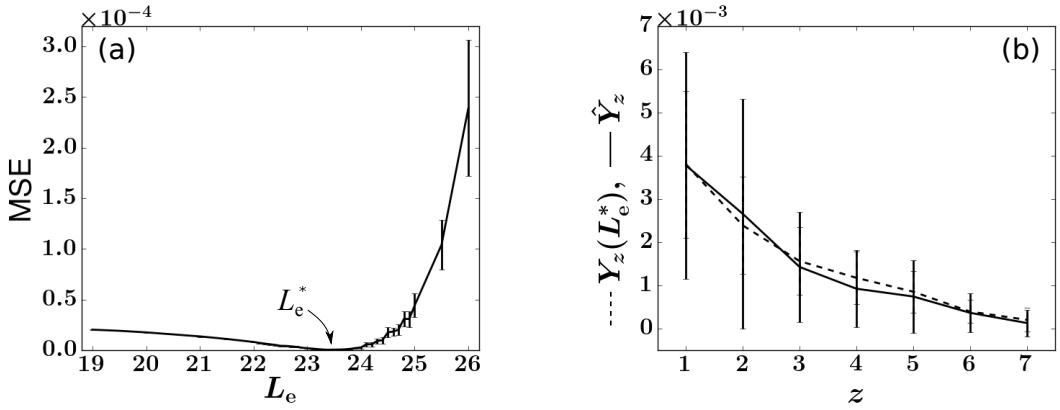

Figure 3.7: MSE($L_{\rm e}$) and the optimal fitting of $\hat{Y}_z$

Finding the least squares estimate (LSE) $L_{\rm e}^*$ for animal RQ5427 by fitting the simulated Y_z to the experimental $\hat{Y}_z$. The values of ($\lambda,\ C_{\rm h},\ r_{\rm n}$) are chosen to be (0.99, 500, 2.5). Simulations with $\{h_i\}$ set to $\{\hat{y}_i\}H_{\rm ss}^+$ instead of drawing from $P(h)$ generate similar results. (a) The LSE is $L_{\rm e}^* = 23.4$. Averages and standard deviations (error bars) of the 200 MSEs are plotted. (b) Comparisons between the experimental (solid) $\hat{Y}_z$ and simulated (dashed) Y_z with fixed $L_{\rm e}^* = 23.4$. The error bars are determined by considering the standard deviation of the average abundances (y_i or $\hat{y}_i$) of all clones exhibiting z absences.

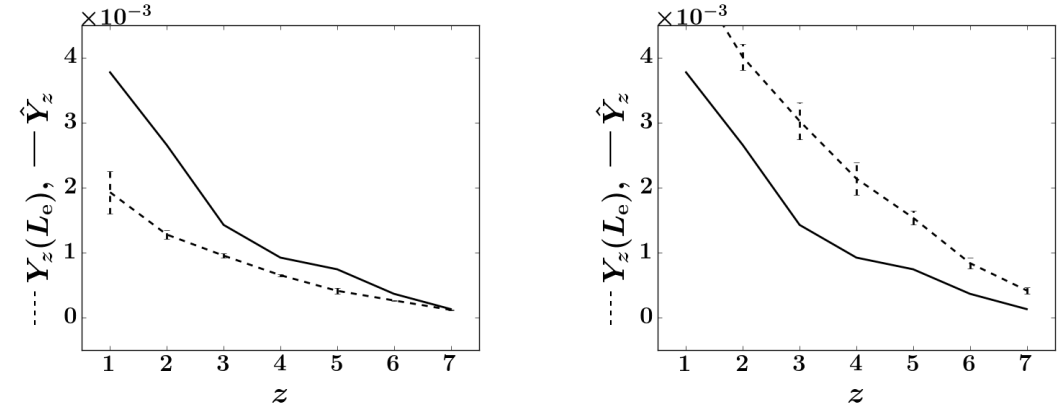

Figure 3.8: Fitting of $\hat{Y}_z$ when $L_{\rm e} \neq L_{\rm e}^* = 23.4$

In both cases of $L_{\rm e} = 22.4 < L_{\rm e}^*$ (a) and $L_{\rm e} = 24.4 > L_{\rm e}^*$ (b), Y_z shows systematical bias from $\hat{Y}_z$ across all values of z.

Comparison of variability from simple sampling and best-fit model. We can check how our LSE result performs against the null hypothesis that clone size variations arise only from random sampling. An estimate of sampling-induced variability can be obtained by assuming a specific number of peripheral blood granulocytes of tag i and randomly drawing an experimentally determined fraction $\varepsilon(t_j)$ of peripheral blood cells. This is repeated J times from a constant peripheral pool $\{m_i\}$. Each draw results in $s_i(t_j)$ cells of clone i in the simulated sample. Normalizing by $S^+(t_j)$, the total number of tagged cells in the sample, we can define the rescaled mean abundance $y_i = \frac{1}{J}\sum_{j=1}^{J} f_i(t_j)$ and the rescaled standard deviation $\sigma_i = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(f_i(t_j) - y_i)^2}$ for each clone i. The simulated quantities $\ln y_i$ and σ_i associated with each clone i and its experimental sampling fraction $\varepsilon(t_j)$ are indicated by the green triangles in Figure 3.9(a). The corresponding values $\ln \hat{y}_i$ and $\hat{\sigma}_i$ derived from the data shown in Figure 3.1(b) are indicated by the blue dots. This simple heuristic test shows that the experimental fluctuations in clone abundances are significantly larger than that generated from random sampling alone and that additional mechanisms are responsible for the fluctuation of clone abundances in peripheral blood. Figure 3.9(b) shows the fluctuations in clone abundances obtained from random sampling of fluctuating mature clones simulated from our model, using LSE parameter values. Here, the variability is a convolution of the fluctuations arising from intrinsic burstiness and from random sampling. The total variability fits those of the experimental data well except for several large-sized outlier clones.

Insensitivity of analysis to HSC configurations. We demonstrate the weak dependence of our least squares estimate to λ, the parameter controlling the shape of $P(h,t)$, as shown in Figure 3.10. For each λ, we sample a fixed number ($C_\mathrm{h} = 500$) of h_i from the theoretical distribution $P(h,t)$ and let L_e vary between 19 and 28. We then simulate the model 200 times and find 200 MSEs at each value of $L_\mathrm{e} \in [19, 28]$. The averages of the 200 MSE's at each value of L_e are compared and the L_e^* that corresponds to the minimal average MSE is selected. The selected L_e^* as a function of λ is plotted in Figure 3.10(a). Figure 3.10(b) shows the averages and standard deviations of MSE(L_e^*) at each value of λ.

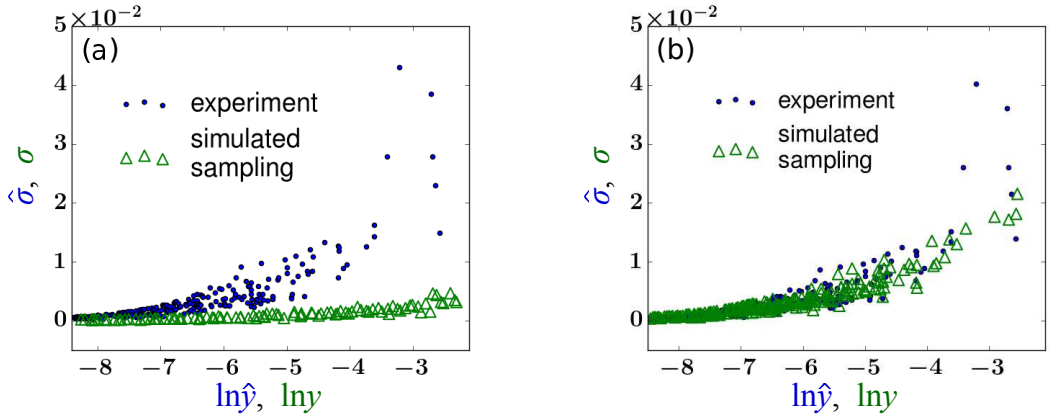

Figure 3.9: Averages and standard deviations of clonal abundances in animal RQ5427 (a) A plot of the standard deviation $\hat{\sigma}_i$ *vs.* the log of the mean $\hat{y}_i$, extracted from abundance data (blue dots). For comparison, clonal tags distributed within the peripheral blood cells were randomly sampled (with the same sampling fraction $\varepsilon(t_j)$ at times t_j as in the experiment). The analogous quantity σ_i shown by the green triangles indicate a much lower standard deviation for a given value of $\ln y_i$. This simple test implies that the clonal variability across time cannot be explained by random sampling. (b) The same test is performed after applying our model with the LSE parameter $L_{\mathrm{e}} = 23.4$ (and the average of parameters listed in Table 1).

We then repeat the simulations with $C_{\mathrm{h}} = 1000$. These results together show that L_{e}^* is insensitive to the distribution of h_i. This insensitivity might be understood by noticing that Y_z is defined as the *mean* of the values of y_i that are associated with z absences (dashed curve in Figure 3.4), and is not necessarily sensitive to how these values are distributed (vertically distributed markers at each value of z in Figure 3.4). Instead, Y_z encodes the intrinsic correlation between y_i and z_i and how much burstiness is transmitted to a clone's $f_i(t_j)$ from its h_i.

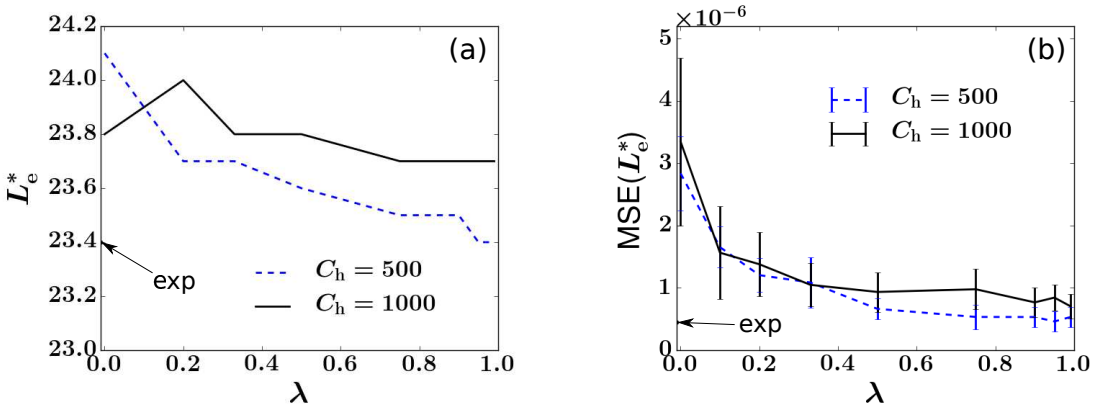

Figure 3.10: The objective function is insensitive to λ and $C_{\rm h}$

The LSE $L_{\rm e}^*$ is insensitive to the geometric distribution factor $\lambda > 0$ and to $C_{\rm h} \gg 1$. This implies that for a wide range of values of λ and $C_{\rm h}$ the LSEs are insensitive to the HSC configuration $\{h_i\}$. (a) $L_{\rm e}^*$'s found at each value of λ. (b) Averages and standard deviations (error bars) of MSE($L_{\rm e}^*$) as a function of λ. The LSE and MSE($L_{\rm e}^*$) values associated with self-consistently using $\{h_i\}/H^+ = \{\hat{y}_i\}$ from experimental data are marked by arrows and "exp."

In Figure 3.11, we plot simulated datasets and their Y_z features under various combination of model parameters. It can be observed that Y_z is insensitive to λ but quite sensitive to $L_{\rm e}$.

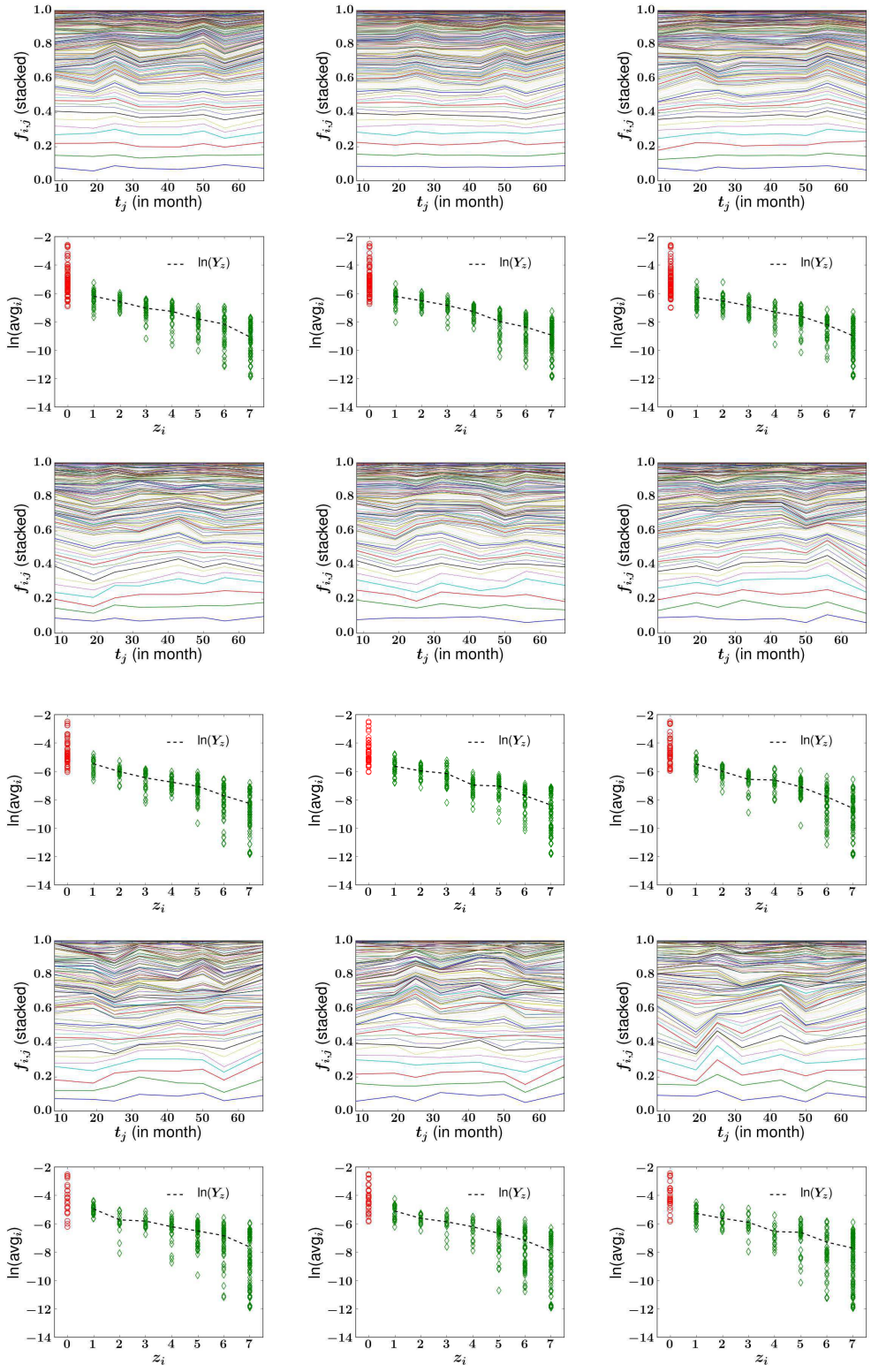

Figure 3.11: More simulations under various λ and L_{e}

Parameters: $\lambda = 0,\ 0.5,\ 0.99$ from left to right, $L_{\mathrm{e}} = 22.4,\ 23.4,\ 24.4$ from top to bottom.

To conclude, though it is generally impossible to recover the exact $\{h_i\}$ configuration, we find the HSC self-renewal-induced geometric distribution in Eq. (3.7) with factor $\lambda \geq 0.5$ generates consistent comparisons with the sampled data.

Data analysis and fitting for animals 2RC003 and RQ3570. The data from the three different rhesus macaques vary in their numbers of tagged clones transplanted and the lengths of the experiments. For animal RQ5427/2RC003/RQ3570, there are 536/1371/442 clones that are detected at least once within 67/103/38 months. The fraction of cells in all tracked clones in animal RQ5427/2RC003/RQ3570 was approximated by the average fraction of cells that were EGFP+ marked over time, around 0.052/0.049/0.086 (the ratios between green square and blue triangle markers in Figures 3.1(a), 3.12(a), and 3.13(a)), respectively. Figures 3.12 and 3.13 also show the clone abundances, the MSE functions, and the statistics of $Y(z)$.

Despite differences among the animals and the large variability in the estimated values of α and $H_{\rm ss}$ individually reported in the literature [SKL07, ZLM12, GKC15], the estimates of $(A^+_{\rm ss})^*$ and $L^*_{\rm e}$ are rather similar across the three animals. For animal 2RC003, the optimal estimates are $L^*_{\rm e} \approx 25.0$, while for animal RQ3570, $L^*_{\rm e} = 24.0$. The corresponding estimates for A^*, after considering the constraint Eq. (3.13) and the EGFP+ ratios in Table 2, are 282.7, 136.7, and 224.4.

We also compared how the simulated LSE $Y_z(L^*_{\rm e})$ fits the experimental $\hat{Y}_z$ for all three animals. Note that for each specific z, the value of Y_z is the conditional mean of the values of y_i for which clones i exhibits exactly z absences. To evaluate the "quality" of fitting $Y_z(L^*_{\rm e})$ to $\hat{Y}_z$, one can directly perform a two-sample t-test between the two sets of values $y_i(L^*_{\rm e})$ and $\hat{y}_i$ that contribute to each value of z. The group of $\hat{y}_i$ values corresponding to each value of z is shown by the vertical cluster of diamonds in Figure 3.4, while the corresponding set of values of $y_i(L^*_{\rm e})$ are generated by simulations. For each z value, we performed 200 simulations and collected the values $y_i(L^*_{\rm e})$ of all clones i that exhibit z absences and contribute to $Y_z(L^*_{\rm e})$. We ensured at least 30 values of $y_i(L^*_{\rm e})$ for each z and performed the t-test with the measured set $\hat{y}_i$ containing clones that exhibit the same number of absences z. Performing this t-test

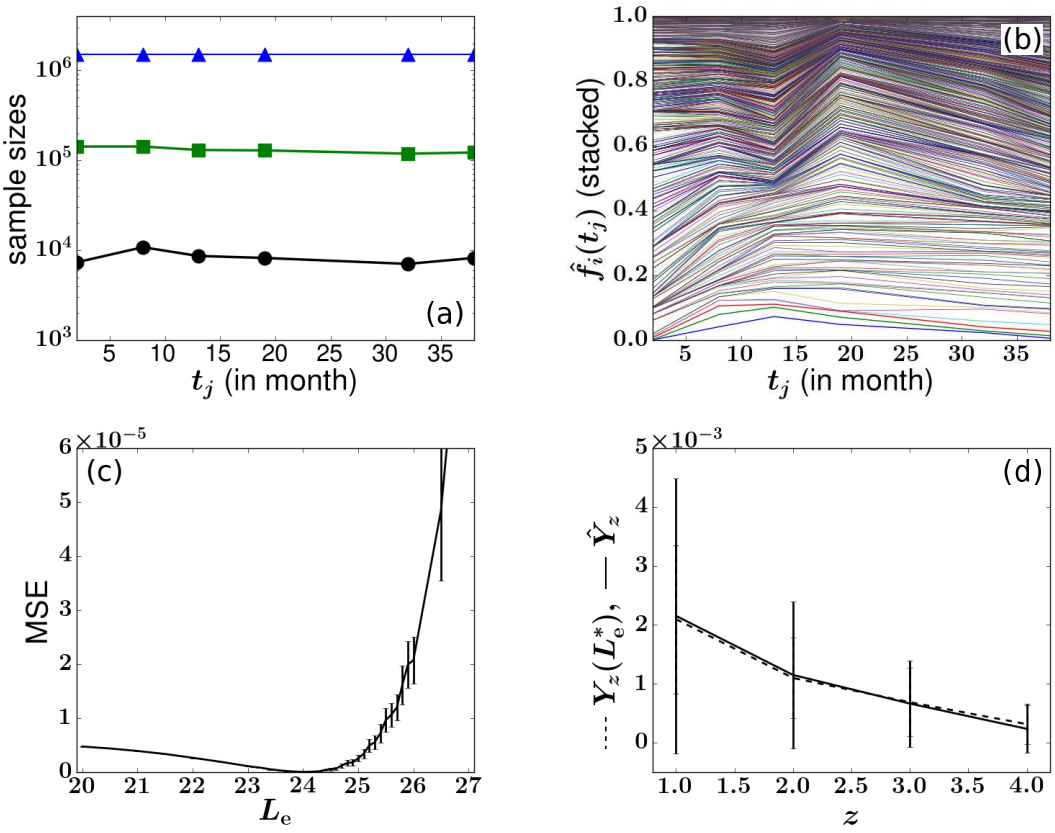

Figure 3.12: Results for animal 2RC003

(a-b) Experimental data for animal 2RC003. (c) Difference between experimental $\hat{Y}_z$ and simulated $Y_z(L_e)$ as a function of L_e. The values of h_i's are set to be equal to $H^+\hat{y}_i$ and the model was simulated 200 times at each value of L_e. Other parameters are taken from Tables 1 and 2. The LSE $L_e^* = 25.0$ and $(A_{ss}^+)^* = 6.7$. (d) Comparison of the optimal Y_z to the experimental $\hat{Y}_z$.

for all $1 \leq z \leq J-2$, we generate $200(J-2)$ p-values. Data for $z = J-1$ is too noisy and was not included. A p-value $p < 0.05$ would indicate that the two sample means $Y_z(L_e^*)$ and $\hat{Y}_z$ are not likely to be considered identical; if instead $p \geq 0.05$, the null hypothesis of equal sample means cannot be rejected. For animals RQ5427 and RQ3570, 93.5% and 99.4% of the p-values are larger than 0.05, while for animal 2RC003, the fraction is 76.6%. This result is consistent with the eroded fitting quality with increasing experimental time, where the slow decrease in the number $C_h(t)$ of HSC clones in animal 2RC003 cannot be neglected. As evident from Figure 3.12(a), several clones start to dominate after month 64;

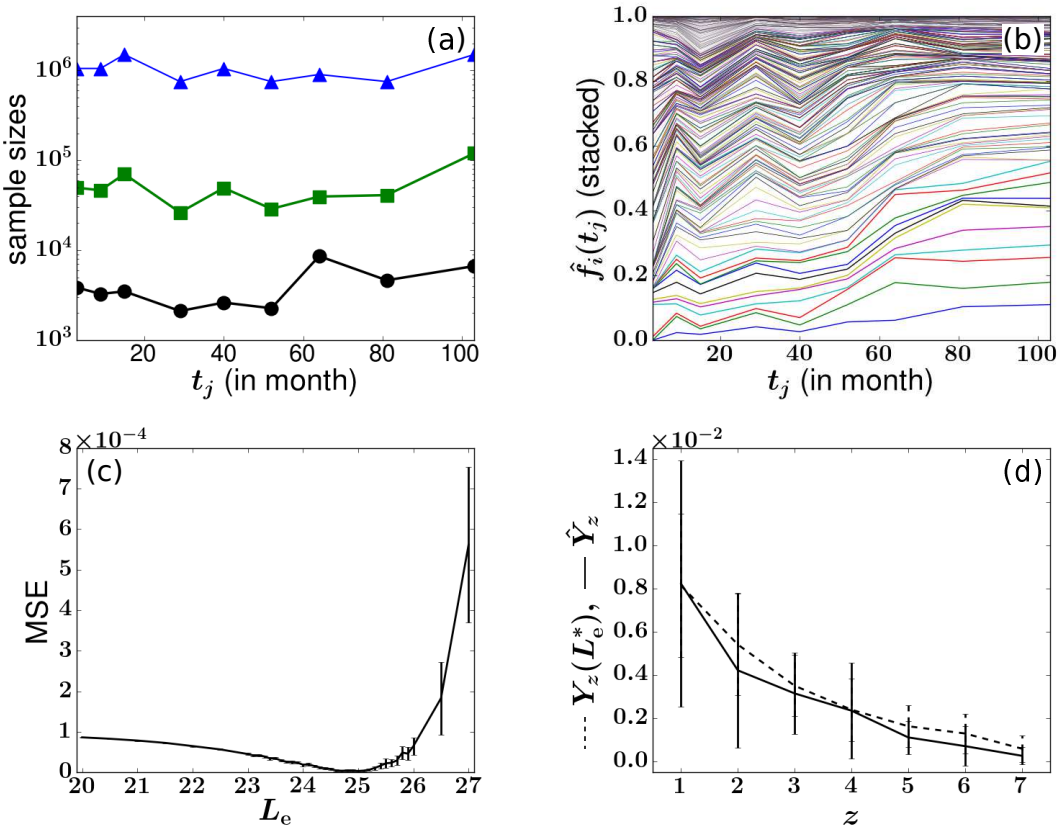

Figure 3.13: Results for animal RQ3570

Experimental data (a-b) and fitting results (c-d) for animal RQ3570. The values of h_i's are set to be equal to $H^+\hat{y}_i$. Other parameters are taken from Tables 1 and 2. The LSE fitting results are $L_e^* = 24.0$ and $(A_{ss}^+)^* = 19.3$.

this coarsening phenomenon is not evident in the data of the other two rhesus macaques. Animal RQ3570 was sacrificed at month 38 so no obvious coarsening is observed and no clones strongly dominate (see Figure 3.13). A summary of the parameters and fitting results for all animals is given in Table 2.

3.4 Discussion

In this study, we analyzed a decade-long clonal tracking experiment in rhesus macaques and developed mechanistic and statistical models that helped us understand two salient features

Parameter	Reference range or LSE value		
	RQ5427	**2RC003**	**RQ3570**
$\hat{C}_{\mathrm{s}}$	536	442	1371
$(A_{\mathrm{ss}}^{+})^{*}$	14.7	6.7	19.3
A_{ss}^{*}	282.7	136.7	224.4
L_{e}^{*}	23.4	25.0	24.0
M_{ss}	3.2×10^{9}	4.6×10^{9}	3.8×10^{9}
$S^{+}(t_j)$	$(5.0 - 30) \times 10^{3}$	$(2.1 - 8.6) \times 10^{3}$	$(7.0 - 10.8) \times 10^{3}$
EGFP+ ratio	0.052	0.049	0.086
$\varepsilon(t_j)$	$(2.8 - 20) \times 10^{-5}$	$(1.2 - 4.2) \times 10^{-5}$	$(2.4 - 3.0) \times 10^{-5}$
Δt_j	$150 - 330$	$180 - 660$	$150 - 260$

Table 3.2: Summary of fitting results for the three rhesus macaques

Summary of specific parameter values for rhesus macaques 2RC003 and RQ3570 derived from experimental measurements [KKP14] or obtained by calculations (L_{e}^{*} and $(A_{\mathrm{ss}}^{+})^{*}$).

of clone abundance data: the heterogeneous (nonuniform) distribution of clone sizes and the temporal fluctuation of clone sizes. Below, we further discuss the implications of our results, the structure of our mechanistic model, and the potential effects of including additional biological processes.

Comparison to previous studies: The long-term clonal tracking data we analyzed were generated from a huge number of initially tagged HSPCs ($C_{\mathrm{h}}(0) \sim 10^{6} - 10^{7}$) [KKP14], a large number of observed clones ($C_{\mathrm{s}} \sim 10^{2} - 10^{3}$), small numbers of sequenced cells ($10^{3} - 10^{4}$), and infrequent sampling. This presents significant challenges to the modeling and analysis over previous studies that mostly focused on one or a few clones [CBE12,MSB12,VBZ13,BKB15].

In a previous analysis, Goyal et al. [GKC15] aggregated the clone abundance data across *all* mature cell types and studied the distribution of the *number* of clones of specific size. At each time point, they ordered the clones according to their sizes. Thus, the ordering can change across samples as some clones expand while others diminish. They found that the cumulative clone number distribution (defined as the number of clones of a specific size or less) of the size-ordered clones become stationary as soon as a few months after

transplantation. They proposed a neutral birth-death description of progenitor cells and fitted the *expected* value of clone counts in each sample by assuming $h_i \equiv 1\,\forall i$ ($P(h,t) = \mathbb{1}(h,1)$) and tuning parameters in the downstream progenitor and mature cell compartments. By focusing on aggregate clone counts, this study could not distinguish the dynamics of individual clones nor could it predict the persistence of clone sizes over time. Since individual clone sizes ($h_i,\ n_i,\ m_i,\ s_i$ of the same tag i) were not tracked, mechanisms driving the dynamics, and in particular, the variability and fluctuations of *individual* clone sizes that drive disappearances and reappearances, remain unresolved [GKC15].

In our model, heterogeneity of clone sizes is explicitly generated by stochastic HSC self-renewal of cells of each tag and extinctions and resurrections arise from a generation-limited progenitor proliferation assumption. We infer model parameters as listed in Table 2. Combining the results with previous experimental and theoretical estimates of $H_{\rm ss} \approx 1.1 \times 10^4 - 2.2 \times 10^4$ [ACM02, GKC15] results in $\alpha = 0.0045 - 0.027$, slightly larger than, but still consistent with, the estimates $\alpha = 0.0013 - 0.009$ by Shepherd *et al.* [SKL07]. Previous studies that modeled total peripheral blood population estimated $\alpha \approx 0.022$ and $H_{\rm ss} \approx 1.1 \times 10^6$/kg for dog and $\alpha \approx 0.044$ and $H_{\rm ss} \approx 1.1 \times 10^6$/kg for human [ZLM12]. These estimates yield a value of $\alpha H_{\rm ss}$ about $10^2 - 10^3$ times greater than ours, which is nonetheless consistent with our steady-state constraint Eq. (3.13) because they assumed a much smaller $L \approx 15 - 18$ for dog and $16 - 21$ for human. This difference in the estimates of L may be partially attributed to the transplant conditions under which the rhesus macaque experiments were performed [KKP14]. Alternative model assumptions and differing values of other parameters may also contribute to this difference. For example, the extremely large value of $H_{\rm ss} \approx 10^7$ used in [BBM03] will naturally decrease their estimate for $L_{\rm e}^*$ relative to that of our analysis.

Model structure, sensitivity to parameters, and cellular heterogeneity: Uncertainties in values of parameters such as $\mu_{\rm h},\ p_{\rm h},\ K_{\rm h}$, and other factors that tune the symmetric-asymmetric modes of HSC differentiation or involve HSC activation processes [WLO08] will impart uncertainty in determining $P(h)$ and $\{h_i\}$. We have assumed $P(h)$ satisfies a master

equation and depends only two effective parameters λ and $C_{\rm h}$. However, we have demonstrated that the statistical properties of Y_z are quite insensitive to the upstream configuration $\{h_i\}$ and hence to λ and $C_{\rm h}$ for a wide range of their values (see Figure 3.10). In other words, very little information in $\{h_i\}$ is retained in the sampled abundances $\hat{f}(t_j)$ after HSCs differentiate and trigger random bursty peripheral blood cell population dynamics.

Another feature we have ignored in our neutral model is cellular heterogeneity such as tag-dependent differentiation, proliferation, and death rates. Cellular heterogeneity in HSC differentiation rates could be described by different α_i for each clone i and the total differentiation rate would be $A_{\rm ss}^+ = \sum_{i=1}^{C_{\rm h}} \alpha_i h_i$. Differences in α_i can be subsumed into a modified configuration $\{h_i\}$ which, as we have seen, does not strongly influence our parameter estimation based on the Y_z statistics. Thus, given the available data and how information is lost along the stages of hematopoiesis and sampling, the present quasi-steady state analyses cannot resolve heterogeneity across HSC clones.

We have not investigated how cellular heterogeneity in progenitor and mature cells would affect our results, but clone-dependences in their birth and death rates could affect sizes and durations of population bursts and quantitatively affect our analysis. However, unless the statistics of inter-burst times are highly variable across clones, we do not expect cellular heterogeneity to qualitatively affect our conclusions.

Changing downstream parameters such as $\mu_{\rm m}$ or invoking alternative mechanisms of terminal differentiation can affect the shape of clonal bursts. We show in **Effective parameters** that these effects can be subsumed into the effective maximum progenitor generation $L_{\rm e}$. We have performed additional simulations to confirm that changing $\mu_{\rm m} = 2$ will not influence the fitting of $A_{\rm ss}^+$ but but increases $L_{\rm e}^*$ by one. In other words, inference of $(A_{\rm ss}^+)^*$ is robust against many upstream and downstream parameters, indicating that the intrinsic clone size fluctuations observed in the experimental data strongly constrain the total rate of HSC differentiation. On the other hand, uncovering the actual maximal generation L^* from $L_{\rm e}^*$ is possible only when uncertainties in these other parameters are resolved.

Of course there is a nearly endless list of details such cellular heterogeneity and more

complex biology that we did not include, but given the noisy data, we propose and quantify the simplest explanation for the observed heterogeneous clone abundances and the temporal "extinctions and resurrections." The key ingredients in our mechanistic model are HSC self-renewal (quantified by the effective parameter λ), intermittent HSC differentiation (quantified by the parameter $A^+_{\rm ss}$), and an effective maximum progenitor generation (quantified by the effective parameter $L_{\rm e}$). Although we cannot fully resolve λ from data, the obvious mismatch between experiment and our model when λ is small shows that a certain level of HSC clone-size heterogeneity (larger $\lambda \approx 1$) is necessary to match the sampled data. Similarly, we cannot fully resolve α and $H^+_{\rm ss}$, but their product, the total tagged HSC differentiation rate $A^+_{\rm ss} = \alpha H^+_{\rm ss}$, is one of the key parameters constrained by our modeling.

Effective parameters There are differing reports on the measured death rates for circulating granulocytes. We have used the most recently reported value $\mu_{\rm m} = 1$ per day for human. The effect of changing the value of $\mu_{\rm m} \to \mu'_{\rm m}$ on our analysis is a reinterpretation of $L_{\rm e}$. By rewriting Eq. (16) as $A^+_{\rm ss} 2^{L_{\rm e}} = M^+_{\rm ss}\mu_{\rm m} = M^+_{\rm ss}\mu'_{\rm m}\left(\frac{\mu_{\rm m}}{\mu'_{\rm m}}\right)$, we rearrange the expression to $A^+_{\rm ss} 2^{L_{\rm e}+\log_2(\mu'_{\rm m}/\mu_{\rm m})} = M^+_{\rm ss}\mu'_{\rm m}$ and find $L'_{\rm e} = L_{\rm e} + \log_2(\mu'_{\rm m}/\mu_{\rm m})$. For example, $\mu'_{\rm m} = 2$ would lead to $L'_{\rm e} = L_{\rm e} + 1$, where one additional round of progenitor doubling compensates for the doubled loss rate of mature granulocytes. One may argue that the change in $\mu_{\rm m}$ can also be compensated for by doubling $A^+_{\rm ss}$, which would have a different effect on the burstiness of the model compared to doubling $L_{\rm e}$. However, when re-fitting the data with $\mu'_{\rm m} = 2$ or 0.2, we observed that $(A^+_{\rm ss})^*$ did not change much, with most of the effect of modifying $\mu_{\rm m}$ absorbed by changes in $L^*_{\rm e}$.

Similarly, uncertainties in other parameters can also be subsumed into $L_{\rm e}$. For example, setting $\mu^{(L)}_{\rm n} = \omega > 0$ implies that only half of the generation-L progenitors contribute to the peripheral blood. For a model with $\mu^{(L)}_{\rm n} = 0$ to generate an equivalent effect, we can halve the number of mature cells by using an effective maximum generation parameter $L'_{\rm e} = L_{\rm e} - 1$. This indicates that the intrinsic clone size fluctuations demonstrated in the experimental data strongly constrain $A^+_{\rm ss}$.

Another possible modification of our mechanistic model is to allow for the possibility of

symmetric HSC differentiation. The effect of symmetric differentiation can again be subsumed into the parameter L_{e} without qualitatively affecting our analysis. Assume a proportion $0 \leq q \leq 1$ of HSC differentiations are symmetric, producing on average $1+q$ generation-0 progenitor cells. After L_{e} rounds of proliferation, the $1+q$ generation-0 progenitors produce on average $(1+q) \times 2^{L_{\mathrm{e}}}$ mature cells. This is equivalent to an exclusively asymmetric differentiation model ($q = 0$) with $L'_{\mathrm{e}} = L_{\mathrm{e}} + \log_2(q+1)$. We also expect symmetric differentiation to slightly increase the speed of coarsening since each HSC differentiation is also accompanied by the HSC's death and clones represented by a single HSC would disappear under symmetric differentiation. However, given the small rate α of HSC differentiation, the large number C_{h} of clones, and the insensitivity of our results to the distribution h_i, the data cannot quantitatively resolve the symmetric-asymmetric modes of HSC differentiation.

Clonal stability *vs* clonal succession. Our model reduction was based on the separation of timescales of the slow HSC dynamics and the fast clonal aging dynamics. Since HSC clone sizes vary extremely slowly for primates ($\sim \mathcal{O}(10^2)$ months), we ignored the homeostatic births/deaths of HSCs when fitting the temporal clonal variations. This is partially justified by visual inspection of Figs 3.1(b), 3.12(b), and 3.13(b) that no significant variations of large clones' abundances is observed before 60 months. Instead, the random intermittent HSC differentiation events induce relatively short ($\sim \mathcal{O}(1)$ months) bursts of granulopoietic progeny that contribute strongly to temporal fluctuations of clone sizes. Such behavior is consistent to the "clonal stability" hypothesis [APS95, PPB96, MGD06], which assumes that a fixed group of HSCs randomly contribute to an organism's blood production at all times.

The alternative hypothesis of "clonal succession" [JL90, DKC96, SRC14] assumes that different groups of HSCs are sequentially recruited to the blood production at different times. This hypothesis would only be consistent with our model under a different set of parameters where HSCs self-renew/die at a rate comparable to that of $\Delta\tau_{\mathrm{b}}$, the duration of a granulocyte burst. For example, murine HSC turnover rates μ_{h} are hypothesized to be 10-fold higher than those in primates while the clonal aging dynamics (and its timescale $\Delta\tau_{\mathrm{b}}$) are relatively conserved across species [CBG11]. According to our result, such a 10-fold

increase in HSC death rate would lead to a 10-fold increase in HSC clone extinction rate, bringing the lifespans of HSC clones closer to the (progenitor) clonal aging timescale $\Delta\tau_{\rm b}$. This interpretation is consistent with the fact that hematopoiesis in large primates have been described in terms of "clonal stability" while hematopoiesis in mice have been described in terms of "clonal succession" [JL90,APS95,DKC96,PPB96,MGD06,SRC14]. We thus predict that with even longer tracking (> 100 months), the "clonal succession" mechanism could also be significant in primates.

3.5 Appendices

3.5.1 Proof of Eq. (3.7):

To solve Eq. (3.6) for $\frac{{\rm d}P(h,t)}{{\rm d}t}$, we transform the equation using the probability generating function $Q(s,t) = \sum_{h=0}^{\infty} P(h,t)s^h$. We have also neglected the subscript i because our model is "neutral" and $P(h,t)$ can describe the size of any HSC clone i. If the HSC self-renewal rate is approximated as $r_{\rm h}(H(t)) \equiv r_{\rm h}(t)$, the solution for $Q(s,t)$ takes on the following form [Wan05]:

$$Q(s,t) = 1 - \frac{s-1}{(s-1)\phi(t) - \psi(t)}, \tag{3.17}$$

where

$$\psi(t) = e^{-\int_0^t (r_{\rm h}(t') - \mu_{\rm h}){\rm d}t'} \text{ and } \phi(t) = \int_0^t r_{\rm h}(t')\psi(t'){\rm d}t'. \tag{3.18}$$

Note that for $h \geq 1$,

$$Q^{(h)}(s,t) = \frac{\partial^h Q(s,t)}{\partial s^h} = \frac{h!(-\phi(t))^{h-1}\psi}{[(s-1)\phi(t) - \psi(t)]^{h+1}} \text{ and } P(h,t) = \frac{Q^{(h)}(0,t)}{h!} = \frac{\phi^{h-1}(t)\psi(t)}{(\phi(t)+\psi(t))^{h+1}}. \tag{3.19}$$

These solutions obey the initial condition $P(h,0) = \mathbb{1}(h,1)$ and as $t \to \infty$, $\psi(t) \to \psi(\infty) \in (0,1)$, $\phi \to \infty$, and $P(h,t) \to 0$. For $h = 0$, $P(0,t) = 1 - \frac{1}{\phi(t)+\psi(t)}$ and $P(0, t \to \infty) \to 1$, indicating eventual extinction at long times [Wan05, YSK15].

Using forms given in Eq. (3.19), since both ϕ and ψ are independent of h, we can define

$$\frac{P(h+1,t)}{P(h,t)} = \frac{\phi(t)}{\phi(t)+\psi(t)} \equiv \lambda(t). \tag{3.20}$$

Thus, the probability distribution $P(h,t)$ can be written as

$$P(h,t) = \frac{1}{\phi(t)+\psi(t)}\frac{\psi(t)}{\phi(t)+\psi(t)}\left(\frac{\phi(t)}{\phi(t)+\psi(t)}\right)^{h-1} = (1-P(0,h))(1-\lambda(t))\lambda(t)^{h-1}. \quad (3.21)$$

3.5.2 Mean-field approximation for $\frac{\mathrm{d}P(h,t)}{\mathrm{d}t}$

In this section, we validate two approximations used to analyze the dynamics of any clone's $h(t)$, the population of a single clone in the HSC pool. The first approximation is that the dynamics of $h(t)$ does not affect those of $H(t)$, thus $r_{\mathrm{h}}(H)$ (decoupling). This "mean-field" approximation is implemented by assuming the growth rate $r_{\mathrm{h}}(H)$ to be a "parametrically driven" force via $r_{\mathrm{h}}(H(t)) \approx r_{\mathrm{h}}(t)$ in Eq. (3.6). It is valid under $h \ll H$ (qualitatively verified by the sampled data), but breaks down if any clone dominates the whole stem cell population (see Chapter 3). The second approximation is that the fluctuations of $r_{\mathrm{h}}(H(t))$ can be neglected. At steady state, if fluctuations in $H(t)$ are sufficiently small, we can set all kinetic parameters (*e.g.*, $r_{\mathrm{h}},\ \mu_{\mathrm{h}},\ \omega$) in Eqs. (3.6), (3.8) and (3.9) as constants.

To justify these two approximations, we consider a stochastic-differential-equation description of the clonal dynamics, which is a random birth-death process with a regulated birth rate [Gar85, AA03],

$$\frac{\mathrm{d}h}{\mathrm{d}t} = (r_{\mathrm{h}}(H)-\mu_{\mathrm{h}})h + \sqrt{(r_{\mathrm{h}}(H)+\mu_{\mathrm{h}})h}\,\xi(t) \quad (3.22)$$

where $\xi(t)$ represents a Gaussian white noise process. If we decompose $r_{\mathrm{h}}(H(t)) = \bar{r}_{\mathrm{h}}(t)+\mathbf{c}(t)$ where $\bar{r}_{\mathrm{h}}(t) = r_{\mathrm{h}}(H(t))$ with the mean-field value $H(t)$ determined by the solution of Eq. (3.1) and $\mathbf{c}(t)$ a residual noise term with mean zero, we find

$$\frac{\mathrm{d}h}{\mathrm{d}t} = (\bar{r}_{\mathrm{h}}(t)-\mu_{\mathrm{h}})h + \mathbf{c}(t)h + \sqrt{(r_{\mathrm{h}}(H(t))+\mu_{\mathrm{h}})h}\xi(t) \quad (3.23)$$

The dominate contribution to $\frac{\mathrm{d}h}{\mathrm{d}t}$ can be found by comparing the magnitudes of the three terms on the right-hand-side of Eq. (3.23).

Shortly following transplantation, $H(t) \ll H^*$, where $r_{\mathrm{h}}(H^*) = \mu_{\mathrm{h}}$ defines the steady-state total population H^*. During these shorter times $\bar{r}_{\mathrm{h}}(t) > \mu_{\mathrm{h}}$ and the magnitude of the regulated growth term $(\bar{r}_{\mathrm{h}}(t)-\mu_{\mathrm{h}})h \sim \mathcal{O}(h)$. At long times when steady state holds, self-renewal and death balance each other and $(\bar{r}_{\mathrm{h}}-\mu_{\mathrm{h}})h \approx 0$. The second term $\mathbf{c}(t)h$ contains

the fluctuations in $r_{\rm h}(H)$ induced by fluctuations in H, which are of order $\mathcal{O}(\sqrt{H})$. The magnitude of $\mathbf{c}h$ thus depends on the specific form of $r_{\rm h}(H)$. The third term $\sqrt{(r_{\rm h}+\mu_{\rm h})h}$ represents the effect of intrinsic noise in the birth-death process and is of order $\mathcal{O}(\sqrt{h})$.

First, since $h \ll H$, $\mathrm{d}r_{\rm h}/\mathrm{d}h \sim (h/H)\mathrm{d}r_{\rm h}/\mathrm{d}H \ll \mathrm{d}r_{\rm h}/\mathrm{d}H$ and we can assume $r_{\rm h}(H) \approx r_{\rm h}(t)$ at all times. To justify our second approximation, we need to show $\mathbf{c}(t)h$ can be neglected at steady state. At steady-state, the first term vanishes so we need to show that $\mathbf{c}(t)h$ is much smaller than $\mathcal{O}(\sqrt{h})$, the magnitude of the third term. Actually, we will show that $\mathbf{c}h$ is of order $\mathcal{O}(\frac{h}{\sqrt{H}})$ for general functions $r_{\rm h}(H)$ as long as certain conditions are met. Equivalently, we show the noise term $r_{\rm h}(H)-\mu_{\rm h}$ is of order $\mathcal{O}((H^*)^{-\frac{1}{2}})$. At steady-state, we expand $r_{\rm h}(H)$ about $r_{\rm h}(H^*)=\mu_{\rm h}$:

$$r_{\rm h}(H) = \mu_{\rm h} + \sum_{k=1}^{\infty} \frac{1}{k!}(H-H^*)^k \frac{\mathrm{d}^k r_{\rm h}(H)}{\mathrm{d}H^k}\bigg|_{H=H^*} \tag{3.24}$$

Two conditions have to be met: (i) the second term on the right-hand-side of Eq. (3.24) is of order $\mathcal{O}((H^*)^{-\frac{1}{2}})$ while (ii) the last summation term is much smaller (say, of order $o((H^*)^{-\frac{1}{2}})$).

Next, we will show that these two conditions are met in both Logistic-type and Hill-type carrying capacity cases by demonstrating $|\frac{\mathrm{d}^k r_{\rm h}}{\mathrm{d}H^k}| \lesssim \mathcal{O}((H^*)^{-k})$. If it does, condition (i) is directly satisfied and condition (ii) is satisfied through Taylor's theorem.

With Logistic carrying capacity, the birth rate can be written as $r_{\rm h}(H) = -(p_{\rm h}-\mu_{\rm h})\frac{H}{H^*} + p_{\rm h}$ and satisfies

$$\left|\frac{\mathrm{d}r_{\rm h}(H)}{\mathrm{d}H}\right| = \frac{p_{\rm h}-\mu_{\rm h}}{H^*} \ll (H^*)^{-\frac{1}{2}}, \qquad \left|\frac{\mathrm{d}^h r_{\rm h}(H)}{\mathrm{d}H^h}\right| = 0 \quad \text{for } h = 2,3,4... \tag{3.25}$$

The Hill-type carrying capacity with coefficient c and half-saturation size $K_{\rm h}$ is written as:

$$r_{\rm h}(H) = \frac{p_{\rm h}K_{\rm h}^c}{H^c + K_{\rm h}^c}. \tag{3.26}$$

Note $K_{\rm h}$ is of the same scale as H^* and their relation is $H^* = \left(\frac{p_{\rm h}}{\mu_{\rm h}}-1\right)^{\frac{1}{c}} K_{\rm h}^c \equiv \gamma K_{\rm h}$. We have

$$\left|\frac{\mathrm{d}r_{\rm h}(H)}{\mathrm{d}H}\right| = \frac{p_{\rm h}K_{\rm h}^c \cdot cH^{c-1}}{(K_{\rm h}^c + H^c)^2}\bigg|_{H=H^*} \sim \frac{1}{H^*} \tag{3.27}$$

and

$$\left|\frac{\mathrm{d}^2 r_{\mathrm{h}}(H)}{\mathrm{d}H^2}\right| = p_{\mathrm{h}} c K_{\mathrm{h}}^{c} \frac{[(c-1)\gamma^{c-2} - (c+1)\gamma^{3c-2} - 2\gamma^{2c-2}]K_{\mathrm{h}}^{3c-2}}{(1+\gamma^{c})^4 K_{\mathrm{h}}^{4s}} \sim \frac{1}{K_{\mathrm{h}}^2} \sim \frac{1}{(H^*)^2}. \tag{3.28}$$

More formally, we can show that if the order of $\frac{d^c r_{\mathrm{h}}}{dH^c}$ is $\mathcal{O}(\frac{1}{H^c})$, then $\frac{d^{c+1} r_{\mathrm{h}}}{dH^{c+1}}$ is of order $\mathcal{O}(\frac{1}{H^{c+1}})$, because

$$\frac{d}{dH}\left(\frac{H^b + ...}{H^a + ...}\right) = \frac{(bH^{b-1} + ...)(H^a + ...) - (H^b + ...)(aH^{a-1} + ...)}{(H^a + ...)^2} \sim \frac{H^{b-1} + ...}{H^a + ...} \tag{3.29}$$

where a and b ($a \neq b$) are the highest orders of denominator and nominator respectively.

3.5.3 Example of alternative model:

An alternative model to the one we have analyzed allows younger-generation progenitor cells ($\ell < L$) to differentiate into peripheral blood. Since each generation can differentiate with rate ω, the progenitor cell dynamics is slightly modified from those in our main model:

$$\frac{\mathrm{d}n^{(\ell)}(t)}{\mathrm{d}t} = \begin{cases} \mathrm{Poisson}(\alpha h(t)) - (r_{\mathrm{n}} + \mu_{\mathrm{n}} + \omega)n^{(0)}(t), & \ell = 0, \\ 2r_{\mathrm{n}} n^{(\ell-1)}(t) - (r_{\mathrm{n}} + \mu_{\mathrm{n}} + \omega)n^{(\ell)}(t), & 1 \leq \ell \leq L-1, \\ 2r_{\mathrm{n}} n^{(L-1)}(t) - (\omega + \mu_{\mathrm{n}}^{(L)})n^{(L)}(t), & \ell = L. \end{cases} \tag{3.30}$$

Moreover, the dynamics of the mature peripheral blood obeys

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \sum_{\ell=0}^{L} \omega n^{(\ell)}(t) - \mu_{\mathrm{m}} m(t). \tag{3.31}$$

The solution to Eqs. (3.30) and (3.31) following a single differentiation event is

$$\begin{aligned} n_{\mathrm{b}}^{(\ell)}(t) &= \frac{(2r_{\mathrm{n}})^{\ell}}{\ell!} t^{\ell} e^{-(r_{\mathrm{n}} + \mu_{\mathrm{n}} + \omega)t}, \\ n_{\mathrm{b}}^{(L)}(t) &= e^{(r_{\mathrm{n}} - \mu_{\mathrm{n}} - \omega)t}\left[1 - \frac{\gamma(L+1, 2r_{\mathrm{n}}t)}{L!}\right], \\ m_{\mathrm{b}}(t) &= \omega \int_0^t \sum_{\ell=0}^{L} n_{\mathrm{b}}^{(\ell)}(\tau) e^{-\mu_{\mathrm{m}}(t-\tau)} \mathrm{d}\tau \end{aligned} \tag{3.32}$$

These results can be applied to the model and analyzed and simulated using the same procedures as described earlier. However, certain parameters have to be re-interpreted. For

example, using the same value of $\omega = 0.16$ will significantly increase the effective death rate for progenitor cells of each generation. Fortunately, as we will show later, this alternative mechanism should not affect our main conclusion as the parameter-fitting results are not sensitive to the exact shape of cell bursts.

3.5.4 Proof of extinction time:

As a function of the initial number h of HSCs in a clone, the mean extinction time (MET) $T(h)$ under the steady-state approximation $r_{\rm h} = \mu_{\rm h}$ obeys [Gar85, All10]

$$[T(h+1) - T(h)]\mu_{\rm h} h - [T(h) - T(h-1)]\mu_{\rm h} h = -1. \tag{3.33}$$

with an absorbing boundary condition $T(0) = 0$. By iterating Eq. (3.33), we find

$$T(h+1) - T(h) = T(1) - \frac{1}{\mu_{\rm h}} \sum_{k=1}^{h} \frac{1}{k}, \tag{3.34}$$

which can be again iterated to obtain

$$T(h) = hT(1) - \frac{1}{\mu_{\rm h}} \sum_{k=1}^{h-1} \sum_{\ell=1}^{k} \frac{1}{\ell}. \tag{3.35}$$

To solve for $T(1)$, we invoke a reflecting boundary condition $T(H_{\rm ss}) - T(H_{\rm ss} - 1) = 1/(\mu_{\rm h} H_{\rm ss})$ [DSS05], where

$$T(H_{\rm ss}) = H_{\rm ss} T(1) - \frac{1}{\mu_{\rm h}} \sum_{k=1}^{H_{\rm ss}-1} \sum_{\ell=1}^{k} \frac{1}{\ell}, \qquad T(H_{\rm ss} - 1) = (H_{\rm ss} - 1)T(1) - \frac{1}{\mu_{\rm h}} \sum_{k=1}^{H_{\rm ss}-2} \sum_{\ell=1}^{k} \frac{1}{\ell}, \tag{3.36}$$

to find

$$T(1) = \frac{1}{\mu_{\rm h}} \sum_{\ell=1}^{H_{\rm ss}} \frac{1}{\ell}. \tag{3.37}$$

Upon using Eq. (3.37) in Eq. (3.35), we find

$$T(h) = \frac{h}{\mu_{\rm h}} \sum_{k=1}^{H_{\rm ss}} \frac{1}{k} - \frac{1}{\mu_{\rm h}} \sum_{k=1}^{h-1} \sum_{\ell=1}^{k} \frac{1}{\ell} \equiv T_{\rm discrete}(h), \tag{3.38}$$

which is the MET for a discrete system.

We can also approximate $T(h)$ by considering h as a continuous variable and replace the summations in Eq. (3.38) by integrations to find a simpler, more insightful approximation

to $T(h)$:

$$\begin{aligned} T_{\text{continous}}(h) &= \frac{h}{\mu_{\text{h}}} \int_{\ell=1}^{H_{\text{ss}}} \frac{\mathrm{d}\ell}{\ell} - \frac{1}{\mu_{\text{h}}} \int_{k=1}^{h-1} \mathrm{d}k \int_{\ell=1}^{k} \frac{\mathrm{d}\ell}{\ell} \\ &= \frac{h \ln H_{\text{ss}} - (h-1)\ln(h-1) + h - 2}{\mu_{\text{h}}} \\ &\approx \frac{h}{\mu_{\text{h}}} \left(\ln \frac{H_{\text{ss}}}{h} + 1 \right), \end{aligned} \tag{3.39}$$

where we have used $\int^x (1/x')\mathrm{d}x' = \ln x$ and $\int^x \ln x' \mathrm{d}x' = x \ln x - x$. The continuous approximation to the MET matches the exact result quite well (relative error $\lesssim 5\%$) for all values of h.

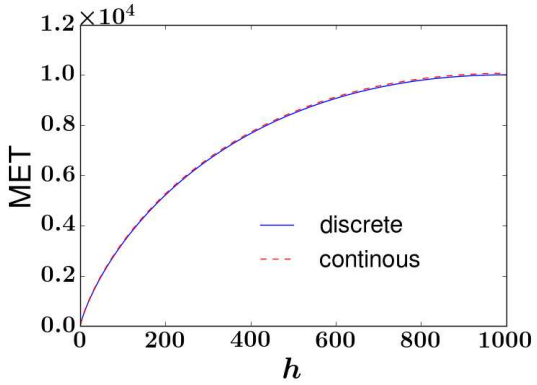

Figure 3.14: Mean extinction time as a function of stem cell clone size h

3.5.5 Simulation scheme:

To generate predictions, we first choose values of $\theta_{\text{model}} = \{\lambda, C_{\text{h}}, r_{\text{n}}, L_{\text{e}}\}$ and simulate our model, including sampling, to find $s_i(t_j)$. Each realization of a simulation of the model is performed by

1. Specify the static HSC clone size distribution $P(h)$ by choosing the pair (λ, C_{h}) and draw $\{h_i\}$ from the geometric distribution C_{h} times using the Python package np.random.geometric. Normalize to construct the configuration $\{h_i\}/H_{\text{ss}}^{+} \equiv \{\frac{h_i}{\sum_{i=1}^{C_{\text{h}}} h_i}\}$. Alternatively, we can also use the data $\hat{y}_i$ to approximate the configuration $\{h_i\}/H_{\text{ss}}^{+}$.

2. Fixing all parameters θ_{model}, construct the total clone i differentiation rate $\alpha h_i \equiv A^+_{\text{ss}} h_i / H^+_{\text{ss}} = 2^{-L_{\text{e}}} \hat{M}^+_{\text{ss}} \mu_{\text{m}} h_i / H^+_{\text{ss}}$ for each clone i. Generate realizations of sets of HSC differentiation event times $\{\tau_k^{(i)}\}$ for each clone i based on the rate $\alpha h_i = A^+_{\text{ss}} h_i / H^+_{\text{ss}}$.

3. Evaluate Eqs. (3.11). Sum up the peripheral blood bursts initiated by each differentiation event of each clone i to find $m_i(t) = \sum_k m_{\text{b}}(t - \tau_k^{(i)})$.

4. Sample a fraction $\varepsilon(t_j) = \frac{\hat{S}^+(t_j)}{\hat{M}^+(t_j)}$ of the total peripheral cell count $M^+(t_j) = \sum_i m_i(t_j)$. Here, $\hat{S}^+(t_j)$, $\hat{M}^+(t_j)$, and the times t_j are defined by the experiment (we used the Python package numpy.random.binomial). The cell counts of each clone are $s_i(t_j)$. Use the simulated total tagged cell counts in the samples $S^+(t_j) = \sum_i s_i(t_j)$ to normalize $\frac{s_i(t_j)}{S^+(t_j)} = f_i(t_j)$. Up to this point, we have generated a data matrix $f_i(t_j)$ of size $C_{\text{h}} \times J$.

5. Increment L_{e} within the desired interval and repeat steps 2-4 200 times. For each value of L_{e}, the 200 simulations generate 200 $f_i(t_j)$ matrices. These repeats are to ensure that the noise induced from drawing values of h_i from $P(h)$ and sampling $s_i(t_j)$ from $m_i(t_j)$ do not significantly corrupt our parameter estimation.

The simulated, model-derived configurations $f_i(t_j)$ are then compared with experimentally measured values $\hat{f}_i(t_j)$. The parameter L_{e} that minimizes the mean-squared error will be chosen as the least-squares estimate L^*_{e}.

3.5.6 PCR bias

Each blood sample drawn from rhesus macaque RQ5427 contains about 10 μg genomic DNA [KKP14]. After PCR amplification, intensive sequencing, and data filtering, the total number S^+_j of quantifiable VISs correspond to $5 \times 10^3 \sim 3 \times 10^4$ tagged cells. The sample ratio is defined by $\varepsilon_j \equiv S^+_j / M^+_{\text{ss}} = 3 \times 10^{-5} \sim 2 \times 10^{-4}$ where $M^+_{\text{ss}} \approx 1.6 \times 10^8 \gg S^+_j \gg 1$ are the total numbers of VIS-tagged cells in the PB pool and in the j^{th} sample respectively. The time intervals $\Delta t_j \equiv t_{j+1} - t_j$ between consecutive samples ranged between 5 and 11 months. Besides the bursty dynamics described above, the data shown in Figure 3.1(a) are subject to the effects of small sampling size, uncertainty and bias induced by experiment

procedures such as PCR amplification and data filtering, and low sampling frequencies that may screen out hidden dynamics occurring between consecutive samples.

To explore how sample size affects $f_i(t_j)$, let us study a specific clone whose abundance in the peripheral blood is $m_i(t_j)/M_{\rm ss}^+ = 0.01$ at a specific sample time t_j, corresponding to $m_i(t_j) = 1.6 \times 10^6$. Sampling includes three detailed steps. In each step, the number of sampled cells obeys a binomial distribution with mean pS and standard deviation $\sqrt{p(1-p)S}$ if the number of trials is S ("sample size") and the success probability ("clone frequency") is p in that step. If the average number of sampled cells is large, the binomial sampling distribution can be approximated by a Gaussian distribution, where the relative standard deviation (RSD) defined as the standard deviation divided by the mean $\sqrt{\frac{1-p}{pS}}$ can be calculated to evaluate how noisy this sampling step is.

- Each time, a 10 μg blood sample were taken from the rhesus macaque, in which about 7.7×10^4 EGFP+ cells were obtained. The sampling ratio is about $7.7 \times 10^4 / 1.6 \times 10^8 \approx 4.8 \times 10^{-4}$. So, the average sampled size of this clone is 770 and the standard deviation is 27.7. The RSD is 3.6%.

- PCR amplification. A previous study has shown that a PCR starting with k copies of DNA will provide an estimate of mean k and standard deviation $\sqrt{\frac{2-q}{q}k}$ [PJ96], where q is the amplification factor. In the current experiment, the correlation between the relative frequencies and dilution factor in this type of clone-specific PCR amplification is between 0.989 to 0.999 [KKP14], corresponding to a $q \geq 1.5$. So if the input is 770 cells, the standard deviation is less than 16, resulting in RSD less than 2%.

- Second sampling with size 7.7×10^3 was taken from the amplified sequence pool. If the input is $0.01 \times 7700 = 77$, then the average is 77 and the standard deviation is 8.7; the RSD is 11%.

Thus in this system PCR generates a smaller uncertainty than sampling. We will approximate this three-step sampling by a one-step sampling of size $S^+ = 7700$ cells directly from the PB pool $M_{\rm ss}^+ = 1.6 \times 10^8$ in the rest of the paper.

3.5.7 Alternative statistical measure

We developed our data analysis based on the statistics of the quantity y_i, the time averaged relative clone sizes for those clones exhibiting z absences across their longitudinal samples. While reasonable parameter estimates were obtained from fitting to data, we also considered alternative objective functions. Specifically, we looked at the standard deviation $\sigma_i = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(f_i(t_j) - y_i)^2}$ quantifying the temporal fluctuations of the relative sizes of each clone i. The way we construct an alternative objective function is similar to the way we constructed Y_z. Recall for Y_z, we calculated the average abundance across only those clones with the same $z_i = z$ absences across time. However, unlike z_i which takes a finite set of discrete values $\{1,\ 2,\ ...,\ J-1\}$, σ_i is a continuous variable so we have to artificially bin their values. Instead, we bin clones with similar y_i and study the average of their associated σ_i's. Since the distribution y_i is non-linear with a long tail, we evaluated $\ln y_i$ to obtain the near-linear distribution shown in Figure 3.15(a) and sorted $\ln y_i$ into equal-width bins and calculated the average of the associated σ_i's. Dividing the values of $\ln y_i$ into bins labeled by k, we compute

$$U_k = \frac{\sum_i \sigma_i \mathbb{1}(\text{clone } i \in \text{bin } k)}{\sum_i \mathbb{1}(\text{clone } i \in \text{bin } k)} \tag{3.40}$$

in analogy with the definition of Y_z. The objective function can be straightforwardly defined as

$$\text{MSE}_\sigma(\theta_{\text{model}}) = \sum_k (U_k(\theta_{\text{model}}) - \hat{U}_k)^2. \tag{3.41}$$

It is also unclear how to set upper and lower bounds on the range of y_i for comparison (in contrast to the natural bound on $1 \leq z \leq J-1$) because an unconstrained set of clones will be sensitive to the underlying h_i distribution (an undesirable property). In Figure 3.15(b) we fit the data from animal RQ5427 using MSE_σ and find $L_{\text{e}}^* \approx 24.4$, consistent with our previous estimate using Y_z.

While it is also possible to choose σ_i as a measure of clone population fluctuations, we list several advantages of $\hat{z}_i$ over σ_i for the current dataset. Note that the number of disappearances z_i of each individual clone are defined on a finite set of integers (unlike the continuously measured σ_i), making it easier to bin clones with the same z values. Different

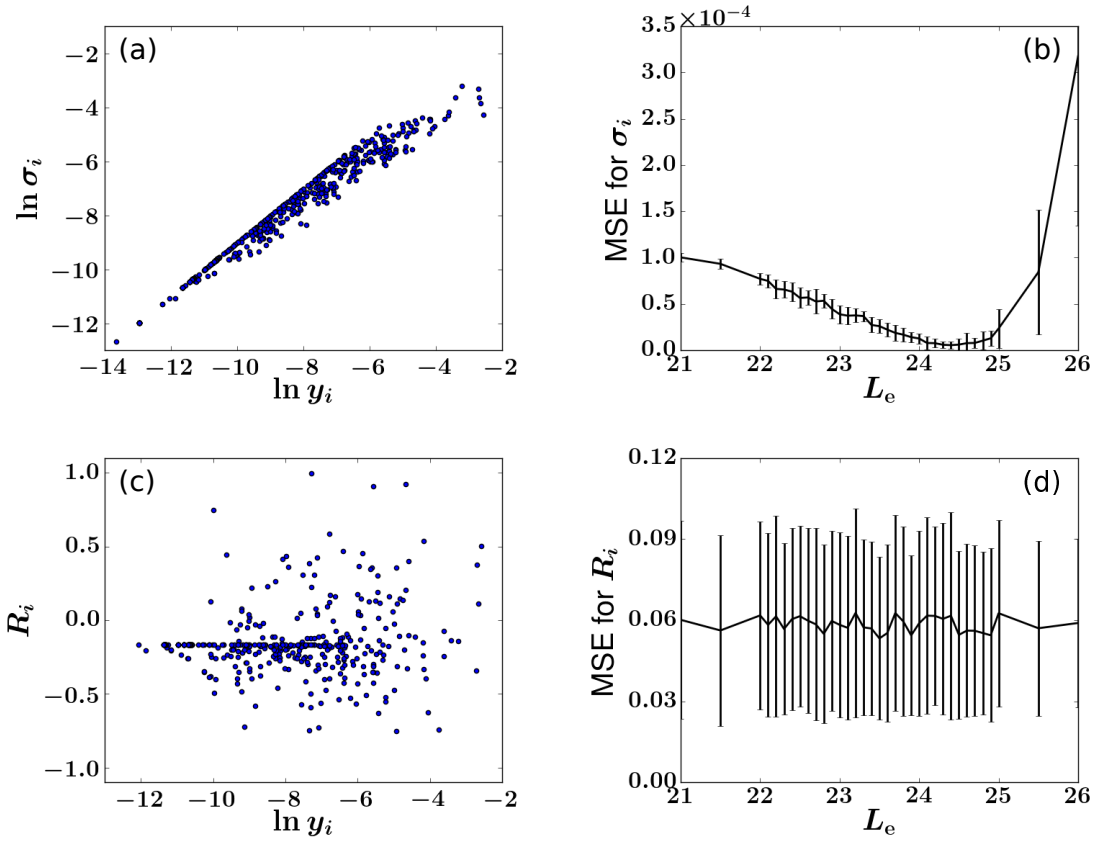

Figure 3.15: Alternative statistical measures of clonal abundances

Statistics of the two alternative fluctuation measures and their fitting results. Each dot represents a clone. (a) Log standard deviation *vs* log average abundances. Clones are near-linearly distributed in the log average abundance space. (b) Objective function MSE_σ *vs.* L_e. Clones of similar y_i are binned and their averaged σ_i were used to compute U_k. (c) Autocorrelations R_i *vs.* log of average abundances u_i. There is no clear pattern in the distribution of R_i's. (d) MSE_R *vs.* L_e. This objective function cannot resolve the LSE L_e^*.

clones i will exhibit different time-averaged abundances y_i, but may have the same value of z_i. As shown in Figure 3.4, the larger $\hat{z}_i$ is, the smaller the corresponding $\ln \hat{y}_i$ tends to be. The robust correlation between z_i and y_i encodes the level of fluctuations for a clone of certain size. For a given y_i, the larger z_i, the more "bursty" the dynamics, implying a smaller number of tagged HSC differentiations per unit time (a smaller A_ss^+).

Another advantage of using z_i statistics emerges when fitting model results to the pattern

of the measured data in Figure 3.4. Average sizes y_i (and the underlying h_i) associated with clones having $1 \leq z \leq 7$ all contain at least one absence. This constraint naturally controls the upper and lower bounds of h_i in a particular z bin ($1 \leq z \leq 7$), based on the burstiness of the model. Exact knowledge of the configuration $\{h_i\}$ is not required for fitting these y_i data. Thus, dividing clones into z bins provides us with a natural way to exclude unconstrained clones sizes. In other words, the theoretical values of y_i (and the underlying h_i) associated with bin $z_i = 0$ can be arbitrarily and unreasonably large and such a possibility should be excluded. Similarly, all y_i below a threshold size generate $z_i = J$ (clones that never appeared in the sampled blood) and does not provide any statistical power. This advantage of using z_i can also be confirmed by visual inspection of Figure 3.9(b). Several very large clones do not follow the general statistical pattern and show extremely large variances. Without manually filtering out these clones, our fitting in Figure 3.15(b) results in a larger $L_{\rm e}^* = 24.4$ than the $L_{\rm e}^* = 23.4$ obtained earlier using Y_z statistics.

Finally, another option for comparing model with data is to use correlation functions. In this approach, the sampling gap Δt_j varies between $5-11$ months so the usual autocorrelation function with equal time gaps cannot be rigorously defined. We use the one-sample-gap autocorrelation function

$$R_i = \frac{1}{(J-1)\sigma_i^2} \sum_{j=1}^{J-1} (f_i(t_j) - y_i)(f_i(t_{j+1}) - y_i), \tag{3.42}$$

and bin values of $\ln y_i$ in analogy to Eq. (3.40) to define

$$W_k = \frac{\sum_i R_i \mathbb{1}(\text{clone } i \in \text{bin } k)}{\sum_i \mathbb{1}(\text{clone } i \in \text{bin } k)} \tag{3.43}$$

and construct an autocorrelation-based objective function

$$\text{MSE}_R(\theta_{\text{model}}) = \sum_k (W_k(\theta_{\text{model}}) - \hat{W}_k)^2. \tag{3.44}$$

Since the inter-sample intervals Δt_j are larger than a typical burst size $\Delta\tau_{\rm b} \approx 32$ days, so cells in different samples likely originate from different HSC differentiation events. Thus, the fluctuations of clone sizes are uncorrelated from sample to sample, as shown in Figure 3.15(c). The values of R_i are randomly distributed between -1 and 1, and centered about the

line $R = \frac{1}{2-J}$, corresponding to the majority of clones that have $z_i = J - 1$ (only 1 non-zero sample). Data fitting using R_i and MSE_R is ill-conditioned and cannot resolve L_{e}^*, as shown in Figure 3.15(d).

3.6 Extended studies

3.6.1 Simulating clone samples under different time gaps

One of the major constraint of taking blood samples from primates is that the frequency (or time gap) of sampling is usually low (or large) for ethical and financial reasons. It is interesting to get a sense of how such gap actually affects the sampled data by taking advantage of computer simulations on which we can tune the gap arbitrarily small. We simulated one realization of $m_i(t)$ in our optimized model ($L_{\mathrm{e}} = 23.4,\ \lambda = 0.99$) and saved this dataset. We then take samples $f_i(t_j)$ of all clones from the saved data. This process is repeated for four times, each with a different sampling gap of 10 month, 1 month (the typical time scale for a "burst"), 10 days, and 1 day (the typical lifespan for a granulocyte). The results are plotted in Figure 3.16.

We observe that as sampling gap decreases, the observed level of fluctuations increases. So indeed, rich dynamics may be hidden between infrequent samples. Nevertheless, this uncertainly does not bring systematical bias to our fitting results since we use the same time gaps as in the experiment. Moreover, as shown by Figure 3.7(a), our proposed statistics Y_z is able to capture the subtle level of clonal fluctuation (which is controlled by L_{e}) emerging on that specific scale of sampling gap. As shown by Figure 3.8, simulated data generated under a different $L_{\mathrm{e}} \neq L_{\mathrm{e}}^*$ would not make a good match to the experimental data.

3.6.2 Reconstructing $\{h_i\}$

In our model, the dynamics of any clone may be approximately summarized by

$$f_i(t) \sim \mathcal{F}(h_i(t)). \tag{3.45}$$

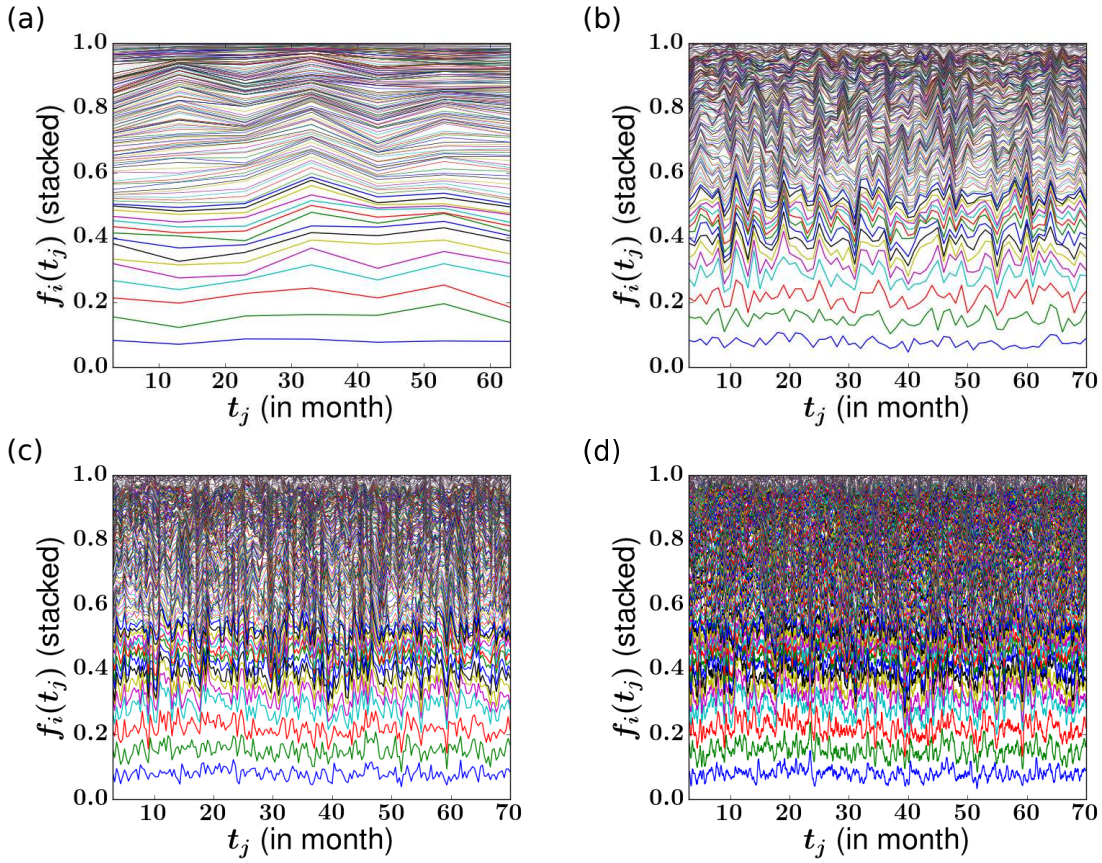

Figure 3.16: Simulated samples under different sampling gaps

In the optimized model ($L_e = 23.4,\ \lambda = 0.99$), we simulated one realization of $m_i(t)$ and sampled it with sampling gaps 10 month (a), 1 month (b), 10 days (c), and 1 day (d).

This is approximate since $f_i(t)$ is contributed by the stem cell activities in the time period $t - \tau_b \leq \tau \leq t$ where τ_b is the characteristic time lag between a stem cell's differentiation and the death of its progeny granulocytes (see Subsection 3.6.3 for more details). What we investigated in the main body of this chapter was choosing the Y_z statistic that is sensitive to the bursty function $\mathcal{F}$ but is robust to the unknown $\{h_i\}$ configuration. In Figure 3.17, we plot how the sampled abundance f_i of a clone correlates to its fraction $g_i \equiv h_i/H^+$ in the stem cell pool. Because of the stochasticity in $\mathcal{F}$, g_i corresponds to a distribution of f_i. The average of f_i is equal to g_i, simply because the clonal dynamics are neutral. The (relative) variance, however, can be quite large, especially for small clones.

After obtaining the bursty dynamics $\mathcal{F}$, a natural next step is to explore the possibility

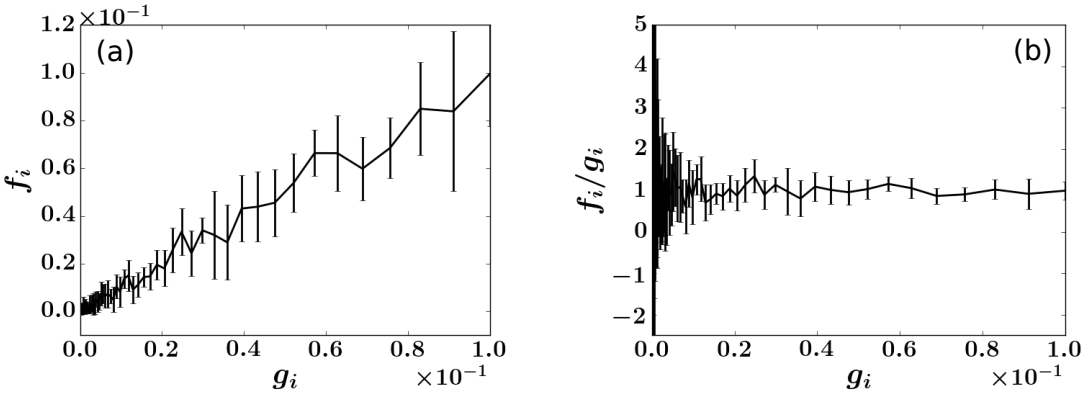

Figure 3.17: Correlations of stem cell clone fractions and sampled fractions

Simulated statistics of (a) f_i and (b) f_i/g_i. We have set $L_\mathrm{e} = 24$. The range $0 < g \lesssim 0.1$ is set to be constant throughout the simulation. For each of the 100 values of f_i a simulation of the full model is performed and eight sampled $f_i(t_j)$ are taken, from which we plot the mean f_i (solid line) and standard deviations (error bars).

of reconstructing $\{h_i\}$ from blood samples. For this purpose, we define the fraction of clones with average abundance f as the "count of average" (COA)

$$\mathrm{COA}(f) = \frac{1}{C_\mathrm{s}} \sum_{i=1}^{C_\mathrm{s}} \mathbb{1}\left(\frac{1}{J} \sum_{j=1}^{J} f_i(t_j), f \right). \tag{3.46}$$

By fixing a certain L_e (or A^+) and adjusting λ, we can fit the $\mathrm{COA}(f)$ from our model with that obtained from data. The fraction of all clones at each time point that have abundance f ("average of count", or AOC), defined by

$$\mathrm{AOC}(f) \equiv \frac{1}{J} \sum_{j=1}^{J} \left[\frac{1}{C_\mathrm{s}} \sum_{i=1}^{C_\mathrm{s}} \mathbb{1}\left(f_i(t_j), f \right) \right], \tag{3.47}$$

was previously considered by [GKC15]. Compared to COA, AOC does not track individual clones across time since it aggregates the counts at each time point, allowing individual clones to exchange their contributions. Previously, AOC was found to fit well by assuming a homogeneous-size stem cell pool $h_i \equiv 1$ (i.e. setting $\lambda = 0$ in our model) [GKC15]. This result suggests that fitting AOC does not require knowledge of $\{h_i\}$ but only requires adjusting the downstream dynamical parameters. In comparison, fitting COA requires tuning both A^+

and λ. Although it is easier to fit, AOC(f) cannot distinguish individual clones as it does not encode clone identity information across time samples. For investigating the *persistence in time* of clone sizes that are determined by g_i and encoded in $f_i(t_j)$, COA(f) should be a more suitable metric for our system even though constrains the dynamics of each individual clone. Such persistence of clone sizes $f_i(t_j)$ over time encodes the information of the true distribution of HSC clone sizes g_i. We expect that AOC is less sensitive to λ, the controlling factor of the g_i distribution, than COA, which is confirmed by Figure 3.18. On the other hand, AOC is more sensitive to A^+, implying that the non-uniform distribution of AOC emerges as a result of the stochasticity in the HSC differentiation and subsequent processes. The magnitude of A^+ controls the distribution of AOC and is insensitive to the distribution of g_i, which is consistent to the analysis in the previous study [GKC15]. The easier fitting of AOC relative to COA, specifically for the long term experiment on animal 2RC003, indicates that a slow evolution of the clone abundance distribution of the HSC pool.

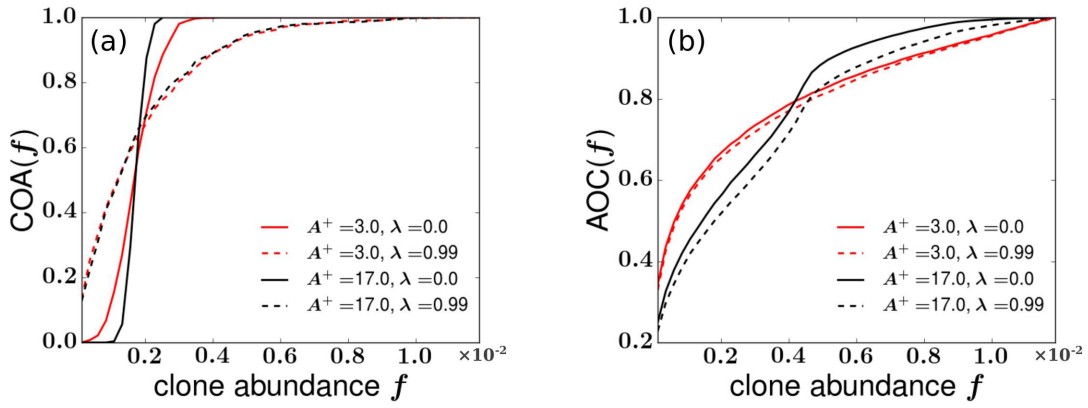

Figure 3.18: Different statistical features encoded in AOC and COA

The dashed and dotted curves represent simulations with $\lambda = 0$ and $\lambda = 0.99$, respectively. The red curve is associated with $A^+ = 3$ while the black curve was generated using $A^+ = 17$. (a) Simulated COAs are more sensitive to λ than to A^+. (b) AOC derived from simulations are found to be sensitive to changes in A^+, but not to changes in λ.

3.6.3 De-convolving the multiplicative noise

In subsection 3.6.2, we talked about the difficulty of modeling the clonal dynamics which includes the hidden $\{h_i\}$ configuration and the bursty dynamics $\mathcal{F}$. Yet another confounding source of the problem is the fluctuations of $\{h_i\}$ over time. The observed clonal fluctuations are contributed by stochasticity in three consecutive stages:

$$\text{Stem cell clone fluctuations} + \text{Progenitor/mature cell bursty dynamics} + \text{Sampling effect.}$$

This is the so-called "multiplicative noise" problem, where the level of fluctuations of the next stage depends on the state of the current stage. For example, the larger m_i is, the less relative variance we would expect to see in s_i over samples. But since m_i itself is fluctuating in time, it is difficult to determine how much of the observed variance is contributed by the sampling effect. The way h_i affects m_i is similar.

Besides the difficulty of de-convolving the observed variances, there is no analytic function for measuring noise induced by the bursty dynamics, unlike sampling noise which is relatively easy to model. In the main text, we simplify the problem by ignoring the fluctuations of $\{h_i\}$ and focusing on the bursty dynamics. This inevitably exaggerates the estimate for L_{e}^*, which can be severe for fitting the murine data since mice HSCs are speculated to turnover much faster than those of primates [CBG11]. Figure 3.19 shows how the fluctuations of h_i may contribute even more to the observed variance than the bursty dynamics and the sampling effect for clone i. The red solid curve is a hypothetical realization of the relative abundance g_i between samples j and $j+1$. The variability simply due to the bursty dynamics and sampling is described by a sharper distribution (black curve) while the experimental variation is described by the broader, red distribution. For large clones, this suggests stochastic fluctuations in h_i contribute to most of the variance between $f_i(t_j)$ and $f_i(t_{j+1})$.

To model fluctuation of HSCs, a major difficulty is still that the distribution of $\{h_i\}$ after the initial phase of fast HSC self-renewal is unknown. To bypass this difficulty, we again study statistical patterns that emerges from the observed data regardless of the underlying $\{h_i\}$. This time, we pair a clone's sampled abundances in consecutive samples to study

$$P(f_i(t_{j+1})|f_i(t_j)), \tag{3.48}$$

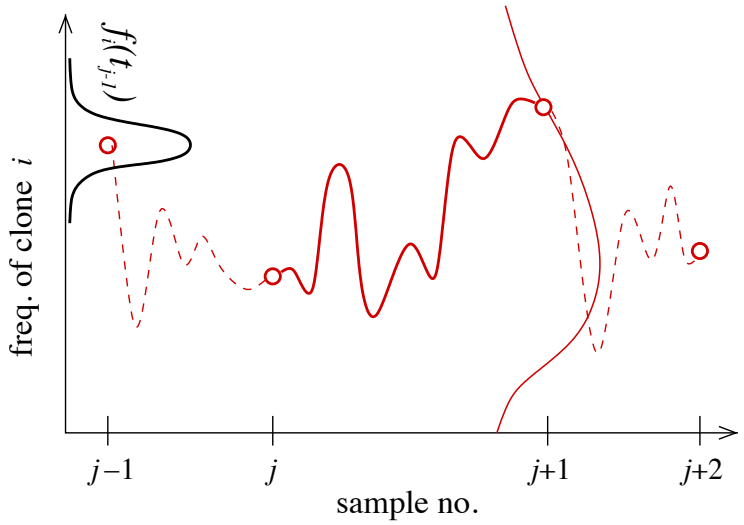

Figure 3.19: Evolving stem cell clones may induce large variances

Thick solid red line describes the trajectory of a stem cell clone's fraction $g_i(t) = h_i/H^+$ over time. Thin solid red line denotes the probability distribution of $g_i(t_{j+1})$ given a specific $g_i(t_j)$. Dashed red line denotes possible trajectories of the stem cell clone's fraction before t_j and after t_{j+1}. The narrower black distribution of the sampled $f_i(t_j)$ arises from variation due to bursty dynamics and sampling.

the probability distribution of $f_i(t_{j+1})$ given $f_i(t_j)$. An immediate advantage of using this statistics instead of collecting clone-wise variables (y_i, σ_i, and z_i) is that more data points can be obtained ($C_\mathrm{s} \times (J-1)$ compared to C_s before). .

Another advantage emerging from considering this statistic is that one can now model the change of $\{h_i\}$ across samples. Consider the following inference/derivation procedure that maps $f_i(t_j)$ to a distribution of $f_i(t_{j+1})$

$$f_i(t_j) \xrightarrow{①} h_i(t_j - \Delta\tau_\mathrm{b}) \xrightarrow{②} h_i(t_{j+1} - \Delta\tau_\mathrm{b}) \xrightarrow{③} f_i(t_{j+1}). \tag{3.49}$$

Transition ① denotes reconstructing $h_i(t_j - \Delta\tau_\mathrm{b})$ from $f_i(t_j)$. This is possible (but difficult) because the observed $f_i(t_j)$ comes essentially from the cumulative activities of stem cells in the same clone during the time period $(t_j - \Delta\tau_\mathrm{b},\ t_j)$. Given $h_i(t_j - \Delta\tau_\mathrm{b})$, one can calculate the probability of observing a specific trajectory $h_i(t_j - \Delta\tau_\mathrm{b} < t < t_j)$, which determines the Poissonian rate of stem cell differentiations during that period. These differentiations

generate progenitor and mature cell bursts successively in time, which stack up at time t_j to determine $m(t_j)$ and thus $s(t_j)$. The full probability reads

$$P(h_i(t_j - \Delta\tau_\mathrm{b})) \cdot \sum_{\{h_i(t) \text{ trajectories}\}} P(h_i(t) \text{ trajectory}|h_i(t_j - \Delta\tau_\mathrm{b})) \cdot \sum_{m_i(t_j)} P(m_i(t_j)|h_i(t) \text{ trajectory}) \cdot P(f_i(t_j)|m_i(t_j)). \tag{3.50}$$

To realistically approximate this probability, we assume that h_i does not change over the period $\Delta\tau_\mathrm{b} \approx 1$ month. Note that it is NOT assuming that h_i does not change over $\Delta\tau_j = 5 - 11$ months, the gap between two consecutive samples, thus imposing a less stringent constraint to the model. Now use Bayes' rule

$$P(h_i(t_j)|f_i(t_j)) \cdot P(f_i(t_j)) = P(f_i(t_j)|h_i(t_j)) \cdot P(h_i(t_j)) \tag{3.51}$$

where $P(f_i(t_j)) = 1$ and $P(f_i(t_j)|h_i(t_j))$ comes from the bursty dynamics (controlled by L_e). The prior $P(h_i(t_j))$ is a geometric distribution with unknown parameter λ.

Step ② is the random walk of h_i from time t_j to t_{j+1} with rate μ_h. The relative persistence h_i is the only reason why $f_i(t_{j+1})$ correlates with $f_i(t_j)$, because NO granulocyte/progenitor that appears in/contributes to the j-th sample will still appear in/contribute to the $(j+1)$-th sample, thanks to their short lifespans. Step ③ is the usual bursty dynamics.

Overall, to simulate Eq. (3.49), we can specify a set of key parameters $\theta = \{\mu_\mathrm{h}, \lambda, L_\mathrm{e}\}$. Given θ and any desired $f_i(t_j)$, we can construct a distribution of $P(h_i(t_j)|f_i(t_j))$ according to Eq. (3.51). This distribution will then be employed to generate a distribution of $P(h_i(t_{j+1})$, which is then used to simulate a distribution of $f_i(t_{j+1})$ with the same L_e. We can compare this simulated distribution to the experiment and find an optimal θ^* by a grid search. The range of $f_i(t_j)$ will be chosen to make sure that the experimentally observed abundances $\hat{f}_i(t_j)$ are fully covered.

3.6.4 Tracking the long-term evolution of HSC clones

Tracking the change of the $\{h_i\}$ configuration is one of the primary goals of this study. In the main text, we considered intensive HSC self-renewal in the short term and assumed

static HSC pool in the long term as a "baseline" model. Based on the results of the baseline model, we are now in a position to investigate the (slow) birth-death process as well as other random or deterministic mechanisms that may affect HSC clone sizes in homeostasis, such as telomere shortening (HSC "aging" [SRM11]), HSC dormancy and reactivation [WLO08, WLT09, TEW10], and changes of lineage bias in the hematopoietic output [KKP14], *etc.*

Our first explorative strategy is to identify "outlier" clones that significantly deviate from our baseline model. As shown in Figure 3.9, projecting time series data of clones onto the average-standard deviation space makes it easy to identify clones that have much higher/lower level of fluctuations than what would be expected from our optimally fitted model. We pick four of these outlier clones that have large y_i (red squares in Figure 3.20(a)) and plot their abundance data in all samples (Figure 3.20(b)). These clones show a qualitative trend to expand their sizes over time (except for the last sample). To compare them with "normal" clones, we pick five "normal" ones (blue circles in Figure 3.20(a)) that have comparably large y_i and plot their their abundance data (Figure 3.20(c)), which do not show an obvious trend to increase or decrease.

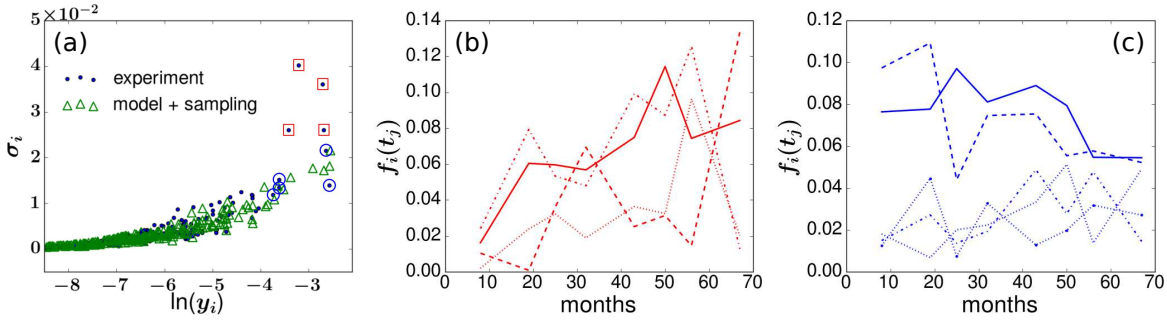

Figure 3.20: Identifying "outlier" clones

Four "outlier" clones (red circles) and five "normal" clones (blue rectangle) are identified from the scatter plots (a). Abundances of the outlier clones (b) and the chosen normal clones (c) in the samples.

According to our model, such long-term clone size expansion implies expansion of their corresponding stem cell populations, which suggests non-neutral dynamics of HSCs. However, since the total number of samples is small, it is difficult to study each individual outlier

clones. Before more sample data become available, our analysis is still performing statistical analysis on as many clones' data as possible. In subsection 3.6.3, we have considered the correlation between consecutive samples for each clone. In order to further extract the long-term trend of clonal dynamics, we further study the inter-sample statistics of clones between non-consecutive samples. If the HSC pool changes considerably over time, such statistics will correlate with the length $t_k - t_j$ of the time interval between any two samples. Here t_j and t_k are not necessarily times of consecutive samples. Figure 3.21(a) shows the means and standard deviations of the distribution $P(t_k|t_j)$ between two observed abundances of the same clone measured at different times t_j and t_k. Note that different pairs of sampling times can have the same gap; for example, both $(t_j = 8,\ t_k = 17)$ and $(t_j = 17,\ t_k = 26)$ correspond to the same interval of 9 months. For similar gaps (e.g. 6 months and 7 months), we group them into the same gap bin. In our dataset for animal RQ5427, $t_k - t_j$ ranges from $6-7$ months to 59 months, depending on t_j and t_k. We also bin clones of similar sizes $f_i(t_j)$ together and study the statistics of the corresponding $f_i(t_k)$ values. The rationale relies on the neutral assumption that clones of the similar sizes behave similarly over a similar amount of time.

Figure 3.21(a) shows that averages of $f_i(t_k)$ calculated within different sample gaps distribute around and are generally close to $f_i(t_j)$. This pattern implies neutral clone dynamics and clone size persistencies. Deviation of $f_i(t_k)$ from $f_i(t_j)$ generally increases as the sample gap increases. A "deterministic" evolution pattern of clones is clearest inside the bin $f_i(t_j) \approx 0.07$: Averages of $f_i(t_k)$ calculated under $24 \sim 25$ months center around $f_i(t_j)$. There is then an obvious rise of $f_i(t_k)$ averages during months $31 \sim 42$, which is followed by an apparent drop at months $31 \sim 59$, suggesting a possible "ballistic" change of HSC clone sizes over time. In the largest bin at $f_i(t_j) \approx 0.11$, $f_i(t_k)$ averages calculated within all sample gaps fall under $f_i(t_j)$. The phenomenon suggests that clone sizes only temporarily reach such large values and would very likely drop to a lower value in later samples. Dynamical pattern in the bin $f_i(t_j) \approx 0.05$ seems random. Clones all fall under $f_i(t_k)$ in the bin $f_i(t_j) \approx 0.03$. In both bins, significant drops are observed at months $48 \sim 59$. For bins at $f_i(t_j) < 0.02$, dynamical patterns tend to be more noisy. Overall, the changes of

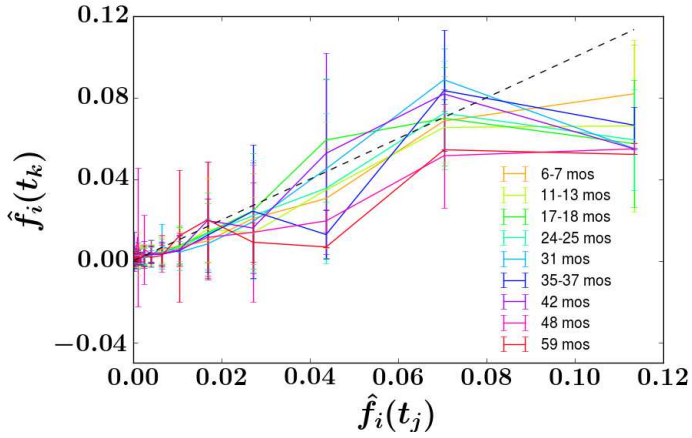

Figure 3.21: Statistics of $f_i(t_k)$ given $f_i(t_j)$ from the experimental data

Different sample gaps $t_k - t_j$ were considered. Because the values of $f_i(t_j)$ are sparse and the relation between $f_i(t_k)$ and $f_i(t_j)$ is stochastic, we discretize $f_i(t_j)$ values into bins and study the distribution of the corresponding $f_i(t_k)$ values. The bin sizes change adaptively with $f_i(t_j)$ values, as there are more $f_i(t_j)$ data points that are close to 0 (small) than close to 0.14 (large). We calculated and plotted the average and standard deviation of $f_i(t_k)$ values in each bin as shown above. The black dashed line denotes $f_i(t_j) = f_i(t_k)$.

HSC sizes between neighboring samples ($6 \sim 11$ month gap) are small and neutral. Possible deterministic mechanisms that induce permanent change (in the HSC pool), if exists, has a characteristic timescale $30 \sim 60$ months. Such deterministic mechanisms, if working on a similar time scale as the random birth and death process, cannot be resolved in the current experimental data set. In Figure 3.22, we plot simulated $m_i(t)$ (solid curves) from a neutral model that includes random birth and death (a) or telomere-length-controlled HSC aging (b) of $h_i(t)$ (dashed curve). Before the ultimate extinction of $h_i(t)$, there is no qualitative difference in the pattern of $m_i(t)$ comparing to that in our baseline model (static h_i). However, such ultimate extinction is not observed from the current dataset.

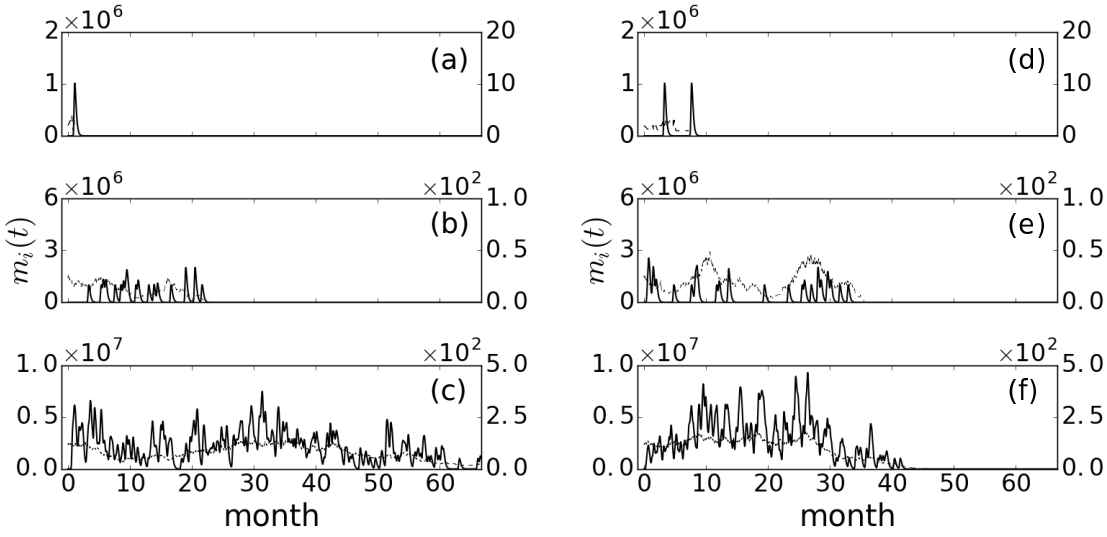

Figure 3.22: Bursty dynamics of $m_i(t)$ under evolving $h_i(t)$

Two types of mechanisms of the stem cell evolution are imposed: h_i (dotted line) changes randomly in time (a) or h_i "ages" (with maximally 60 rounds of self-renewal) after each self-renewal and differentiation (b).

3.6.5 Other datasets

Recently, clonal data from another HSPC tracking experiment (Dunbar's group) on rhesus macaques have become available [KEW17]. Instead of using the integration site information of lentivirus (Chen's group), oligonucleotide barcodes were used to identify various clones. Such barcodes contain 6-base pair library ID followed by a 35-base pair region of highly diverse combinations of base pairs. HSPCs from four macaques were tagged and cell data from various clones were sampled for 10-20 times for up to 49 months. We plot the total numbers of appearing granulocyte clones in each sample *vs.* the months at which the samples were taken in both datasets in Figure 3.23. Though the lengths of tracking are generally shorter in this dataset, sampling is more frequent ($1-8$ months *vs* $6-11$ months), sample sizes are generally larger (about 4×10^6 *vs* 10^3-10^4), and more clones are tracked ($3000-5000$ *vs* $500-1200$).

We check the average-standard deviation pattern of the dataset in Figure 3.24. The

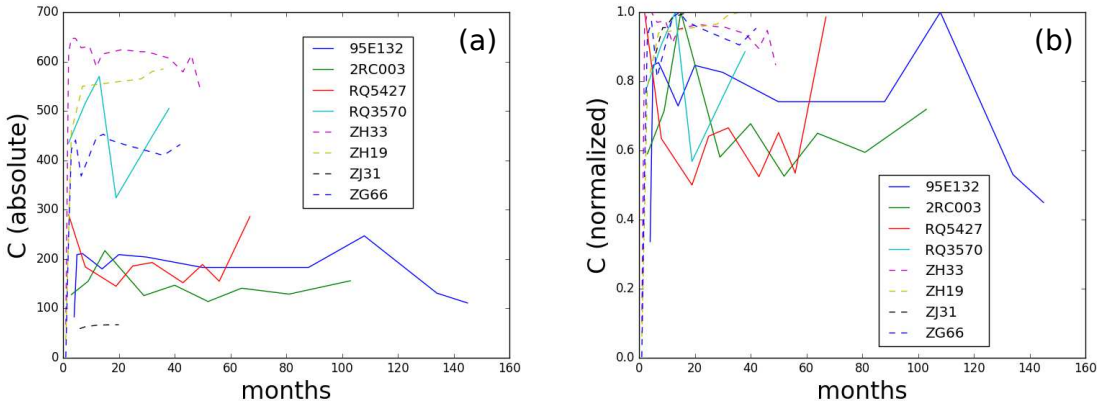

Figure 3.23: Numbers of detected clones in 8 rhesus macaques

sampled clones (red dots) are distributed in a range of smaller average abundances (x-axis) than those from Chen's group (blue dots), probably because more clones are tracked (so the average fraction of each clone was "suppressed" by normalization). Also, sampling-alone-induced fluctuations are much smaller (green triangles) , a direct outcome of the much larger sample sizes. There is still a gap between fluctuations induced by sampling alone and those observed from the experimental data.

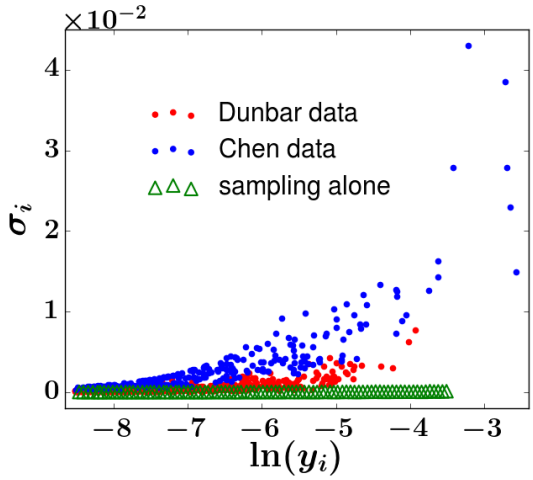

Figure 3.24: Averages and standard deviations of clonal abundances in animal ZH33. Data of ZH33 from the Dunbar group (red dots), RQ5427 from Chen's group (blue dots), and simulations from the null hypothesis model of "sampling alone" (green triangles) based on sampling parameters from Dunbar's group.

Another feature of the dataset is that cells of various types or lineages (e.g. granulocytes, monocytes, T cells) were separately sequenced, which allows an additional aspect of hematopoiesis to be investigated. It was observed [KEW17] that the granulocyte (Gran) lineage shows a high correlation with the Monocyte (Mono) lineage, validating the common knowledge that both belong to the myeloid lineage. Monos are also ideal objects to study since they also have relative shorter lifespan in circulation and do not proliferate in periphery, thus capable of indicating recent activities of hematopoietic stem cells. Monocytes have three subgroups by phenotype: classical, non-classical, and intermediate, where the classical group accounts for 80-90% of peripheral blood monocytes [BMH17]. Using *in vivo* human deuterium labeling, it was shown that the average waiting time to release post-mitotic monocytes from bone marrow is 1.6 days and that of circulation of about 1 day [PZF17]. In peripheral blood, Grans are around $15 - 20$ folds more than Monos [CQD09].

The positive correlation between Grans and Monos was quantitatively modeled by employing simple multi-compartment birth-death-migration process [XKG16], which however only focus on the scale-free correlation measure between different lineages, which is easier to fit. However, the observed averages and covariances, which are scale-dependent quantities, require a more detailed model to account for. Our progenitor aging model has been shown to capture the scale-dependent averages and variances quite well. We can push further our results by modeling the covariances and with the inclusion of the Monos data.

In our previous model, each burst of progenitor cells is constrained to the granulocyte lineage. However, since blood cells of various types share common ancestor at some stage of the progenitor differentiation (at least, they share the same HSC), different lineages have to "branch out" at some point of the progenitor proliferation. Thus our granulocyte burst ignores "byproduct cells" of other types, whose effect can be subsumed into progenitor cell death. To include Monos, it is important to find the "branching point" of the myeloid lineage after which progenitor cells commit to either Grans or Monos. Such tree-like structure is naturally formulated by our progenitor aging model, which will be shown in subsection 3.6.6.

Two extremal cases are shown by Figure 3.25, where the branching point is set at $\ell_{\mathrm{b}} = 0$ (a) and $\ell_{\mathrm{b}} = L$ (b). In the first case, a generation-0 progenitor cell immediately needs

to decide which lineage it would go, i.e. Grans or Monos. The subsequent expansion of progenitor cells would only include either one of the lineages. We expect negative correlation between the cell counts of the two lineages. In the second case, a progenitor cell would not make a decision until the last generation L, so all cells previous generations are common progenitors for both lineages. We would instead expect positive correlation.

In real data, we expect such pattern to be clear only among small clones. For large clones, on the other hand, the correlation should always be close to 1 since the total number of HSC differentiation is so high that cell bursts overlap. Data from small clones, however, would be severely corrupted by sampling errors. We plot the experimental data in Figure 3.26. Observation confirms that large clones tend to have positive correlation between their Gran and Mono outputs. As the average abundance drops, correlations start to spread out among $(0, 1)$. One can expect sampling effect to play a role in corrupting the pattern. The challenge is again to extract model-induced correlation information from sampling noise.

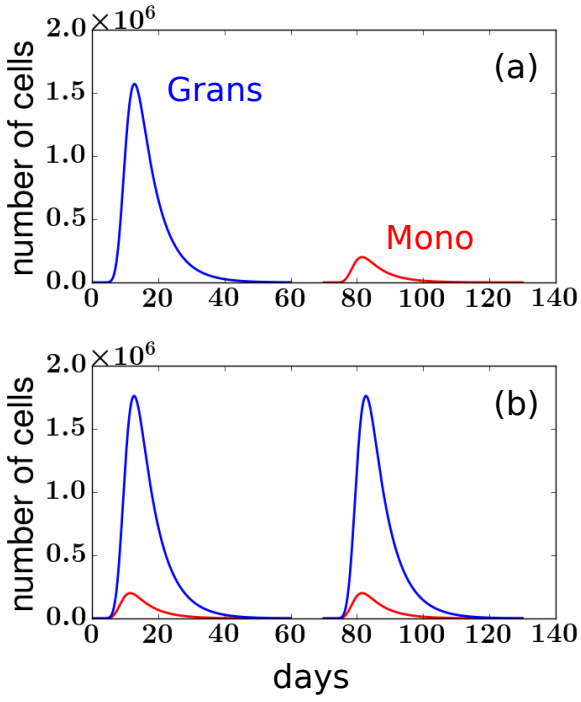

Figure 3.25: Modeling correlations between Gran and Mono bursts

Two patterns of bursty dynamics of granulocytes and monocytes.

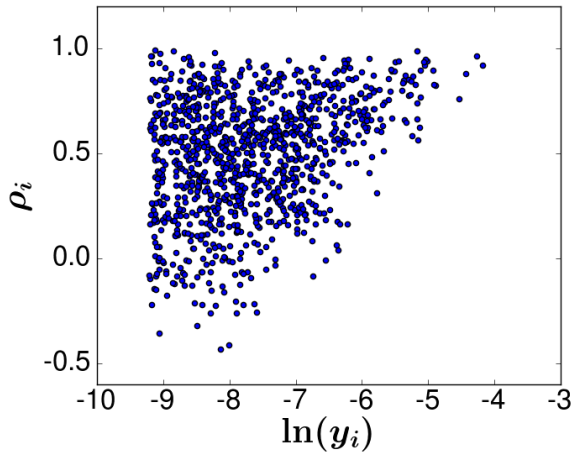

Figure 3.26: Correlations of granulocytes and monocytes in animal ZH33

Scatter plot of granulocyte clones in the average-correlation space for rhesus macaque ZH33 from Dunbar's group

3.6.6 Lineage tree of myeloid progenitor cells

An earlier statistical work proposed several simple "lineage tree" structures that specify how cells of various lineages are derived from a stem cell [XKG16]. However, progenitor "aging" was not involved, but cells from different compartments undergo random homogeneous birth-death-migration processes. Given an initial stem/progenitor cell, all progenies' averages, variances and covariances were calculated through the probability generating function (PGF) method. In particular, the scale-free correlations were extracted to compare experiment and simulations. The problem was formulated by the probability generating function of the system at any state $\{n_\ell\}$ ($0 \leq \ell \leq L$, where $L+1$ is the total number of cell types), starting from a type-i cell,

$$F_i(\mathbf{z};t) = \sum_{n_0=0}^{\infty} ... \sum_{n_L=0}^{\infty} P_i(\mathbf{n};t) z_1^{n_1} ... z_L^{n_L}. \tag{3.52}$$

Denote A_i as the generating function for the dynamics of the i^{th} type cell. For example, in the modeled plotted in Figure 3.27 which contains $L = 5$ types of cells, one obtains

$$A_0[\mathbf{z}] = \mu_0 + \nu_{01} z_1 + \lambda_0 z_0^2,$$

$$A_1[\mathbf{z}] = \mu_1 + \nu_{12} z_2 + \nu_{13} z_3 + \lambda_1 z_1^2, \qquad A_2[\mathbf{z}] = \mu_2 + \nu_{24} z_4 + \lambda_2 z_2^2,$$

$$A_3[\mathbf{z}] = \mu_3 + \nu_{35} z_5 + \lambda_3 z_3^2, \qquad A_4[\mathbf{z}] = \mu_4, \qquad A_5[\mathbf{z}] = \mu_5.$$

Another example is plotted in Figure 3.28, where lineage branching is naturally integrated

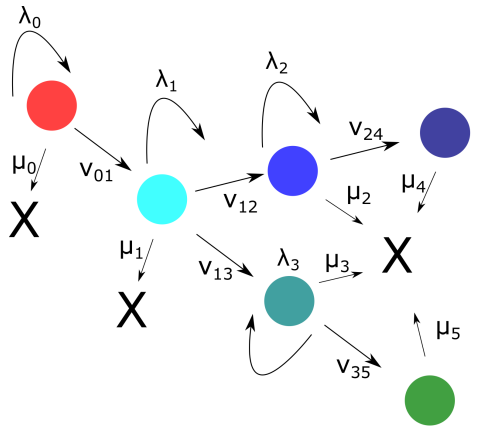

Figure 3.27: Branching model for simple Gran and Mono dynamics

Red ball ($i = 0$) is hematopoietic stem cell (HSC). Light blue ball ($i = 1$) is multi-potent progenitor (MPP) for Gran and Mono. Mid blue ball ($i = 2$) is uni-potent progenitor (UPP) for Gran. Dark blue ball ($i = 4$) is differentiated Gran. Mid green ball ($i = 3$) is UPP for Mono. Dark green ball ($i = 5$) is differentiated Mono.

into a generation-limited progenitor cell model. There are L_0 generations of myeloid progenitors, L_1 generations of Gran progenitors, and L_2 generations of Mono progenitors. One obtains

$$A_1[\mathbf{z}] = \mu_0 + \lambda_0 z_1^2, \qquad ..., \qquad A_{L_0-1}[\mathbf{z}] = \mu_0 + \lambda_{L_0-1} z_L^2,$$

$$A_{L_0}[\mathbf{z}] = \mu_0 + \nu_{01} z_{L_0+1} + \nu_{02} z_{L_0+L_1+1},$$

$$A_{L_0+1}[\mathbf{z}] = \mu_1 + \lambda_1 z_{L_0+2}^2, \qquad ..., \qquad A_{L_0+L_1}[\mathbf{z}] = \mu_1 + \omega_1,$$

$$A_{L_0+L_1+1}[\mathbf{z}] = \mu_2 + \lambda_2 z_{L_0+L_1+2}^2, \qquad ..., \qquad A_{L_0+L_1+L_2}[\mathbf{z}] = \mu_2 + \omega_2.$$

In this model, the stochastic self-renewal of progenitors is replaced by a limited number of

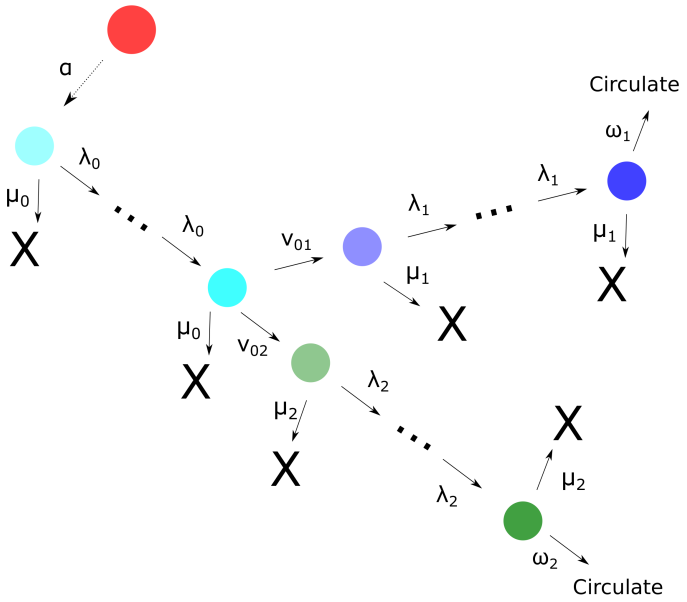

Figure 3.28: Branching model combined with progenitor cell aging

Red ball is HSC. Light blue ball are common progenitors for Grans and Monos ($\ell = 1, 2, 3, ..., L_0$). The blue ball branch is the Gran lineage ($\ell = L_0 + 1, L_0 + 2, L_0 + 3, ..., L_0 + L_1$). The green ball branch is the Mono lineage ($\ell = L_0 + L_1 + 1, L_0 + L_1 + 2, L_0 + L_1 + 3, ..., L_0 + L_1 + L_2$).

cell divisions. More "bursty" dynamics are expected in this latter model.

To solve the time evolution of Eq. (3.52), further define $u_i(\mathbf{z}) = A_i[\mathbf{z}] - A_i[\mathbf{1}]z^i$, which describes the dynamics of how all other types of cells affect the i^{th} type. One obtains [XKG16]

$$\frac{\partial F_i(\mathbf{z};t)}{\partial t} = u_i(F_1(\mathbf{z};t), ..., F_L(\mathbf{z};t)). \tag{3.53}$$

Based on this formula, and the definition of $F_i(\mathbf{z};t)$, one obtains the average number of type k and covariance of type k, ℓ cells, given one type-0 cell at time 0, as

$$M_{k|0} = \frac{\partial}{\partial z_k} F_0(\mathbf{z};t)\bigg|_{\mathbf{z}=\mathbf{1}}, \qquad U_{k\ell|0} = \frac{\partial^2 F_0}{\partial z_k \partial z_\ell}\bigg|_{\mathbf{z}=\mathbf{1}}. \tag{3.54}$$

Evolution equations for the average of and the covariance between various lineage cell num-

bers can then be calculated:

$$\frac{\partial M_{j|i}}{\partial t} = \frac{\partial F_i}{\partial t \partial z_j}\bigg|_{\mathbf{z}=1} = \sum_k \frac{\partial u_i}{\partial z_k}\frac{\partial F_k}{\partial z_j}\bigg|_{\mathbf{z}=\mathbf{1}} = \sum_k \frac{\partial u_i}{\partial z_k} M_{j|k}, \tag{3.55}$$

$$\frac{\partial U_{jk|i}}{\partial t} = \frac{\partial^3 F_i}{\partial t \partial z_j \partial z_k}\bigg|_{\mathbf{z}=1} = \sum_{m=1} \left(\frac{\partial u_i}{\partial F_m}\frac{\partial^2 F_m}{\partial z_j \partial z_k} \right) + \sum_{m,n=1} \left(\frac{\partial^2 u_i}{\partial F_m \partial F_n}\frac{\partial F_m}{\partial z_j}\frac{\partial F_n}{\partial z_k} \right) \bigg|_{\mathbf{z}=\mathbf{1}}. \tag{3.56}$$

CHAPTER 4

Density-dependence-induced phase transition in clone abundances

4.1 Introduction

During our study of the multi-compartment clonal dynamics in Chapter 3, we studied the multi-clonal dynamics in the stem cell pool under a birth-death process with density- dependency (or carrying capacity). This model is a typical high-dimensional model if the size of each clone constitutes a dimension. High-dimensional stochastic models are important across many fields of science and arise often in biological and medical contexts, including immunology [ZES13], ecology [MEG07], cellular barcoding experiments [GKC15], etc. For example, applications of DNA-tagging technology allow *in vivo* tracking of multiple hematopoietic clones, each of which was derived from a unique hematopoietic stem cell that carries a unique DNA tag [ZES13, KKP14, SRC14, BPS16, KEW17], generating clonal-tracking data of very high dimensions. Here the number of cells n_i carrying the i^{th} label (or the size of the i^{th} clone) represents the i^{th} dimension. Multispecies ecological communities are another example of high-dimensional systems. For example, if the compartment of interest is an island, then n_i quantifies the number of animals of species i on the island. T cells in the jawed vertebrates can also classified into multiple subpopulations, each corresponding to different T cell receptor subtypes and produced by the thymus. Here, n_i denotes the number of T cells in an organism that carry the i^{th} receptor. In this setting the huge number of different T cell receptors ($1 \leq i \leq \Omega$, $\Omega \sim 10^6 - 10^8$) present in an organism allows its adaptive immune system to recognize and respond to a wide range of unknown antigens that it might encounter.

The simplest single-compartment mathematical structure that is common to all the multispecies systems mentioned above is the birth-death-immigration (BDI) processes shown in Figure 4.1 The source of immigration into the system is a fixed population of H different

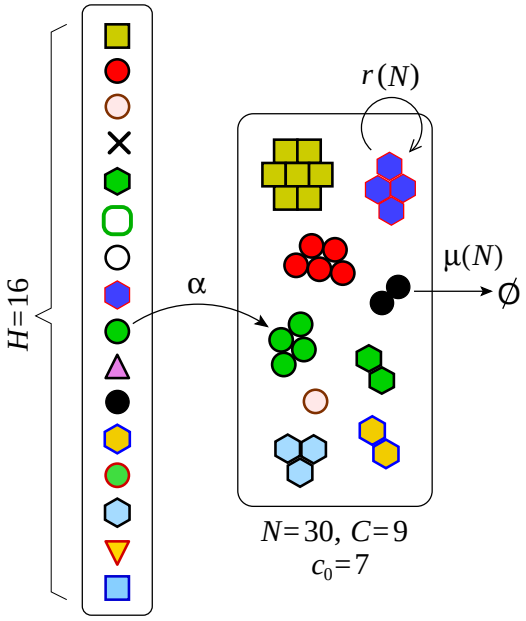

Figure 4.1: Birth-death-immigration model

A simple H-species birth-death-immigration process in which an "external mainland" always contains H individuals, each of a unique species. Each species immigrates into the system with rate α but is immediately replaced on the mainland. All individuals in the system can proliferate with rate $r(N)$ and dies with rate $\mu(N)$, where N is the total population in the system. A specific configuration with $H = 16$ and $N = 30$ is depicted. Here, $C = 9$ represents the *number of different clones* that exist in the system.

individuals. After immigration into the system, the individuals can proliferate with rate $r(N)$ and die with rate $\mu(N)$, both possibly function of the total population N. In the configuration shown in Figure 4.1, the number of individuals of each species is $n_1 = 7, n_2 = 5, n_3 = n_4 = 4, n_5 = 3, n_6 = n_7 = n_8 = 2, n_9 = 1$, where we have defined clones i according to decreasing population. In this work, we will interchangeably use the terms "clones" and "species" that are distinguishable by "labels", whether they distinguish different DNA tags, T-cell receptor types, or species identities.

Such high-dimensional stochastic systems are generally difficult to study because of the "curse of dimensionality." A description using the full probability distribution $P(\mathbf{n}) \equiv P(\{n_1, n_2, ..., n_H\})$, where H denotes the total number of labels, is unintuitive and typically computationally intractable. It also contains more information than necessary if we consider only neutral clones and clone identity is not relevant.

Describing the system in terms of moments such as $\langle n_i \rangle$ and $\langle n_i n_j \rangle$ reduces the model complexity and allows one to track the dynamics of specific clones [XKG16], but does not directly capture the clone size distribution resulting from the relevant stochastic processes. Another approach is to use single-quantity metrics such as species richness $R \equiv \sum_{k=1}^{\infty} \sum_{i=1}^{H} \mathbb{1}(n_i, k)$, Simpson's diversity, Shannon's diversity, or the Gini index to describe and compare various ecological communities. Here $\mathbb{1}(x, y)$ is the identity function which takes value 1 when $x = y$ or 0 otherwise. Such diversity measures can be overly simplistic and can lead to different conclusions depending on the diversity index used. Thus, a description of intermediate complexity is desired.

In ecology, a commonly used measure is the species abundance distribution (SAD) that counts the number of different species encountered in a community [MEG07]. In the language of clonal dynamics, it is the count of the number of clones that are each represented by k individuals as depicted in Figure 4.2 [GKC15]:

$$c_k = \sum_{i=1}^{H} \mathbb{1}(n_i, k), \tag{4.1}$$

The clone or species count c_k is a one-dimensional vector of numbers indexed by $k = 0, 1, ...$ and gives a more comprehensive picture of how the clone/species are distributed compared to that of a single index. The clone count c_k is also constrained by construction: By definition we also have the normalization and

$$\sum_{k=0}^{\infty} c_k = H, \qquad \sum_{k=0}^{\infty} k c_k = N. \tag{4.2}$$

Clone counts can be useful in describing numbers of rare or abundant clones especially when clone identity is not important. Clone- or species-counts are widely applied to characterize systems of various scales, including gene-barcoded, virally tagged, or TCR-decorated [JHH13,

ZES13,QLC14,GKC15,DMW16] cellular clones, microbial populations [HWH03,HBJ06], and ecological species [Hub01, MEG07, GT05].

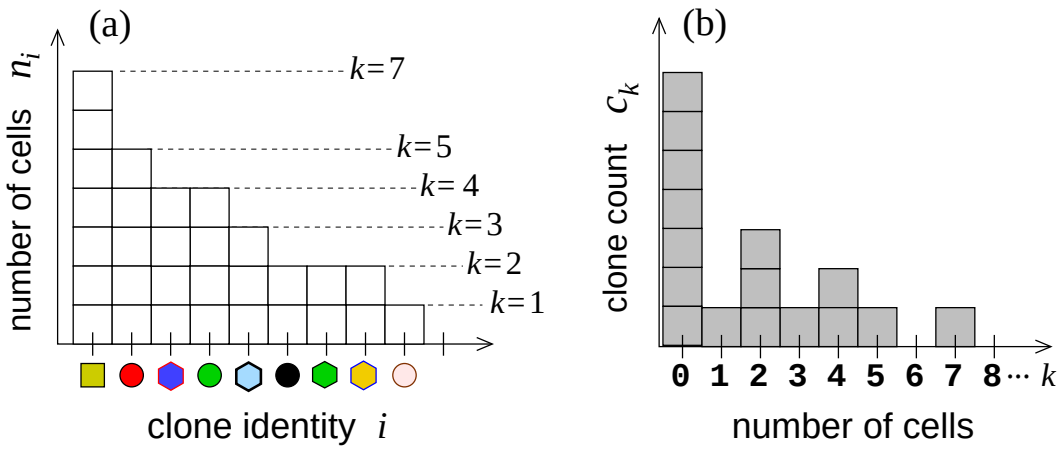

Figure 4.2: Clone count statistics

Definition of clone counts corresponding to the configuration in Figure 4.1. (a) In the cell-count representation n_i is the number of cells of clone i detected in a sample. (b) c_k is the number of different clones that are represented by exactly k cells in a sample. A given set $\{n_i\}$ uniquely determines the corresponding $c_k \equiv \sum_{i=1}^{\infty} \mathbb{1}(n_i, k)$. However, one cannot uniquely determine n_i from c_k since clone identity information is lost when transforming from n_i to c_k.

A universally observed feature in empirical studies across all these fields is a "hollow curve distribution" for c_k, where few highly populous species and many low-population species arise [MEG07]. Theoretical studies attempt to explain these observations by proposing various physical models, including neutral models with constant immigration, birth, and death rates [Mot32,FCW43,Ken48a], time-dependent birth and death [Ken48b], cell-wise and clone-wise heterogeneities [DMW16], and intra-species carrying capacities [VBH05]. Many of these studies neglect interactions or apply a mean-field approximation.

Although some theories considered competition for resources (such as T cell proliferation competing for stimuli from self-peptides [LCH16, DMW16, EGC16]), resources themselves were modeled as evolving variables of the system and explicit solutions exist only in very simple cases. Other studies have considered intra-species carrying capacities that limit the size of an individual clone. However, populations in each clone are assumed to evolve in-

dependently of those in other clones. None of these previous studies have treated global interactions that correlate population dynamics across all clones/species, a commonly considered factor in population dynamics [PQP08].

In this paper, we introduce a simple idea of transforming the problem of calculating the q^{th} moment of $\{c_k\}$ to the problem of solving a $(q+1)$-dimensional process described by the population vector $\{n_1,\ ,n_2,\ ...,\ n_q,\ N'\}$. This process can be further approximated by a q-dimensional Moran model that imposes a fixed total population. We use this approach to accurately calculate the 1^{st} and 2^{nd} moments of c_k under functional forms of the carrying capacity. We then exploit ideas from energy landscapes to identify the key parameters that control a phase transition of the general multi-species BDI process and explain the failure of previously used mean-field assumptions.

4.2 Classical formulation and mean-field assumption

Here, we develop dynamical model for the stochastic BDI process depicted in Figure 4.1. In the language of clonally tracked stem cell differentiation, the probability of an stem cell that carries any specific tag asymmetrically differentiating to produce a progenitor cell within infinitesimal time $\mathrm{d}t$ is $\alpha \mathrm{d}t + o(\mathrm{d}t)$. We will assume that there is a fixed number H of "source" or "founder" individuals The probability for any progenitor cell to divide into two new identical progenitor cells (birth) within $\mathrm{d}t$ is $r\mathrm{d}t + o(\mathrm{d}t)$ and the probability of it dying in $\mathrm{d}t$ is $\mu \mathrm{d}t + o(\mathrm{d}t)$.

We will further assume the particle dynamics are coupled in a clone-independent way, leading to identical (but not independent) statistics of the populations of each clone. The canonical implementation of such a "global" interaction is through a birth rate $r(N)$ and/or death rate $\mu(N)$ that depend only on the total population $N \equiv \sum_{i=1}^{H} n_i$. Thus, the total population N can be "decoupled" and completely described by its own master equation,

$$\begin{aligned}\frac{\partial P(N,t)}{\partial t} = \alpha[P(N-1) - P(N)] + r(N)\left[(N-1)P(N-1) - NP(N)\right] \\ +\mu(N)\left[(N+1)P(N+1) - NP(N)\right],\end{aligned} \tag{4.3}$$

from which moments of N can be computed. The higher-dimensional master equation obeyed

by the full multispecies distribution $P(n_1, n_2, \ldots, n_H; t)$ is explicitly given in Subsection 4.7.1.

Let us denote the ensemble (not time) average of a quantity by $\langle\cdot\rangle$. Thus, $\langle n_i(t)\rangle \equiv \sum_{n_j} n_i P(\{n_j\}, t)$ represents the expected sizes of the i^{th} clone. By using Eqs. (4.3) and (4.35, we can show that the expected subpopulation $\langle n_i(t)\rangle$ and total population $\langle N(t)\rangle$ for the BDI process obeys

$$\begin{aligned}\frac{\mathrm{d}\langle n_i\rangle}{\mathrm{d}t} &= \alpha + \langle (r(N) - \mu(N))n_i\rangle,\\ \frac{\mathrm{d}\langle N\rangle}{\mathrm{d}t} &= \alpha H + \langle (r(N) - \mu(N))N\rangle.\end{aligned} \tag{4.4}$$

In Subsection 4.7.2, we also explicitly derive the equation for $\langle c_k(t)\rangle$ from the master equation for $P(c_0, c_1, c_2, \ldots; t)$:

$$\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} = \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + \langle r(N)\left[(k-1)c_{k-1} - kc_k\right]\rangle + \langle \mu(N)\left[(k+1)c_{k+1} - kc_k\right]\rangle. \tag{4.5}$$

This evolution equation indicates that immigration (at rate α) of a cell within a clone of size $n_i = k$ increases its size by 1, thereby decreasing c_k by 1 but increasing the number of clones of size $k+1$, c_{k+1}, by 1. Cellular birth and death have similar effects, but their corresponding rates are proportional to the clone size k (the number of cells in the clone). In the rest of this paper we are interested in evaluating the steady-state values of $\langle c_k\rangle$.

4.2.1 Constant rates

In the simplest scenario of constant birth and death rates, one can write [GKC15]

$$\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} = \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + r[(k-1)\langle c_{k-1}\rangle - k\langle c_k\rangle] + \mu[(k+1)\langle c_{k+1}\rangle - k\langle c_k\rangle]. \tag{4.6}$$

If $r < \mu$, a stable steady state can be found:

$$\langle c^*_{k\geq 1}\rangle = \frac{\alpha H}{rk!}\frac{(\frac{r}{\mu})^k(1-\frac{r}{\mu})^{\alpha/r}}{\frac{\alpha}{r}+k}\prod_{\ell=1}^{k}\left(\frac{\alpha}{r}+\ell\right), \quad \langle c^*_0\rangle = H - \sum_{k=1}^{\infty}\langle c^*_k\rangle = H\left(1-\frac{r}{\mu}\right)^{\alpha/r}. \tag{4.7}$$

4.2.2 Carrying capacity and mean-field approximation

Now, assume the birth and death rates depend on N such that $r(N)$ decreases with N and/or $\mu(N)$ increases with N. These forms for $r(N)$ and $\mu(N)$ guarantee that n_i and c_k are

bounded even if $r(0) > \mu(0)$. Terms of the form $\langle r(N)c_k\rangle \neq \langle r(N)\rangle\langle c_k\rangle$ in Eq. (4.5) cannot be factored because $r(N)$ depends on c_k through the stochastic variable $N \equiv \sum_\ell \ell c_\ell$ (Eq. 4.2). Nonetheless, to make headway, a mean-field method is often invoked to simplify Eq. (4.4) and Eq. (4.5). Upon fully factorizing interaction terms such as $\langle r(N)c_k\rangle \approx r(\langle N\rangle)\langle c_k\rangle$ and $\langle r(N)N\rangle \approx r(\langle N\rangle)\langle N\rangle$, we can approximate Eqs. (4.4) and (4.5) as

$$\frac{\mathrm{d}\langle N\rangle}{\mathrm{d}t} \approx \alpha H + \langle (r(\langle N\rangle) - \mu(\langle N\rangle))\langle N\rangle \equiv f(\langle N\rangle), \tag{4.8}$$

$$\begin{aligned}\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} \approx{}& \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + r(\langle N\rangle)\left[(k-1)\langle c_{k-1}\rangle - k\langle c_k\rangle\right] \\ &+ \mu(\langle N\rangle)\left[(k+1)\langle c_{k+1}\rangle - k\langle c_k\rangle\right].\end{aligned} \tag{4.9}$$

By first solving Eq. (4.8) we can input $\langle N(t)\rangle$ into Eq. (4.9) and explicitly solve for $\langle c_k(t)\rangle$. The steady state of $\langle c_k\rangle$ can be reached only after the steady state of $\langle N\rangle$ is reached and $r(\langle N\rangle)$ and $\mu(\langle N\rangle)$ approach constant values. The steady-state solution of Eq. 4.8 is defined by $f(\langle N^*\rangle) = 0$. The requirement that $\langle N^*\rangle > 0$ is equivalent to $\frac{\mathrm{d}f}{\mathrm{d}\langle N\rangle}\Big|_{\langle N^*\rangle} = \left(\frac{\mathrm{d}r}{\mathrm{d}\langle N\rangle} - \frac{\mathrm{d}\mu}{\mathrm{d}\langle N\rangle}\right)\Big|_{\langle N^*\rangle} < 0$.

We show in Subsection 4.7.3 that this deterministic description breaks down after an exponentially long time as the immigration rate α is sufficiently small. The reason is that $N = 0$ becomes effectively an absorbing boundary in the full stochastic model. The $\langle N^*\rangle$ we find from $f(\langle N^*\rangle) = 0$ is actually a quasi-steady state (QSS) even though Eq. (4.8) yields a stable deterministic equilibrium $\langle N^*\rangle$ for physically reasonable functions $r(\langle N\rangle)$ and $\mu(\langle N\rangle)$.

Focusing on evaluating the QSS value of $\langle c_k^*\rangle$ before the final extinction that occurs over exponentially long times, we denote $r(\langle N^*\rangle) \equiv r^*$ and $\mu(\langle N^*\rangle) \equiv \mu^*$ as the rates of birth and death at QSS. The QSS solution $\langle c_k^*\rangle$ is similar in form to that in Eq. (4.7),

$$\langle c_{k\geq 1}^*\rangle = \frac{\alpha H}{r^* k!}\frac{(\frac{r^*}{\mu^*})^k(1-\frac{r^*}{\mu^*})^{\alpha/r^*}}{\frac{\alpha}{r^*}+k}\prod_{\ell=1}^{k}\left(\frac{\alpha}{r^*}+\ell\right), \quad \langle c_0^*\rangle = H - \sum_{k=1}^{\infty}\langle c_k^*\rangle = H\left(1-\frac{r^*}{\mu^*}\right)^{\alpha/r^*}. \tag{4.10}$$

Here, $\langle c_k^*\rangle$ above corresponds to the expected QSS clone-count under the mean-field approximation which we expect to be different from the exact solution.

4.2.3 Failure of the mean-field approximation

To explicitly investigate the errors incurred under a mean-field assumption, we will be use a concrete Logistic growth law for the total population defined by

$$r(N) = p\left(1 - \frac{N}{K}\right), \qquad \mu(N) = \mu, \tag{4.11}$$

where p is the maximal birth rate and K is the carrying capacity parameter. The QSS solution for the total population is

$$\langle N^* \rangle = \frac{K}{2}\left(1 - \frac{\mu}{p}\right)\left[1 + \sqrt{1 + \frac{4\alpha H p}{(p-\mu)^2 K}}\right]. \tag{4.12}$$

In many examples, such as progenitor cell dynamics, $K \gg 1$. To ensure large populations $\langle N^* \rangle$, p, μ, and $p - \mu$ must be comparable in magnitude. For example, we can have $p = 2,\ \mu = 1$ but not $p = 2, \mu = 1.999999$. We will focus on the parameter range $\alpha H \ll pK$, so the population is mainly supported by birth rather than immigration.

We are now in a position to use $\langle N^* \rangle$ to determine r^* and evaluate the mean-field approximation for $\langle c_k^* \rangle$ (Eq. (4.10)). In Figure 4.3 we compare numerically evaluated mean-field solution $\langle c_k^* \rangle$ with Monte-Carlo simulations of the underlying BDI process for various values of α.

Clearly, for small $\alpha = 10^{-8}$ as in Figure 4.3(a), Eq. (4.10) fails to capture the peak arising in $\langle c_k^* \rangle$ near $k = \langle N^* \rangle$. In the singular limit $\alpha \to 0$, the mean-feield solution $\langle c_{k\geq 1}^* \rangle \to 0$ and $\langle c_0^* \rangle \to H$ but nonetheless by construction, satisfies $\sum_{k=1}^{\infty} k \langle c_k^* \rangle \to \langle N^* \rangle$.

However, in the exact (simulated) $\langle c_k^* \rangle$ the small peak at large size $k \approx \langle N^* \rangle$ signals that a single large clone has come to dominate the total population. The number of clones not in the system is thus $\langle c_0^* \rangle \approx H - 1$. One clone, typically the first to have immigrated, has taken over the system squeezing out all others that try to immigrate when the immigration rate α is small. At higher immigration rates, the structure of the simulated $\langle c_k^* \rangle$ is non-monotonic with respect to α. When α is still relatively small as in (b), the simulated $\langle c_k^* \rangle$ is dominated by small clones but with a slow decay with size k. As immigration increases, larger clones do not have the opportunity to establish and more intermediate-sized clones arise at the expense

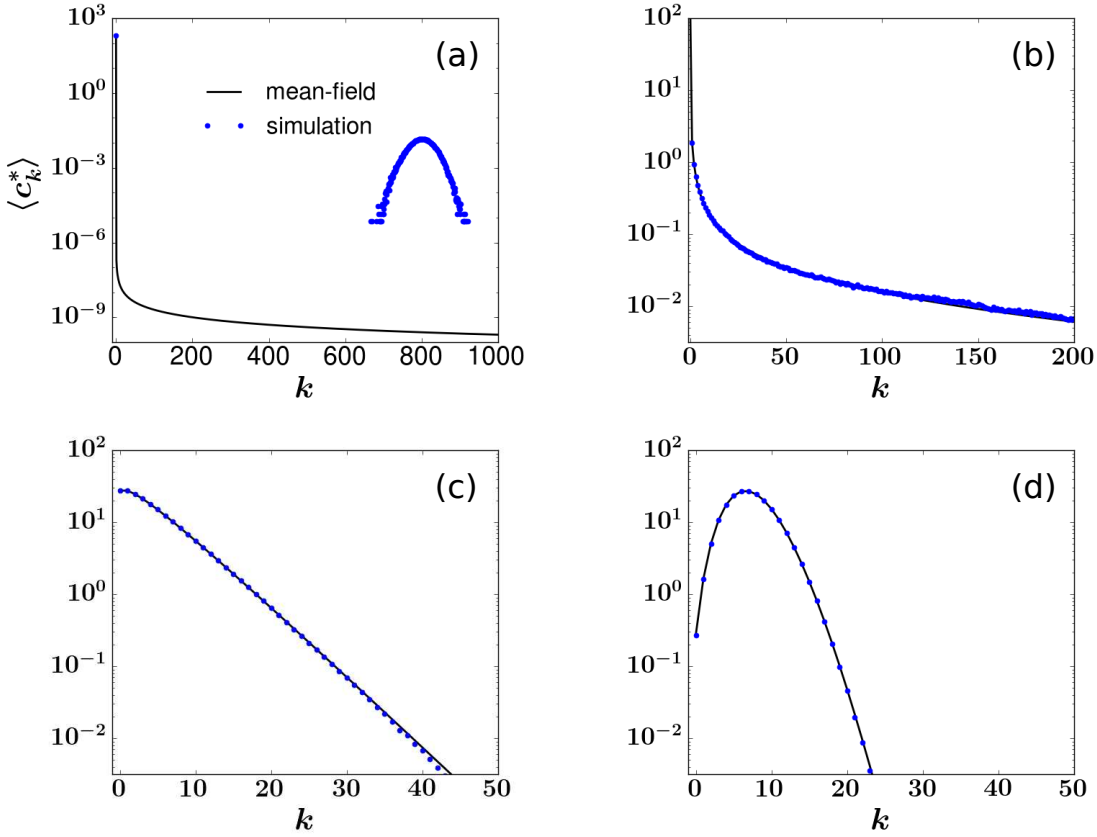

Figure 4.3: Simulation and mean-field results under Logistic birth

Comparison of steady-state clone-count distributions from simulations (black dots) to the mean-field model (dashed curves) $\bar{c}_k^*$ using the logistic growth model of Eq. (4.11) and (a) $\alpha = 10^{-8}$, (b) $\alpha = 0.1$, (c) $\alpha = 10$, and (d) $\alpha = 60$. Other parameters used are $\mu = 10,\ p = 20,\ K = 1600,\ H = 200$. The mean-field approximation $\langle c_k^* \rangle$ breaks down for small α in (a). Also, note the log scale and the absence of simulations that capture the rare configurations where $k \neq \langle N^* \rangle$ (a).

of very large clones but the simulated result $\langle c_k^* \rangle$ remains monotonic. For even larger α, a peak at size $k \ll \langle N^* \rangle$ develops and even fewer large clones arise. All these larger-α cases shown in Figure 4.3(b-d) are accurately described by the mean-field approximation for $\langle c_k^* \rangle$.

4.3 Proposed Model for $\langle \mathbf{c}^q \rangle$

The challenge in solving Eq. (4.5) lies in the non-separable terms $\langle r(N)c_k \rangle$. Even in the simple case of logistic growth where $r(N)$ is linear, the $\langle r(N)c_k \rangle$ terms include second-moments $\langle c_k c_l \rangle$, which usually cannot be approximated by $\langle c_k \rangle \langle c_l \rangle$. If one attempts to solve Eq. (4.5) for the time-dependent or steady-state solution $\langle c_k^* \rangle$, one encounters the so-called "moment closure" problem, where the solution of the 1st moment $\langle c_k \rangle$ depends on 2nd moments $\langle c_k c_\ell \rangle$, which in turn depends on 3rd moments, and so on. There is usually no closed-form solution or easy approximation to such problems. In the rest of this section, we develop an alternative approach.

4.3.1 Transformation of the problem

We recall the definition of the clone count c_k and exploit the cell-count representation. By using the definition in Eq. (4.1), one can easily show that (Subsection 4.7.4)

$$\langle c_k(t) \rangle = HP(n_1 = k; t), \tag{4.13}$$

which is exact if the initial populations of all clones are identical $n_1(0) = n_2(0) = \ldots = n_H(0)$. This assumption does not affect the generality of our later conclusions at QSS as long as different initial distributions $c_k(0)$ converge to a unique $\langle c_k^* \rangle$. Thus, at QSS $\langle c_k^* \rangle = HP(n_1 = k; t \to \infty)$ always holds. Intuitively, the expected fraction of all clones that have size k is the probability that any one clone is of size k.

Eq. (4.13) illustrates the indistinguishability among clones that will be quite useful. We can heuristically write the master equation for the BDI process of a single clone with population n_1 as

$$\begin{aligned} \frac{\mathrm{d}P(n_1)}{\mathrm{d}t} = \alpha[P(n_1 - 1) - P(n_1)] + r(N)\left[(n_1 - 1)P(n_1 - 1) - n_1 P(n_1)\right] \\ +\mu(N)\left[(n_1 + 1)P(n_1 + 1) - n_1 P(n_1)\right] \end{aligned} \tag{4.14}$$

where $N(t)$ represents one trajectory of the random process $N = n_1 + n_2 + \ldots + n_H$ which we might approximate using the deterministic solution to Eq. (4.4). Equation (4.14) has the exact same form as the right-hand side of Eq. (4.9) for $\langle c_k \rangle$. However, in the presence of other

species or clones, it is immediately clear that Eq. (4.14) is not a complete description for n_1 since the variable N depends on the population of the 1st clone. Clonal "independence" breaks down through the $r(N)$ and $\mu(N)$ terms. All clones compete with each other for the limited sources in the environment through regulating their shared birth rate. The full model should be H-dimensional.

Eq. (4.13) is still exact (Subsection 4.7.4) since the clonal dynamics are neutral and all clones start with the same initial size. One still needs to solve any *individual* clone's marginal probability distribution $P(n_1)$ given that all other clones can affect it. Formally, this corresponds to first solving the full distribution $P(\{n_1,\ n_2,\ ...,\ n_H\})$ before summing over all other populations $\{n_2,\ ...,\ n_H\}$. Further consideration indicates that we do not care about the detailed configuration of $\{n_2,\ ...,\ n_H\}$, but rather their combined effects on n_1. Therefore, we can lump species 2 through H into an effective "bath" clone whose size is $N' = n_2 + ... + n_H$. This effective clone has birth rate $r(N) \equiv r(n_1 + N')$, death rate $\mu(n_1 + N')$, and immigration rate $(H-1)\alpha$. Eq. (4.14) is now coupled to a master equation

$$\begin{aligned}\frac{\mathrm{d}P(N')}{\mathrm{d}t} =& \alpha(H-1)[P(N'-1) - P(N')] + r(N)\left[(N'-1)P(N'-1) - N'P(N')\right] \\ &+ \mu(N)\left[(N'+1)P(N'+1) - N'P(N')\right]. \end{aligned} \tag{4.15}$$

One usually combines Eqs. (4.14) and (4.15) together into a 2D master equation

$$\begin{aligned}\frac{\partial P(n_1, N')}{\partial t} =& \alpha[P(n_1 - 1, N') - P(n_1, N')] \\ &+ r(N)\left[(n_1 - 1)P(n_1 - 1, N') - n_1 P(n_1, N')\right] \\ &+ \mu(N)\left[(n_1 + 1)P(n_1 + 1, N') - n_1 P(n_1, N')\right] + \\ &\alpha(H-1)[P(n_1, N' - 1) - P(n_1, N')] \\ &+ r(N)\left[(N'-1)P(n_1, N'-1) - N'P(n_1, N')\right] \\ &+ \mu(N)\left[(N'+1)P(n_1, N'+1) - N'P(n_1, N')\right]. \end{aligned} \tag{4.16}$$

This 2D problem can be approximated by a 1D problem when $n_1 \ll N'$ and $r(N) \approx r(N')$. The birth rate is regulated approximately by the "external" population decoupling it from n_1. Similarly, when $n_1 \gg N'$, $r(N) \approx r(n_1)$ and the birth rate is approximately independent

of N'. In either limit, the problem is approximately one-dimensional and can be modeled using a 1D master equation $P(n_1)$ (Eq. (4.14) with $r(N) \equiv r(N'(t))$ as an external force) or $P(N')$ (Eq. (4.3)) correspondingly. However, when n_1 and N' are comparable in size, one needs to evaluate the full 2D distribution $P(n_1, N')$ and marginalize over N' to obtain $P(n_1) = \sum_{N'=0}^{\infty} P(n_1, N')$ and finally

$$\langle c_k(t) \rangle = HP(n_1 = k; t) = H \sum_{N'=0}^{\infty} P(n_1 = k, N'; t). \tag{4.17}$$

This approach can be extended to higher dimensions to determine higher moments of $c_k(t)$, which are important for characterizing the variability of clone size distributions. Covariances $\mathrm{cov}(c_k, c_\ell) \equiv \langle c_k c_\ell \rangle - \langle c_k \rangle \langle c_\ell \rangle$, in particular, will illustrate the differences between the solutions to the mean-field model (Eq. (4.10)) and the exact model (Eq. (4.5)). In Subsection 4.7.4, we derive relationships between higher moments of c_k and the cell count distributions $P(n_1, n_2, ...)$. Specifically, for the second moments,

$$\langle c_k(t) c_\ell(t) \rangle = H(H-1)P(n_1 = k, n_2 = \ell; t) + \delta_{k,\ell} HP(n_1 = k; t). \tag{4.18}$$

4.3.2 Approximating $P(\{n_1, ..., n_q, N'\})$ by a q-dimensional Moran model

We now try to solve for $P(n_1, N')$. Since this 2D master equation does not usually have analytic solutions, we will show how to approximate $P(n_1, N')$ by a 1D two-species Moran model [PQP08,CN15,CRM16,CN17,CM17] with n_1 individuals of clone 1 and N' individuals of clone 2. The 1D Moran model imposes $n_1 + N' \equiv N$, the total population size, to be a fixed value.

We will first fix the value of N to be the quasi-steady-state value of the original unconstrained BDI process $N \to N^* \coloneqq \langle N^* \rangle$, at which the condition $\alpha H + r(N^*)N^* = \mu(N^*)N^*$ is satisfied. For example, under a logistic birth law (Eq. (4.11), the mean-field approximation Eq. (4.12) yields an accurate value of N^*. At this value of N^*, the growth and death rates take on specific values defined by $r^* \coloneqq r(N^*)$, $\mu(N^*) \coloneqq \mu^*$. In fact, to absolutely fix N^* the stochastic dynamics are driven by completely coupled birth-death events. During each event, one individual is randomly chosen to die and immediately replaced by a new one. This

tethering of birth and death ensures that the total population N^* is fixed. The total rate of a tethered birth-death event is $\frac{1}{2}(\alpha H + r^* N^* + \mu^* N^*) = \mu^* N^*$, where the factor $\frac{1}{2}$ compensates for the fact that two birth-death events occur simultaneously during one tethered event so on average the arrival rate of events has to be halved. Thus, $\mu^* N^*$ is the intrinsic rate of evolution in the Moran model. The master equation for the probability distribution $P_{\mathrm{M}}(n_1; t|N^*)$ of the fixed-N^* two-clone Moran model can be expressed as

$$\begin{aligned}\frac{\partial P_{\mathrm{M}}(n_1;t|N^*)}{\partial t} =& \omega_{12}(n_1-1|N^*)P_{\mathrm{M}}(n_1-1|N^*) + \omega_{21}(n_1+1|N^*)P_{\mathrm{M}}(n_1+1|N^*) \\ &- \left[\omega_{12}(n_1|N^*) + \omega_{21}(n_1|N^*)\right] P_{\mathrm{M}}(n_1|N^*), \end{aligned} \tag{4.19}$$

where the functions $\omega_{ji}(n|N^*)$ denote the rate that a clone i individual is replaced by a clone j individual in a Moran of fixed total population N^*

$$\begin{aligned}\omega_{12}(n|N^*) &= n\left(1-\frac{n}{N^*}\right) r^* + \left(1-\frac{n}{N^*}\right)\alpha \\ &= \mu^* N^* \left[(1-m^*)\frac{n}{N^*}\left(1-\frac{n}{N^*}\right) + m^* Q_1 \left(1-\frac{n}{N^*}\right)\right], \\ \omega_{21}(n|N^*) &= n\left(1-\frac{n}{N^*}\right) r^* + (H-1)\left(1-\frac{n}{N^*}\right)\alpha \\ &= \mu^* N^* \left[(1-m^*)\frac{n}{N^*}\left(1-\frac{n}{N^*}\right) + m^*(1-Q_1)\left(1-\frac{n}{N}\right)\right], \end{aligned} \tag{4.20}$$

where we have further defined

$$m^* \equiv \frac{\alpha H}{\mu^* N^*}, \quad Q_1 = \frac{1}{H}. \tag{4.21}$$

Here, m^* represents the relative total immigration rate and Q_1 is the fixed fraction of clone 1 amongst those from on the "mainland."

In these dynamics, it is clear that the probability of choosing an individual for removal/death from clone 1 and clone 2 are n_1/N^* and $1 - n_1/N^*$, respectively. The newly created individual has probability n_1/N^* to be of clone 1 and $1 - n/N^*$ to be of clone 2, calculated from the state of the model prior to death. Thus, after one event, the population of clone 1 may increase by 1 (if a clone 2 individual is chosen to die, and a clone 1 individual is chosen to be born) or decrease by 1 (if a clone 1 individual is chosen to die, and a clone 2 individual is chosen to be born). The total rate is for one clone augmented by the per-cell

immigration rate α, which is equal to the per-clone immigration rate since the cells initiating immigration are unique (see Figure 4.1). The total immigration into the "bath" clone is thus $(H-1)\alpha$.

To solve Eq. (4.19) in steady state, we use Eqs. (4.20) and invoke the detailed balance condition $\omega_{12}(n_1-1|N^*)P^*_{\rm M}(n_1-1|N^*)=\omega_{21}(n_1|N^*)P^*_{\rm M}(n_1|N^*)$ to obtain

$$P^*_{\rm M}(n_1|N^*)=P^*_{\rm M}(0|N^*)\frac{\omega_{12}(0|N^*)}{\omega_{21}(n_1|N^*)}\prod_{\ell=1}^{n_1-1}\frac{\omega_{12}(\ell|N^*)}{\omega_{21}(\ell|N^*)},\quad P^*_{\rm M}(0|N^*)=\left[\sum_{n_1=0}^{N}\prod_{\ell=1}^{n_1}\frac{\omega_{12}(\ell-1|N^*)}{\omega_{21}(\ell|N^*)}\right]^{-1}. \tag{4.22}$$

For general q-dimensional ($q\geq 2$) Moran models that involve ($q+1\geq 3$) subpopulations, closed form solutions are difficult to obtain. However, we can approximate these models using a diffusion approximation that treats the clonal fractions $x_i=n_i/N^*$ $(1\leq i\leq q)$ as continuous variables. After Taylor-expanding $q-$dimensional discrete master equations and assuming $m^*\equiv\frac{\alpha H}{\mu^* N^*}\ll 1$, a simple $q-$dimensional Fokker-Planck equation can be derived [Kim64, BM07]

$$\frac{\partial P_{\rm M}(\mathbf{x}|N^*)}{\partial t}=\mu^*N^*\left[-\frac{1}{N^*}\sum_{i=1}^{q}\frac{\partial A_i(\mathbf{x})P_{\rm M}(\mathbf{x}|N^*)}{\partial x_i}+\frac{1}{(N^*)^2}\sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\partial^2 B_{ij}(\mathbf{x})P_{\rm M}(\mathbf{x}|N^*)}{\partial x_i\partial x_j}\right] \tag{4.23}$$

where

$$A_i(\mathbf{x})=\sum_{j=1}^{q}m^*(Q_i-x_i),\quad B_{ii}(\mathbf{x})=x_i(1-x_i),\quad B_{ij}(\mathbf{x})=-x_ix_j\ (i\neq j). \tag{4.24}$$

For example, when $q=2$ (three clones), we have $Q_1=Q_2=\frac{1}{H}$, $Q_3=\frac{H-2}{H}$. We explicitly show the derivations for the 1D and 2D Fokker-Planck equations in Subsection 4.7.5. The exact QSS solution of the general q-dimensional diffusion model is known and follows the Dirichlet distribution [BBM07]

$$P^*_{\rm M}(\mathbf{x}|N^*)=\Gamma(2N^*m^*)\prod_{i=1}^{q}\frac{x_i^{2N^*m^*Q_i-1}}{\Gamma(2N^*m^*Q_i)}. \tag{4.25}$$

4.3.3 Relaxing the hard population constraint of the Moran model

While the Moran model can be used to approximate $P(n_1,N')$, it includes an additional hard constraint $n_1+N'=N^*$ that is not imposed in the original BDI model. In fact, n_1

itself can fluctuate above N^*. To relax this constraint and find an improved approximation to the reduced QSS distribution $P^*(\{n_1, n_2, ..., n_q\})$, we simply allow the system size of the Moran process to vary and weight each QSS Moran process by the steady-state probability distribution over the total population

$$P^*(N) = \frac{\prod_{j=1}^{N} \frac{r(j-1)+\alpha H}{\mu(j)}}{\left(\sum_{N=0}^{\infty} \prod_{\ell=1}^{N} \frac{r(\ell-1)+\alpha H}{\mu(\ell)}\right)}, \tag{4.26}$$

which is readily obtained from solving Eq. (4.3), the master equation for the total population of the BDI process. We thus use a whole family of Moran models, each at a different value of N, weighted by $P^*(N)$ to find a "convolved" QSS probability

$$P^*(\{n_1, n_2, ..., n_q\}) = \sum_{N=1}^{\infty} P^*_{\mathrm{M}}\left(\{n_1, n_2, ...n_{q+1}\}|N\right) P^*(N). \tag{4.27}$$

In the above summation the total population of the Moran models can vary around $N = \{1, 2, ..., N^* - 1, N^*, N^* + 1, ...\}$. Different values of the system size will yield different values of the rates $\omega_{ji}(n|N)$ according to Eq. (4.20). In 1D, according to Eq. (4.20), the ratio $\frac{\omega_{12}(\ell|N)}{\omega_{21}(\ell|N)}$ varies with N according to

$$\frac{\omega_{12}(\ell|N)}{\omega_{21}(\ell|N)} = \frac{(1-m^*)\frac{\ell}{N}\left(1-\frac{\ell}{N}\right) + m^* Q_1 \left(1-\frac{\ell}{N}\right)}{(1-m^*)\frac{\ell}{N}\left(1-\frac{\ell}{N}\right) + m^*(1-Q_1)\left(1-\frac{\ell}{N}\right)}, \tag{4.28}$$

where we have kept the intrinsic rates r^* and μ^* and the relative immigration rate m^* *fixed*. The only terms in Eq. (4.28) that vary with N are the probability factors ℓ/N and $1 - \ell/N$. By keeping the r^*, μ^*, and m^* fixed, we preserve the relative rates of tethered birth, death, and immigration that define the original BDI process.

4.4 Results

4.4.1 $\langle c_k \rangle$ and $\langle c_k c_\ell \rangle$ under logistic growth

In Figure 4.4, we plot results from Monte-Carlo simulations, the mean-field solutions to Eq. (4.10), and the numerically weighted solutions of Eq. (4.27) in which $P^*_{\mathrm{M}}(n_1|N)$ is taken from Eq. (4.22) but with varying N as defined in Eq. (4.28). This improved weighted solution yields accurate expected QSS clone count distributions $\langle c_k^* \rangle$.

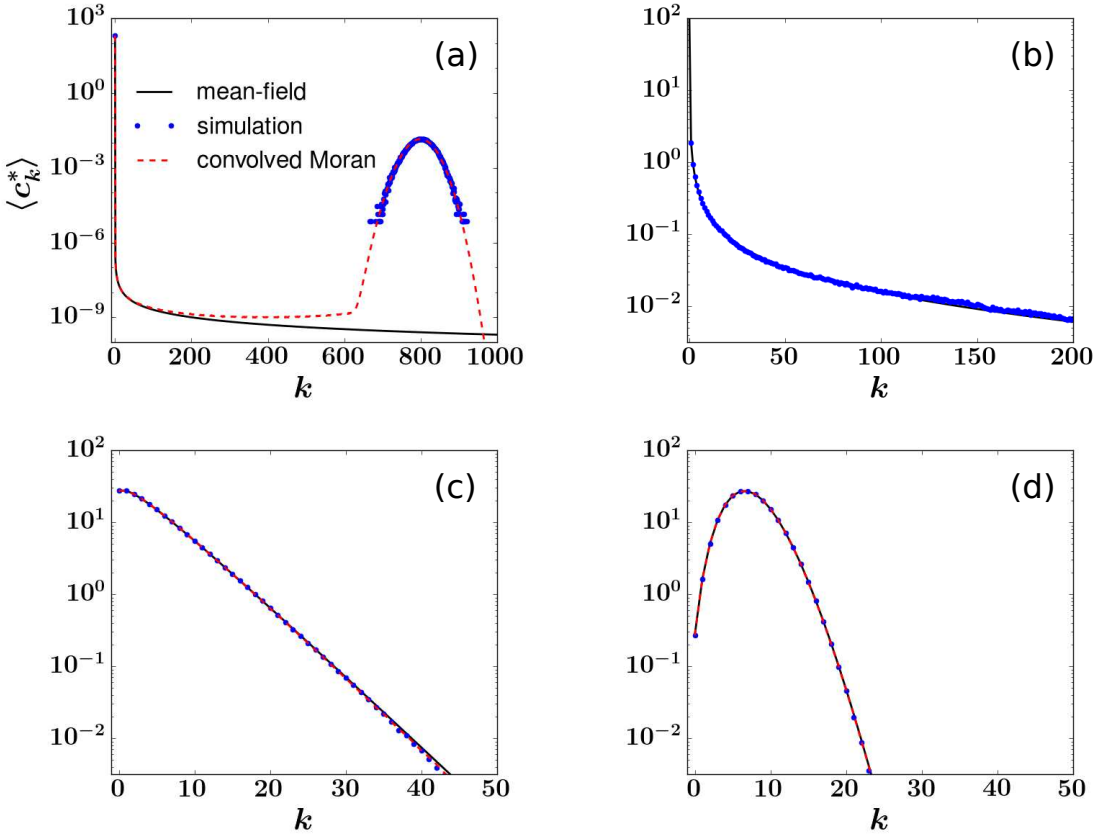

Figure 4.4: Results under Logistic birth

Simulated (blue dots), mean-field (solid black), and weighted Moran (red dashed) approximations of $\langle c_k^* \rangle$ using logistic growth laws and the parameters $\mu = 10,\ p = 20,\ K = 1600,\ H = 200$. Immigration rates used were (a) $\alpha = 10^{-8}$, (b) $\alpha = 0.1$, (c) $\alpha = 10$, and (d) $\alpha = 60$, as in Figure 4.3. The weighted QSS Moran model approach yields a very accurate approximation to the simulated values of $\langle c_k^* \rangle$ for all values of α, including small α as shown in (a).

To calculate the covariance between c_k^* and c_ℓ^* at QSS, we use the 2D ($q = 2$) "continuum" solution given in Eq. (4.25) in the weighting in Eq. (4.27) in order to numerically compute Eq. (4.18).

The covariances $\mathrm{cov}(c_k^*, c_\ell^*) \equiv \langle c_k^* c_\ell^* \rangle - \langle c_k^* \rangle \langle c_\ell^* \rangle$ with $\alpha = 10^{-8}$, both from Monte-Carlo simulations and from our weighted Moran model approximation, are plotted in Figure 4.5. The results provide insight on how the true dynamics for $\langle c_k^* \rangle$ in Eq. (4.5) differs from that of the mean-field description in Eq. (4.10).

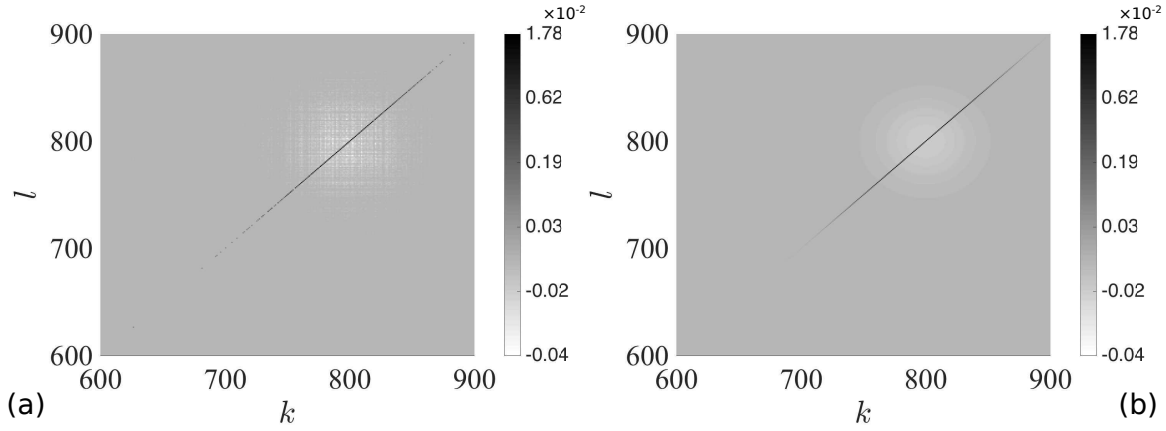

Figure 4.5: Covariances under Logistic birth

$\mathrm{cov}(c_k, c_\ell)$ in the simulations (a) and from our calculations (b). Parameters are $\alpha = 10^{-8},\ \mu = 10,\ p = 20,\ K = 1600,\ H = 200$. Only the interesting ranges of k and ℓ close to $N^* \approx 800$ are shown. The pattern shows that large clone counts are positively self-correlated (black line) but are negatively correlated with neighboring counts (white dots). The grey background shows no correlation between farther-away clone counts.

The large values (black line) along the diagonal $k = \ell$ is indicate $\langle c_k^* c_\ell^* \rangle$ for $k \sim \ell \sim N^* \approx 800$. White regions in the off-diagonal areas imply negative correlation between clone counts of large neighboring sizes. In other words, whenever we observe a clone of 800 individuals in a simulation at any fixed time t (at QSS), we will probably not observe another clone with 801 cells at the same time. Gray areas that are farther away (such as $k = \ell = 600$) represent transient states of the system and have near-zero covariances.

4.4.2 Other forms of global interactions

Since global interactions across all clones mediate the breakdown of the mean-field approximation, we now investigate different forms of regulation imposed through the functions $r(N)$ and $\mu(N)$. To explore how the "stiffness" of different total population constraints affects the expected QSS clone-count vector $\langle c_k^* \rangle$, we consider a simple Hill-type birth function with

Hill coefficient 1:

$$r(N) = \frac{p_2 K_2}{K_2 + N}, \qquad \mu(N) = \mu_2. \tag{4.29}$$

This form imposes a "softer" constraint on the total population N than the Logistic birth function. In order to compare the results with those of the Logistic model in subsection 4.4.1, we use the same values of α and H and use $\mu_2 = \mu,\ p_2 = p$ and $K_2 = K - N^*$ where N^* is the QSS population size obtained from the Logistic model. In this way, the Hill-type model generates the same steady state total population size $N^*_{\rm H} = N^*$. In Figure 4.4.2, we plot results of $\langle c^*_k \rangle$ from our convolved model, from the mean-field approximation, and from simulations.

Now consider a constant birth but density-dependent death rate of the form [PQP08]

$$r(N) = r_3, \qquad \mu(N) = \mu_3(1 + \frac{N}{K_3}). \tag{4.30}$$

Again, we are interested in clonal dynamics near the same N^* as in subsection 4.4.1. Besides the same α and H, we set $K_3 = K,\ r_3 = r^*,\ \mu_3 = \mu/(1 + \frac{N^*}{K_3})$. Simulations and results are plotted in Figure 4.7.

4.4.3 "Stiffness" of regulation

Although the above examples share the same Moran model $P_{\rm M}(x)$, they differ in their $P^*(N)$ distributions. In panels (a) of Figures 4.4, 4.4.2, and 4.7, the differences in the $P(N)$ are illustrated by the different "widths" of the peak near N^*. Generally, a wider peak represents a "softer" (less stiff) total population constraint. According to simulations and numerical solutions of the convolved model shown by the figures, the levels of stiffness of the regulatory effects are ranked by logistic > hill-type > density-dependent death.

As long as $f(N)$ on the right-hand side of Eq. (4.4) is a differentiable function of N, "stiffness" near N^* may be defined by its first derivative

$$|f'_*| = -f'_* \equiv \left.\frac{{\rm d}f(N)}{{\rm d}N}\right|_{N^*} \in [0, +\infty). \tag{4.31}$$

f'_* takes negative value as long as N^* is a locally stable state. The larger $|f'_*|$ is, the more likely the next event will be "compensatory" (e.g. a new birth increases the chance for the next

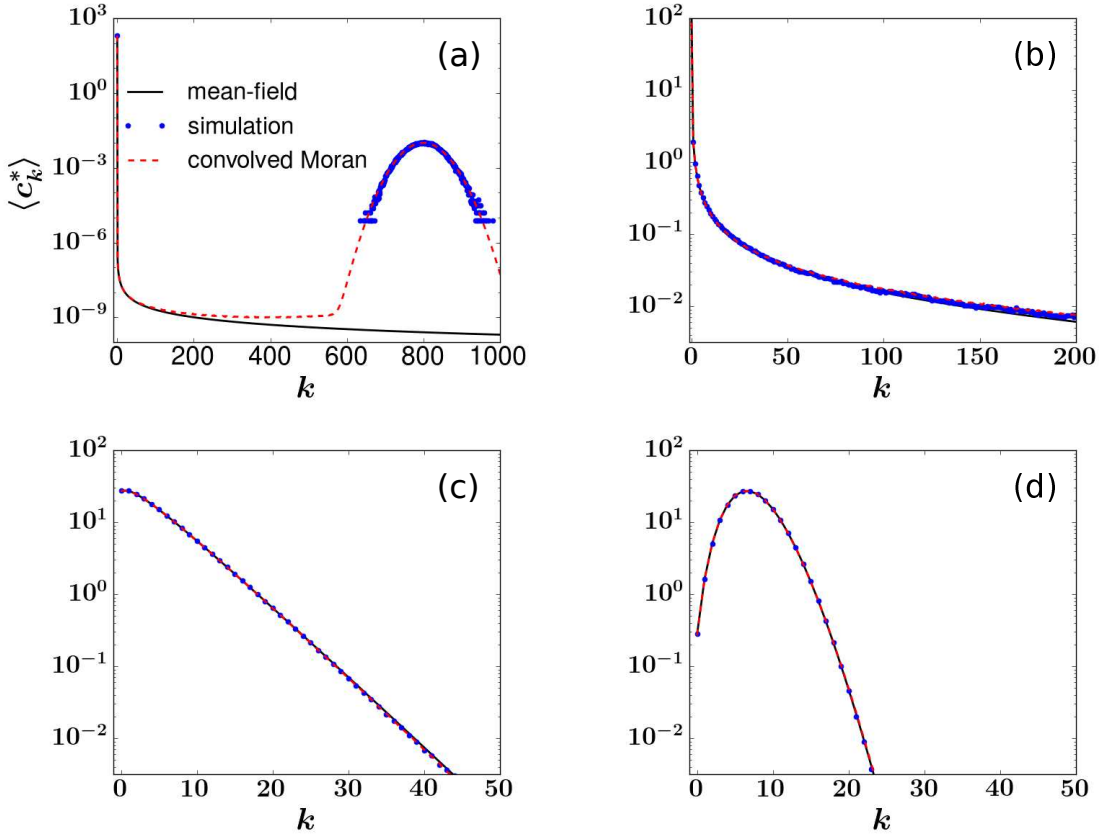

Figure 4.6: Results under Hill-type birth

Comparison of $\langle c_k^* \rangle$ distributions from simulations, mean-field model (green dash dot), and convolved Moran model (red solid) under Hill-type birth and for different immigration rates $\alpha = 10^{-8}$ (a), 0.1 (b), 10 (c), and 60 (d). Other parameters are $\mu_2 = 10,\ p_2 = 20,\ H = 200$. The half-saturation size K_2 was set such that the model has the same N^* as in Figure 4.4. Thus K_2 differs under different α.

event to be death). It measures how much response was given by the model as N randomly increases/decreases. In our three examples (under $\alpha = 10^{-8}$), $f'_* = p - \mu - \frac{2p}{K}N^* = -10$ for Logistic birth, $f'_* = p_2 \frac{(N^*)^2 + 2K_2 N^*}{(N^* + K_2)^2} - \mu_2 = -5$ for Hill-type birth, and $f'_* = r_3 - \mu_3 - \frac{2\mu_3}{K_3}N^* = -3.3$ for density-dependent death, which is consistent to the ranking of the levels of stiffness shown by Figures 4.4(a), 4.4.2(a), and 4.7(a).

We may extend our definition of the stiffness to cases where $f(N)$ is not differentiable. For example, the Moran model has an infinitely "stiff" constraint ($f'_* = -\infty$) which "forces"

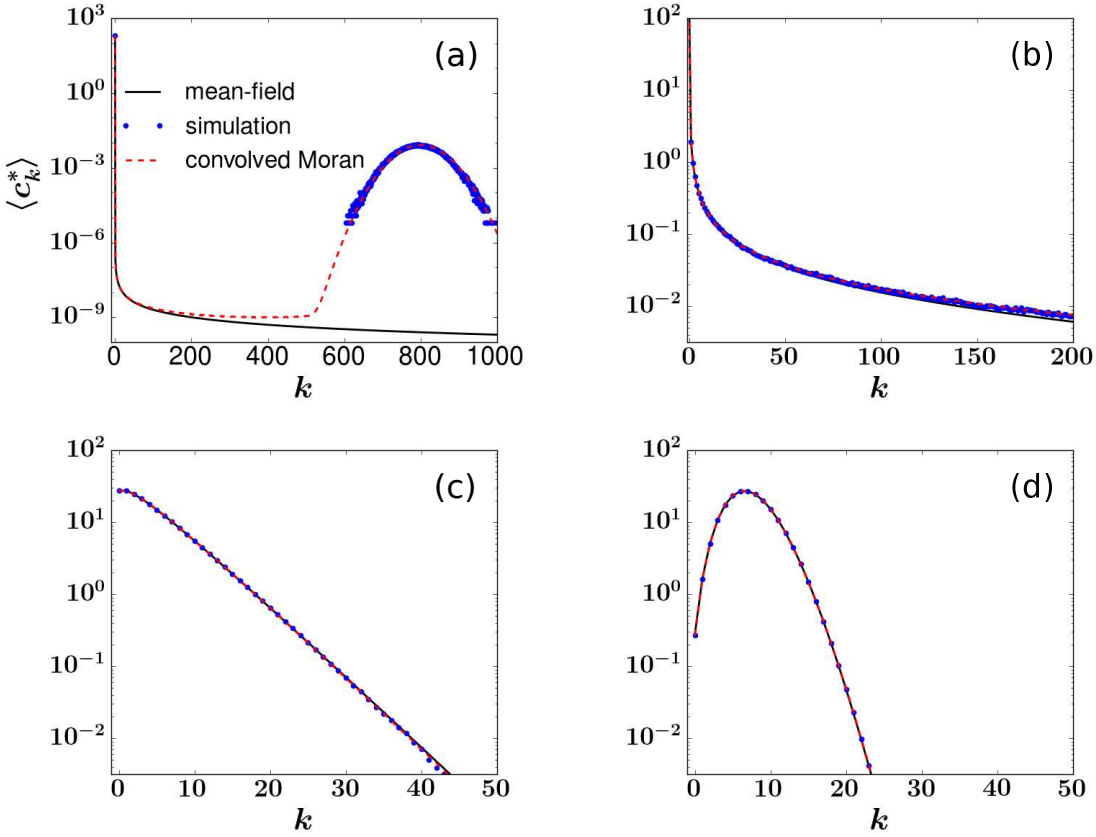

Figure 4.7: Results under density-dependent death

Comparisons of $\langle c_k^* \rangle$ distributions from simulations, mean-field model (green dash dot), and convolved Moran model (red solid) under density-dependent death. Immigration rates are $\alpha = 10^{-8}$ (a), 0.1 (b), 10 (c), and 60 (d). Other parameters are $K_3 = 1600$, $H = 200$. r_3 and μ_3 were set such that the model has the same N^* and QSS rates of birth and death as in Figure 4.4.

an immediately death after a new birth. Nevertheless, as we have shown, even though a regulated BDI model may have much less sensitivity than the Moran model, the latter still provides insights on how such regulatory effect would induce a local stable state at N^*. Also, the larger $|f'_*|$ is, the better the model is approximated by the Moran model. A model with constant birth and death rates, on the other hand, has $f'_* = 0$.

4.5 Phase transition in the clonal distribution

In this section, we aim to interpret the failure of the mean-field equation (Eq. (4.10)) as a result of a "phase-transition" in the clonal statistics. As discussed in Section 4.4, the full high-dimensional BDI model can be approximately decomposed into two components, (1) evolution of different clones' fractions (the Moran model, which is the same across the Logistic birth, Hill-type birth, and density-dependent death) and (2) fluctuations of the total population size (the total population model, which differs for different models). We will focus on the first part since the failure of the mean-field approach essentially arises from the emergence of a clone that represents a large fraction of the whole population.

Phase-transition behavior can be conveniently visualized using a potential energy landscape ϕ. This analytical tool has been widely used in population genetics and developmental biology [Wri32, Wad57, She97, APJ01, Ao09, Orr09, XJJ14]. Its recent development from the physics community has allowed its applications to quantitative and systems biology [Ao04, Qia06, WXW08]. Defined as a measure of "generalized energy", its gradient indicates the direction of evolution of the system and its minima (potential wells) denote local stable states.

4.5.1 Energy landscape as an analytical tool

We employ $\phi(\{x_1, x_2, ...\})$, the energy landscape function at system state $\{x_1 = \frac{n_1}{N^*}, x_2 = \frac{n_2}{N^*}, ...\}$ (N^* is fixed) to study how the Moran-type clonal dynamics vary with parameters. The shape of ϕ across $\{x_1, x_2, ...\}$ characterizes the global stability of the model.

First, write Eq. (4.24) in 1D

$$A(x) = m^* \left(\frac{1}{H} - x\right), \; B(x) = x(1-x). \tag{4.32}$$

The energy function is obtained as [XJJ14]

$$\begin{aligned}\phi(x) &\equiv -\int^x \frac{A(y)}{B(y)/N^*} + \ln B(x) = \left(1 - \frac{\alpha}{\mu^*}\right)\ln(x) + \left(1 - \frac{\alpha(H-1)}{\mu^*}\right)\ln(1-x) \\ &\equiv \frac{1}{H-1}\left[(H-1) - \frac{\alpha}{\alpha_c}\right]\ln(x) + \left(1 - \frac{\alpha}{\alpha_c}\right)\ln(1-x)\end{aligned} \tag{4.33}$$

where we have defined

$$\alpha_{\rm c} = \frac{\mu^*}{H-1}, \tag{4.34}$$

a critical value of α that controls a "phase transition." When $P^*_{\rm M}(x)$ is normalizable, the energy function satisfies $\phi(x) \propto -\ln P^*_{\rm M}$. Since $\ln(0) = -\infty$ and $\ln(1) = 0$, the shape of $\phi(x)$ is determined by the signs of the coefficients $H - 1 - \frac{\alpha}{\alpha_{\rm c}}$ and $1 - \frac{\alpha}{\alpha_{\rm c}}$.

Assuming $\alpha_{\rm c} > 0$ (see Discussion for the special case $\alpha_{\rm c} = 0$), different regimes of the model can be delineated

- When $\alpha < \alpha_{\rm c}$, we have $\frac{\alpha}{\alpha_{\rm c}} < 1$, $\frac{\alpha}{\alpha_{\rm c}} < H - 1$. Two potential wells in $\phi(x)$ emerge at 0 and 1 with two associated basins of attraction as shown in Figure 4.8(a). All clones starts with a small fraction $x_i \approx 0$. One of them has the chance to transit to the attractive peak $x = 1$ and causes the failure of the mean-field description. However, this transition is different from the usual stochastically-driven "escape" in statistical physics (see Discussion). When α is extremely small, the "extinction" state $x = 0$ is an absorbing state for each clone and the mean-field approximation fails severely.

- When $\alpha = \alpha_{\rm c}$, we have $\frac{\alpha}{\alpha_{\rm c}} = 1$, $\frac{\alpha}{\alpha_{\rm c}} < H - 1$. The potential exhibits a well at 0 and a peak at 1. The whole interval $[0, 1]$ is a basin of attraction for $x = 0$ as shown by Figure 4.8(b). The severity of the failure of the mean-field approach is sensitive to α when it is near $\alpha_{\rm c}$.

- When $\alpha_{\rm c} < \alpha \leq (H-1)\alpha_{\rm c}$, we have $\frac{\alpha}{\alpha_{\rm c}} > 1$, $\frac{\alpha}{\alpha_{\rm c}} \leq H - 1$. The landscape remains single-peaked at $x = 0$ but the energy increases much more quickly as x increases (Figure 4.8(c)). The mean-field approach is accurate in this regime.

- When $\alpha > (H-1)\alpha_{\rm c}$, there is a single potential well in $\phi(x)$ appearing at some value $x > 0$. The peak location $x_{\rm peak} = \frac{\alpha - \alpha_{\rm c}(H-1)}{\alpha H - 2\alpha_{\rm c}(H-1)}$ is close to $x = 0$ when $H \gg 1$. Two energy peaks emerge at both $x = 0$ and $x = 1$ and the basin of attraction is the whole $[0, 1]$ interval as shown by Figure 4.8(d). The mean-field approach does not fail here.

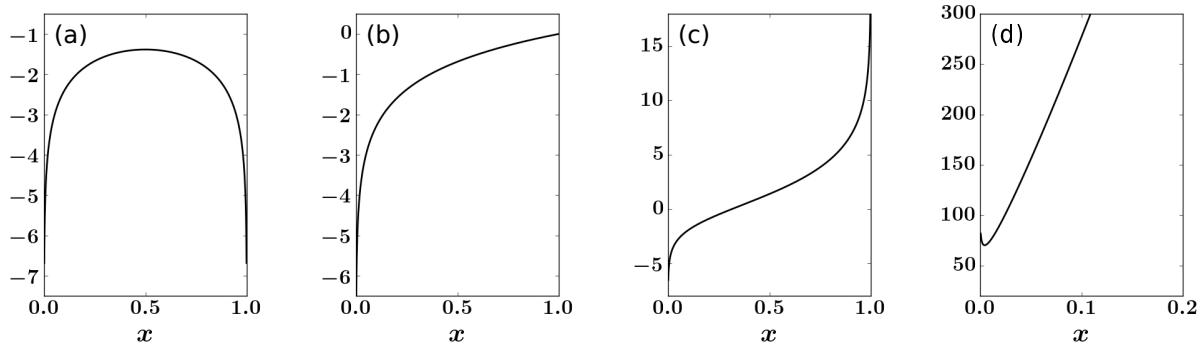

Figure 4.8: Energy landscapes

Energy landscapes $\phi(x)$ as a function of x, the fraction of a specific clone when $\alpha = 10^{-8}$, 0.05 ($= \alpha_c$), 0.2, and 120 in (a-d) respectively. Other parameters are $\mu = 10$, $p = 20$, $K = 1600$, and $H = 200$.

4.5.2 Resolving the effects of α and H

Equipped with the energy landscape, we can examine how clonal dynamics look like when varying the per-stem-cell differentiation rate α and the total number of stem cells H but fixing their product αH. The product αH determines how immigration contributes to the dynamics of the total population N. Varying parameters in this way should not change the dynamics of N but affect the resolved dynamics of individual clones.

This is readily shown by the different shapes of $\phi(x)$ as α and H change. For example, let $\mu = 0.33$. With $H = 2,\ \alpha = 1$, landscape where a positive "most probable" clone size is maintained by high per-clone immigration rate. However, if $H = 200$ and $\alpha = 0.01$, each clone has a low immigration rate. The associated landscape $\phi(x) = 0.97 \ln x - 5 \ln(1 - x)$ exhibits a unique potential well at 0 as all clones tend to vanish.

We can also consider the limit $H \to \infty$ while fixing a finite αH. This limit is appropriate for naive T cell generation by the thymus. While total thymic output αH is finite, there are theoretically $\geq 10^{18}$ different clones (T-cell receptor sequences) that can be generated. In this limit, every immigration leads to a new clone and $k = 0$ is an absorbing boundary for all existing clones. Clone labels keep changing, but the distribution of c_k reaches a QSS. The

energy landscape becomes (taking $\alpha \to 0$ in Eq. (4.33)) $\phi(x) = \ln(x) + \left(1 - \frac{\alpha H}{\mu^*}\right) \ln(1-x)$. There is always an potential well at 0 while the dynamics near 1 depend on the sign of $\frac{2\alpha H}{\mu^*} - 1$.

Recall that the condition for the failure of the mean-field approach is $\alpha < \alpha_c$. From Eq. (4.34), the smaller H is, the larger α_c becomes, thus the more likely the mean-field approach would fail. When $\alpha_c = 0$ (e.g. realized by $\mu^* = 0$), Eq. (4.33) no longer has a valid form. Birth becomes negative ($N^* > K$) to balance immigrations. Nevertheless, we can multiply the landscape function by a constant μ^* without affecting its ability to qualitatively characterize and classify the transient dynamics of the system. Such transient dynamics drive clone sizes towards their steady state distribution, whether it has a valid form or not. We then take the limit $\mu^* \to 0$ and get $\phi \propto -\alpha \ln(x) - \alpha(H-1)\ln(1-x)$, which always has a unique potential well between $(0, 1)$, corresponding to Figure 4.8(d).

4.6 Summary and Discussion

In this paper, we were able to map the q^{th} moment of $\{c_k\}$ to a $(q+1)$-dimensional cell-count BDI model, which is then approximated by a q-dimensional Moran model. The expected distribution and covariances of clone counts were accurately calculated in parameter regimes in which the mean-field approximations break down. By exploiting the concept of energy landscapes, we analytically describe a phase transition in clonal dynamics which explains the failure of the mean-field approach in the original model. Our analysis shows that global (inter-clonal) carrying capacity, when combined with a random sampling mechanism, generates a genetic-drift-like effect [Ewe12] in a Moran model that ultimately destroys the universal power-law distribution of c_k.

In Eq. (4.5), dynamics of any $\langle c_k \rangle$ are controlled by $r(N)$, where $N \equiv \sum_\ell \ell c_\ell$. It involves contributions from all clone populations $\ell = 0, 1, 2, ..., k, ...$ Recall that in many classical scenarios the relative strengths of these effects on the k^{th} dimension decay with distance $|k - \ell|$. For example, in the constant-rate BDI model, only $c_{k\pm1}$ and c_k affect the dynamics of c_k. Here, however, contributions from c_ℓ is proportional to the index ℓ itself instead of

on $|k-\ell|$. This is a type of "long-range" interaction or long-distance coupling, arising as a theoretical challenge across many fields [MS87, Tak12, SLJ12, CVJ16]. Thus, the evolution of $\langle c_k \rangle$ according to Eq. (4.5) can no longer be approximated by diffusive-like dynamics across k. This is clearly shown in Figure (4.5) where correlations between large k and its neighboring states $k \pm 1$ are negative.

High-dimensional diffusion models conveniently provide analytic-form steady-state distributions [Ewe63, Aal89]. However, the boundary values of $P^*_{\mathrm{M}}(\mathbf{x})$ in the diffusion model may not accurately approximate those from the discrete Moran model, especially in higher dimensions. For example, when $2Nm_i \ll 1$ in Eq. (4.25), $P_{\text{diffusion}}(x_i = 0) = +\infty$ but $P_{\text{diffusion}}(x_i = \frac{1}{N}) \approx 0$. Near the boundary, $P_{\text{diffusion}}(0 < x < \frac{1}{N})$ generally changes in a highly non-linear fashion. Only with extremely large N, the probability densities of the discrete and continuous Moran models match. Nevertheless, our result in Figure (4.5) is accurate because the range is far away from the boundaries; also, the second moment $\langle c_k^2 \rangle$ in Eq. (4.18) turns out to be dominated by the first-moment term $H\langle c_k \rangle$, which was calculated based on the exact Eq. (4.22). Small k region under various α's show qualitative consistency to the simulations (data not shown).

Curiously, failure of the mean-field assumption under such simple additional force has not been noticed or explicitly mentioned in the literature. In fact, the consequences of density-dependence have been rarely discussed in the context of species diversity [VBH05]. Such a global effect will break independence between clones and can be difficult to model. On the other hand, failure of predicting a large-size clone by Eq. (4.10), was not mentioned before. First, most empirical studies focused on the small-to-intermediate range of k, where the c_k distribution is well approximated by the mean field model (Eq. (4.10)). As seen in Figures 4.4 and 4.4.2, the mean-field $\bar{c}^*_k$ and our $\langle c^*_k \rangle$ match well in the small k range in all ranges of α. Also, even though one or a few large clones/species were always observed, they might be believed to be favored by selection effect, instead of by the (simpler) explanation of global carrying capacity.

Importantly, global carrying capacity is ubiquitous in cellular populations (e.g. bone marrow for progenitors) and ecological systems (e.g. physical space for all species). This largest

'outlier' clone contains most cells in the compartment and can be biologically more important than all other smaller clones/species in the organism/community. Correctly predicting it provides a correct "null hypothesis" of neutral dynamics. Otherwise, one may incorrectly argue that the existence of such a singular 'outlier' clone suggests a selection effect. It is also required for the mathematical consistency of the model and numerical calculations.

It is worth noting that the initial establishment of the large clone i, denoted by the transition $x_i \approx 0 \to x_i \approx 1$, is different from traditional scenarios where clone i randomly cross the energy peak near $x = 0.5$ and "escapes" to the other attractive basin. Here, as most clones vanish, this clone was randomly selected to be "pushed" towards $x = 1$ to compensate for the loss of all other clones. However, the classical escape problem can still be observed from the model after clone i dominates the whole population. Specifically, consider an event in which a different clone j starts from a small fraction $x_j \approx 0$ at QSS but subsequently replaces clone i ($x_j \approx 1$). The waiting time for such replacement event was obtained by [XJJ14] as $T_{\mathrm{r}} \sim \mathcal{O}(\frac{\mu N^*}{\alpha})$, a much longer time than the waiting time $T_2 \sim \mathcal{O}(\frac{N^*}{\mu})$ (Subsection 4.7.3) for the first dominant clone to emerge in our BDI model.

Future work includes more accurately determining higher-dimensional steady state solutions of the diffusion models, especially near the boundaries; adding clone-wise or cell-wise heterogeneities, where the current "one-VS-others" view may break down since clone labels are no longer exchangeable, and mapping time-dependent BDI models to time-dependent Moran models; and find direct applications of the current results in empirical studies.

4.7 Appendices

4.7.1 Dynamical equation for $P(\mathbf{n},t) \equiv P(n_1,...,n_i,...,n_H;t)$

The high-dimensional master equation obeyed by the full multispecies distribution reads

$$\begin{aligned}\frac{\mathrm{d}P(\mathbf{n})}{\mathrm{d}t} &= \alpha\sum_{i=1}^{H}[P(n_1,...,n_{i-1},n_i-1,n_{i+1},...,n_H)-P(\mathbf{n})]\\ &+\sum_{i=1}^{H}[r(N-1)(n_i-1)P(n_1,...,n_{i-1},n_i-1,n_{i+1},...,n_H)-r(N)n_iP(\mathbf{n})]\\ &+\sum_{i=1}^{H}[\mu(N+1)(n_i+1)P(n_1,...,n_{i-1},n_i+1,n_{i+1},...,n_H)-\mu(N)n_iP(\mathbf{n})] \quad (4.35)\end{aligned}$$

where $N \equiv \sum_{i=1}^{H} n_i$.

4.7.2 Dynamical equations for $\langle c_\ell \rangle$

Define $P(\mathbf{c},t)$ as the probability of observing the configuration $\mathbf{c} = \{c_0, c_1, c_2, ...\}$ at a specific time t. Under constant immigration rate and density-regulated birth and death rates, the evolution of the full probability distribution satisfies the master equation

$$\begin{aligned}\frac{\mathrm{d}P(\mathbf{c})}{\mathrm{d}t} &= -\sum_{k=0}^{\infty}[\alpha+(\mu(N)+r(N))k]c_kP(\mathbf{c})\\ &+\sum_{k=0}^{\infty}(c_{k+1}+1)(k+1)\mu(N+1)P(\{...,c_k-1,c_{k+1}+1,...\})\\ &+\sum_{k=0}^{\infty}(c_k+1)(\alpha+kr(N-1))P(\{...,c_k+1,c_{k+1}-1,...\}). \quad (4.36)\end{aligned}$$

Without loss of generality, let us assume constant $\mu,\ \alpha$ but regulated $r = r(N) = r(\sum_{k=0}^{\infty} kc_k)$. The expected count of clone is (the time argument has been neglected for simplicity)

$$\langle c_\ell \rangle = \sum_{c_\ell=0}^{H} c_\ell P(c_\ell) = \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \ c_\ell P(c_0,c_1,c_2,c_3,...,c_{\ell-1},c_\ell,c_{\ell+1},...). \quad (4.37)$$

Substituting Eq. (4.36) into Eq. (4.37), we obtain

$$\frac{\mathrm{d}\langle c_\ell\rangle}{\mathrm{d}t} = \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \; c_\ell \frac{\mathrm{d}P(c_0,c_1,c_2,c_3,...,c_{k-1},c_k,c_{k+1},...)}{\mathrm{d}t}$$

$$= \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \; c_\ell \Bigg\{ \sum_{k=0}^{\infty} -[\alpha+(\mu+r(N))k]c_k P(c_0,c_1,...,c_{k-1},c_k,c_{k+1},...)$$

$$+\sum_{k=0}^{\infty}(c_{k+1}+1)[(k+1)\mu]P(c_0,c_1,...,c_{k-1},c_k-1,c_{k+1}+1,...)$$

$$+\sum_{k=0}^{\infty}(c_k+1)[\alpha+kr(N-1)]P(c_0,c_1,...,c_{k-1},c_k+1,c_{k+1}-1,...)\Bigg\}. \qquad (4.38)$$

By collecting only terms in Eq. (4.38) that involve $r(N)$, we obtain two summations

$$S_1+S_2 \equiv -\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \; c_\ell \sum_{k=0}^{\infty} r(N)kc_k P(c_0,c_1,...,c_{k-1},c_k,c_{k+1},...)$$

$$+\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \; c_\ell \sum_{k=0}^{\infty} r(N-1)k(c_k+1)P(c_0,c_1,...,c_{k-1},c_k+1,c_{k+1}-1,...).$$

Consider the contribution of the k^{th} terms in both summations:

- When $k<\ell-1$ or $k\geq\ell+1$, the k^{th} term of S_1 becomes

$$-\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... \; c_2 r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_k,...) \qquad (4.39)$$

  and k^{th} term of S_2 becomes

$$\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H}... \; c_\ell r(N-1)(k-1)(c_{k-1}+1)P(c_0,c_1,c_2,...,c_{k-1}+1,c_k-1,...)$$

$$=\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H-1}... \; c_\ell r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_{k-1},c_k,...)$$

$$=\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H}... \; c_\ell r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_{k-1},c_k,...). \qquad (4.40)$$

  The last equality holds because $P(c_k=H)=0$ (if $c_k=H$, then all other $c_{m\neq k}=0$ and the equation still holds).

- When $k=\ell-1$, we have the k^{th} term of S_1

$$-\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}... \; c_\ell r(N)(\ell-1)c_{\ell-1}P(c_0,c_1,c_2,...,c_k,...) \qquad (4.41)$$

and the k^{th} term of S_2

$$\begin{aligned}
&\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}... \, c_\ell r(N-1)(\ell-1)(c_{\ell-1}+1)P(c_0,c_{\ell-1}+1,c_\ell-1,c_3,...,c_k,...)\\
&\quad=\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H-1}... \, (c_\ell+1)r(N)c_{\ell-1}P(c_0,c_1,c_2,c_3,...,c_k,...)\\
&\quad=\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}... \, (c_\ell+1)r(N)c_{\ell-1}P(c_0,c_1,c_2,c_3,...,c_k,...). \qquad (4.42)
\end{aligned}$$

The two terms sum up to

$$\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}\sum_{c_2=0}^{H}\sum_{c_3=0}^{H}...\sum_{c_k=0}^{H}... \, r(N)c_{\ell-1}P(c_0,c_1,c_2,c_3,...,c_k,...)=\langle r(N)c_{\ell-1}\rangle. \qquad (4.43)$$

- When $k=\ell$, we have the k^{th} term of S_1

$$-\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}... \, c_\ell r(N)\ell c_\ell P(c_0,c_1,c_2,...,c_k,...) \qquad (4.44)$$

and the k^{th} term of S_2

$$\begin{aligned}
&\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H}... \, c_\ell r(N-1)\ell(c_\ell+1)P(c_0,c_1,c_\ell+1,c_{\ell+1}-1,...,c_k,...)\\
&\quad=\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H-1}... \, (c_\ell-1)r(N)\ell c_\ell P(c_0,c_1,c_2,c_3,...,c_k,...)\\
&\quad=\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H}... \, (c_\ell-1)r(N)\ell c_\ell P(c_0,c_1,c_2,c_3,...,c_k,...). \qquad (4.45)
\end{aligned}$$

The two terms sum up to

$$\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H}...r(N)\ell(-c_\ell)P(c_0,c_1,c_2,c_3,...,c_k,...)=-\ell\langle r(N)c_\ell\rangle. \qquad (4.46)$$

To sum up, terms that involve $r(N)$ in Eq. (4.38) are simplified as $(\ell-1)\langle r(N)c_{\ell-1}\rangle - \ell\langle r(N)c_\ell\rangle$. Terms involving α and μ can be similarly obtained if they are regulated by N. Together, Eq. (4.38) becomes

$$\frac{\mathrm{d}\langle c_\ell\rangle}{\mathrm{d}t}=\alpha(\langle c_{\ell-1}\rangle-\langle c_\ell\rangle)+[(\ell-1)\langle r(N)c_{\ell-1}\rangle-\ell\langle r(N)c_\ell\rangle]+\mu(N)\left[(\ell+1)\langle c_{\ell+1}\rangle-\ell\langle c_\ell\rangle\right]. \qquad (4.47)$$

4.7.3 Multi-time-scale dynamics of $N(t)$ and c_k

For simplicity, we discuss the model with no immigration ($\alpha = 0$) first. Under this limit, $N^* = (1 - \frac{\mu}{p})K \sim \mathcal{O}(K)$. The deterministic Eq. (4.4) gives a quite good approximation for the typical dynamics for N in its first phase of evolution: $N(0)$ first approaches its QSS N^* quickly. To estimate this time period, one can integrate $\frac{\mathrm{d}N}{\mathrm{d}t}$ in Eq. (4.11) under $\alpha = 0$ to get $N(t) = \frac{N^* N_0}{N_0 + e^{-(p-\mu)t}(N^* - N_0)}$. Thus N approaches N^* in a characteristic timescale $\mathcal{O}(\frac{1}{p-\mu})$.

As N reaches QSS, $r(N)$ also stabilizes to r^*, $\langle c_k \rangle$ can now approach its QSS, which has a high peak at 0 and a low peak near N^*. Corresponds to the observation that one large-size clone emerges persists for a long time while the other $H - 1$ clones vanish after a while. If we define the number of living clones (or "species richness")

$$R \equiv \sum_{k>0} c_k, \tag{4.48}$$

the above process decreases R from its initial value H to 1 also called a "coarsening" process in physics or fixation in population genetics [Ewe12]. We define the waiting time for such fixation to take place as T_1. Since fixation is most relevant to changes of fractions of clones, we study the problem in a Moran model. In a standard textbook such as [Ewe12], the mean time for the i^{th} clone to fix (conditioned on its fixation) is $T_{\text{fix}}(i) \approx -\frac{N^*}{\mu}\frac{N(0)-n_i(0)}{n_i(0)} \ln\left[1 - \frac{n_i(0)}{N(0)}\right]$. The expected time until any arbitrary clone's fixation is then calculated by averaging each clone's fixation times conditioned on its probability to fix ($P_{\text{fix}}(i) = x_i(0)$) as $T_{\text{c}} = \sum_i T_{\text{fix}}(i) P_{\text{fix}}(i) \approx \frac{N^*}{\mu}$.

The last clone, which is just the total population, is stabilized to N^* by the regulatory effect of $f(N)$. It fluctuates around N^* for an exponential long time. The variance ("width") of such fluctuation near N^* can be calculated by invoking the full stochastic model Eq. (4.3) which leads to the solution $P^*(N)$ in Eq. (4.26). The fact that $P^*(0) \neq 0$ (though is typically exponentially small) dictates that N may initiate a huge fluctuate to the absorbing boundary $N = 0$, resulting in the ultimate extinction of the total population [KS07]. Denote T_2 as the waiting time for the extinction of the total population. Its asymptotic estimate $\mathcal{O}(e^{\mathcal{O}(N^*)})$ shows an extremely large magnitude, which can be obtained from the exact formula $T_2 = \sum_{m=1}^{n} a_m$ where $a_m = \frac{1}{\mu m} + \sum_{j=1}^{\Bbbk - m} \frac{1}{\mu(m+j)} \prod_{i=1}^{j} \frac{r_{m+i-1}}{\mu}$ [DSS05]. Here $\Bbbk = 2N^*$ denotes

a reflecting boundary that does not allow k to go infinity, which in our model is realized by a threshold beyond which c_k is practically 0. Numerically, setting $\mathbb{k} = 2N^*,\ 20N^*$ or $200N^*$ generate practically the same T_2 estimates.

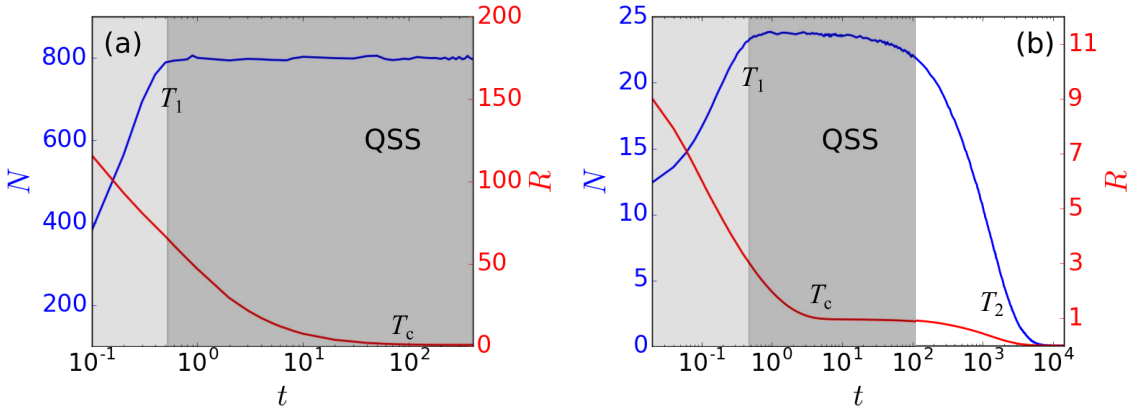

Figure 4.9: Multi-timescale dynamics of N and c_k

Simulations of the multi-timescale dynamics of a small (a) and a large (b) system. Common parameters are $\mu = 10,\ p = 20$. Different parameters are $K = 50,\ H = 11,\ \alpha = 0$ for (a) and $K = 1600,\ H = 200,\ \alpha = 10^{-8}$ for (b).

In Figure 4.9, we plot simulations of the dynamics of both N and R under two different sets of parameters. Values of α are set to be extremely small or 0. Figure 4.9(a) shows that N reaches N^* within 1 unit of time and remains stable for longer than 10^2. Figure 4.9(b) follows a smaller system observed on a longer timescale, where the total population ultimately vanishes as random events accumulate.

So when $\alpha = 0$, we can formally define the QSS to be the time period $T_1 < t < T_2$. When $\alpha > 0$, $N = 0$ is no longer a rigorous absorbing boundary. For extremely small $\alpha \ll (HT_2)^{-1}$, nevertheless, the dynamics is similar to the case $\alpha = 0$ (Figure 4.8(a)) since the waiting time of a single immigration is even longer than the extinction time of the whole population. For large α, cells all clones will persist by fast immigrations and $x = 0$ becomes unstable (Figure 4.8(d)). Intermediate-level immigrations do not stop clones to go extinct, but above regulatory effect may still be considerable such that $x = 0$ is still more stable (Figure 4.8(b-c)).

4.7.4 Moments

The first moment of $\mathbf{c}$ is readily obtained by invoking its definition in Eq. (4.1) as

$$\langle c_k \rangle = \sum_{\mathbf{n}} [I(n_1, k) + I(n_2, k) + ... + I(n_H, k)] P(\mathbf{n}) = H \sum_{\mathbf{n}} I(n_1, k) P(\mathbf{n}) = HP(k)$$

The second moment, when $k \neq \ell$, is obtained as

$$\begin{aligned}
\langle c_k c_l \rangle &= \sum_{\mathbf{n}} [I(n_1, k) + I(n_2, k) + ... + I(n_H, k)][I(n_1, \ell) + I(n_2, \ell) + ... + I(n_H, \ell)] P(\mathbf{n}) \\
&= \sum_{\mathbf{n}} \sum_{i=1}^{H} I(n_i, k)[I(n_1, \ell) + I(n_2, \ell) + ... + I(n_H, \ell)] P(\mathbf{n}) \\
&= \sum_{\mathbf{n}} H \sum_{j \neq 1} I(n_1, k) I(n_j, \ell) P(\mathbf{n}) = H(H-1) \sum_{\mathbf{n}} I(n_1, k) I(n_2, \ell) P(\mathbf{n}) \\
&= H(H-1) P(k, \ell).
\end{aligned}$$

When $k = \ell$, we have

$$\begin{aligned}
\langle c_k c_k \rangle &= \sum_{\mathbf{n}} [I(n_1, k) + I(n_2, k) + ... + I(n_H, k)][I(n_1, k) + I(n_2, k) + ... + I(n_H, k)] P(\mathbf{n}) \\
&= H(H-1) P(k, k) + \sum_{\mathbf{n}} HI(n_1, k) I(n_1, k) P(\mathbf{n}) \\
&= H(H-1) P(k, k) + HP(k).
\end{aligned}$$

The third moment, when $k \neq \ell \neq m$, is obtained as

$$\begin{aligned}
\langle c_k c_\ell c_m \rangle &= \sum_{\mathbf{n}} [I(n_1, k) + ... + I(n_H, k)][I(n_1, \ell) + ... + I(n_H, \ell)][I(n_1, m) + ... + I(n_H, m)] P(\mathbf{n}) \\
&= \sum_{\mathbf{n}} \sum_{i=1}^{H} I(n_i, k)[I(n_1, \ell) + ... + I(n_H, \ell)][I(n_1, m) + ... + I(n_H, m)] P(\mathbf{n}) \\
&= \sum_{\mathbf{n}} HI(n_1, k) \sum_{i \neq 1} I(n_i, \ell) \sum_{j \neq 1, i} I(n_j, m) P(\mathbf{n}) \\
&= H(H-1)(H-2) P(k, \ell, m).
\end{aligned}$$

When $k = \ell \neq m$, we have

$$\begin{aligned}
\langle c_k^2 c_m \rangle &= H(H-1)(H-2) P(k, k, m) + \sum_{\mathbf{n}} [I(n_1, k) + ... + I(n_H, k)][I(n_1, m) + ... + I(n_H, m)] \\
&= H(H-1)(H-2) P(k, k, m) + H(H-1) P(k, m).
\end{aligned}$$

And finally when $k = \ell = m$, we obtain

$$\begin{aligned}\langle c_k^3 \rangle &= H(H-1)(H-2)P(k,k,k) + H(H-1)P(k,k) + \sum_{\mathbf{n}}[I(n_1,k) + ... + I(n_H,k)] \\ &= H(H-1)(H-2)P(k,k,k) + H(H-1)P(k,k) + HP(k).\end{aligned}$$

4.7.5 Diffusion approximation by the Taylor expansion

For simplicity, we replace x_1 with a continuous variables x and neglect the subscript "M" for the Moran model probability P_{M} in the rest of this subsection. Letting $\Delta = \frac{1}{N^*} \to 0$ ($N^* \to \infty$) in Eq. (4.19), we expand the transition rates to the second derivative of x

$$\omega_{12}(x-\Delta)P(x-\Delta) \approx (\omega_{12}P) - \Delta(\omega_{12}P)' + \frac{\Delta^2}{2}(\omega_{12}P)'', \tag{4.49}$$

$$\omega_{21}(x+\Delta)P(x+\Delta) \approx (\omega_{21}P) + \Delta(\omega_{21}P)' + \frac{\Delta^2}{2}(\omega_{21}P)''. \tag{4.50}$$

Substituting them into Eq. (4.19), considering $\omega_{12}(x) = \alpha(1-x) + r^*N^*x(1-x),\ \omega_{21}(x) = \alpha(H-1)x + r^*N^*x(1-x)$ in Eq. (4.20), and canceling out terms, we obtain (when $\alpha H \ll r^*N^*$)

$$\begin{aligned}\mathrm{RHS} &\approx -\Delta[(\omega_{12}-\omega_{21})P]' + \frac{\Delta^2}{2}[(\omega_{12}+\omega_{21})P]'' \\ &= -\frac{\alpha H}{N^*}\frac{\partial}{\partial x}\left(\frac{1}{H} - x\right)P + \frac{1}{2(N^*)^2}\frac{\partial^2}{\partial x^2}[\alpha(1-x) + \alpha(H-1)x + 2r^*N^*x(1-x)]P \\ &\approx -\frac{\alpha H}{N^*}\frac{\partial}{\partial x}\left(\frac{1}{H} - x\right)P + \frac{r^*N^*}{(N^*)^2}\frac{\partial^2}{\partial x^2}x(1-x)P \\ &\approx \mu^*N^*\left[-\frac{1}{N^*}\frac{\partial}{\partial x}m^*\left(\frac{1}{H} - x\right)P_{\mathrm{M}}(x) + \frac{1}{(N^*)^2}\frac{\partial^2}{\partial x^2}x(1-x)P_{\mathrm{M}}(x)\right]\end{aligned} \tag{4.51}$$

where $m^* = \frac{\alpha H}{\mu^* N^*}$ is the fraction of birth that comes from immigration.

For the 2D Moran model, we have

$$\begin{aligned}\frac{\partial P(x_1,x_2)}{\partial t} &= (\omega_{21}P)(x_1+\Delta, x_2-\Delta) + (\omega_{31})P(x_1+\Delta,x_2) + (\omega_{12}P)(x_1-\Delta,x_2+\Delta) \\ &+(\omega_{32}P)(x_1,x_2+\Delta) + (\omega_{13}P)(x_1-\Delta,x_2) + (\omega_{23}P)(x_1,x_2-\Delta) \\ &-[(\omega_{21}+\omega_{31}+\omega_{21}+\omega_{32}+\omega_{13}+\omega_{23})P](x_1,x_2)\end{aligned} \tag{4.52}$$

where

$$\omega_{21} = \alpha x_1 + r^* N^* x_2 x_1, \qquad \omega_{31} = \alpha(H-2)x_1 + r^* N^* x_3 x_1, \tag{4.53}$$

$$\omega_{12} = \alpha x_2 + r^* N^* x_1 x_2, \qquad \omega_{32} = \alpha(H-2)x_2 + r^* N^* x_3 x_2, \tag{4.54}$$

$$\omega_{13} = \alpha x_3 + r^* N^* x_1 x_3, \qquad \omega_{23} = \alpha x_2 + r^* N^* x_3 x_2. \tag{4.55}$$

Invoking the 2D Taylor expansion on Eq. (4.52), we obtain terms like

$$\begin{aligned}(\omega_{21}P)(x_1+\Delta, x_2-\Delta) \approx (\omega_{21}P) + \Delta\left[\frac{\partial(\omega_{21}P)}{\partial x_1} - \frac{\partial(\omega_{21}P)}{\partial x_2}\right] \\ + \frac{\Delta^2}{2}\left[\frac{\partial^2(\omega_{21}P)}{\partial x_1^2} - 2\frac{\partial(\omega_{21}P)}{\partial x_1}\frac{\partial(\omega_{21}P)}{\partial x_2} + \frac{\partial^2(\omega_{21}P)}{\partial x_2^2}\right].\end{aligned}$$

The right-hand side of Eq. (4.52) is thus approximated by

$$\begin{aligned}\text{RHS} \approx{}& \Delta\left[\frac{\partial(\omega_{21}P)}{\partial x_1} - \frac{\partial(\omega_{21}P)}{\partial x_2}\right] + \frac{\Delta^2}{2}\left[\frac{\partial^2(\omega_{21}P)}{\partial x_1^2} - 2\frac{\partial^2(\omega_{21}P)}{\partial x_1 \partial x_2} + \frac{\partial^2(\omega_{21}P)}{\partial x_2^2}\right] \\ &+ \left[\Delta\frac{\partial(\omega_{31}P)}{\partial x_1} + \frac{\Delta^2}{2}\frac{\partial^2(\omega_{31}P)}{\partial x_1^2}\right] + \left[\Delta\frac{\partial(\omega_{32}P)}{\partial x_2} + \frac{\Delta^2}{2}\frac{\partial^2(\omega_{32}P)}{\partial x_2^2}\right] \\ &+ \Delta\left[-\frac{\partial(\omega_{12}P)}{\partial x_1} + \frac{\partial(\omega_{12}P)}{\partial x_2}\right] + \frac{\Delta^2}{2}\left[\frac{\partial^2(\omega_{12}P)}{\partial x_1^2} - 2\frac{\partial^2(\omega_{12}P)}{\partial x_1 \partial x_2} + \frac{\partial^2(\omega_{12}P)}{\partial x_2^2}\right] \\ &+ \left[-\Delta\frac{\partial(\omega_{13}P)}{\partial x_1} + \frac{\Delta^2}{2}\frac{\partial^2(\omega_{13}P)}{\partial x_1^2}\right] + \left[-\Delta\frac{\partial(\omega_{23}P)}{\partial x_2} + \frac{\Delta^2}{2}\frac{\partial^2(\omega_{23}P)}{\partial x_2^2}\right] \\ ={}& \Delta\left[\frac{\partial}{\partial x_1}(\omega_{21}+\omega_{31}-\omega_{12}-\omega_{13})P + \frac{\partial}{\partial x_2}(\omega_{12}+\omega_{32}-\omega_{21}-\omega_{23})P\right] \\ &+ \frac{\Delta^2}{2}\left[\frac{\partial^2}{\partial x_1^2}(\omega_{21}+\omega_{31}+\omega_{12}+\omega_{13})P + \frac{\partial^2}{\partial x_2^2}(\omega_{12}+\omega_{32}+\omega_{21}+\omega_{23})P - 2\frac{\partial^2}{\partial x_1 \partial x_2}(\omega_{12}+\omega_{21})P\right] \\ ={}& \mu^* N^*\left[-\frac{1}{N^*}\sum_{i=1}^{2}\frac{\partial A_i(\mathbf{x})P(\mathbf{x})}{\partial x_i} + \frac{1}{(N^*)^2}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial^2 B_{ij}(\mathbf{x})P(\mathbf{x})}{\partial x_i \partial x_j}\right]\end{aligned} \tag{4.56}$$

where

$$A_i(\mathbf{x}) = \sum_{j=1}^{2} m^*(Q_i - x_i), \quad B_{ii}(\mathbf{x}) = x_i(1-x_i), \quad B_{ij}(\mathbf{x}) = -x_i x_j \ (i \neq j). \tag{4.57}$$

The last step of Eq. (4.56) involves calculations based on Eqs. (4.53-4.55) and the assumption $m^* \ll 1$. For example,

$$\begin{aligned}&\omega_{12} - \omega_{21} + \omega_{13} - \omega_{31} = \alpha(1-x_1) - \alpha H x_1 = \alpha H\left(\frac{1}{H} - x_1\right) \equiv \mu^* N^* \cdot m^*(Q_1 - x_1) \\ &\omega_{21} + \omega_{31} = \alpha(H-1)x_1 + r^* N^* x_1(x_2+x_3) \approx \mu^* N^* \cdot x_1(1-x_1). \\ &\omega_{12} + \omega_{13} = \alpha(1-x_1) + r^* N^* x_1(1-x_1) \approx r^* N^* x_1(1-x_1)\end{aligned} \tag{4.58}$$

CHAPTER 5

Conclusions

In this dissertation, I have described three projects that serve the same overall goal of modeling hematopoietic dynamics from a decade-long clone-tracking dataset, which both carries great scientific values in unraveling single cell-level behaviors of hematopoietic stem and progenitor cells and presents unprecedented technical challenges.

Computationally, I have combined approaches of computer algorithms, statistical analysis, mechanistic modeling, and theory development in order to bridge the gap between the multi-stage stochastic hematopoiesis and the infrequently sampled and noisy downstream data. First, I have developed an empirical rule-based, complexity-optimized, and statistically supported algorithm for improving the quality of the genetically-tagged sequences. The algorithm automatically corrected DNA sequencing errors by grouping similar sequences to an elected "authentic" genome sequence, leading to a 23% data quality boost. The algorithm was integrated as the core component of an in-house data-processing pipeline. Second, to quantitatively understand the clonal dynamics underlying these data, I have built a mathematical framework that successfully maps two principle statistical features of the data, the heterogeneous clone abundances and the highly fluctuating abundances of each individual clone, to two mechanistic features of the hematopoietic system, the short-term HSC self-renewal and the limited proliferation of progenitor cells. Such mapping allows robust inference of the total HSC differentiation rate $100 - 300$ among three rhesus macaques. Third, during the study of size distribution among T cell clones, I noticed an unexpected breakdown of the power law distribution under global carrying capacity (density dependent birth death rates) and small immigration. Further investigation by using global energy landscapes shows that the breakdown is a result of phase transition induced by such density dependence

which leads to the breakdown of the traditional mean-field assumption. I then proposed to transform the system into an equivalent one which was further approximated by a classical Moran model. The results accurately fit the first and second moments of the simulated clone size distributions.

Biologically, first, new predictions can be naturally derived from our computational model and readily tested when larger and more frequently sampled data become available. For example, we predict distinct dynamical patterns between small and large clones (Figure 3.3(b)) that are not caused by sampling noise. Small clones "randomly" contribute to hematopoiesis, while large clones almost "deterministically" appear in peripheral blood. There actually ARE "extinctions and resurrections" of small clones in peripheral blood, which is because of the limited proliferation and long waiting time of HSC differentiation instead of insufficient sampling. Thus, in most of the time, no matter how large the blood sample/how deep the sequencing is, one would not be able to sample small clones. Another prediction is that more frequent sampling can lead to qualitatively different patterns of clonal dynamics as shown by our computer simulations under various sampling frequencies (Figure 3.16). Second, neutral model can explain most of the observed variances (Figure 3.9(b)). As a baseline model, it also helps identify "outliers" (Figure 3.20(a)) that violates the neutral assumption. Such neutrality is the underlying assumption to consider the trajectories (sampled abundances over time) of different clones hypothetically as different realizations of the same clone, on which we can perform modeling and statistical analysis. Third, randomness can generate deterministic heterogeneity in clonal behaviors. The key factor is different timescales. Our model in Chapter 3 has two distinct timescales: the repairing process of bone marrow (BM), and homeostasis. The stochasticity during BM-repairing generates a power-law-like distribution of stem cell clone sizes, which becomes stable in homeostasis. Our theoretical study in Chapter 4, however, predicts that on an even larger time scale (after about thousands of years), the power-law would break down and there will be only one clone left. This prediction was supported by computer simulation but it requires a special animal (whose HSC turnover rates is much faster) to experimentally confirm. This observation also shows that an extremely large-size clone can arise from a neutral model without invoking intrinsic hetero-

geneity. Such large size is beyond the prediction from the usual demographic noise-induced power law distribution, but represents a evolutionary stable state of the system to have one dominating clone. Another biological insight based on such multi-timescale idea has already been mentioned in Section 3.4 is that our model unifies the "clonal stability" hypothesis on primates (extremely slow HSC turnover rate) and "clonal succession" observation on mice (faster HSC turnover rate).

Overall, even though existing data do not yet have the statistical power to accurately resolve the dynamics of each individual clone, a proper combination of feature selection and model simplification makes it possible to extract "signals" from the noisy samples. The collected data indeed contain highly valuable information for understanding the single-cell level behaviors of stem and progenitor cells in a long-term and multi-clonal manner. More data of such type on non-human primates and human have become available and drawn more attentions these days, thanks to the advances in experimental techniques. I hope the obtained results not only help solve the present scientific problems, but may also be useful for studies in relevant areas and interesting for wider audience.

REFERENCES

[AA03] Linda JS Allen and Edward J Allen. "A comparison of three different stochastic population models with regard to persistence time." *Theoretical Population Biology*, **64**(4):439–449, 2003.

[Aal89] Erkki Aalto. "The moran model and validity of the diffusion approximation in population genetics." *Journal of theoretical biology*, **140**(3):317–326, 1989.

[AC09] M Adimy and F Crauste. "Mathematical model of hematopoiesis dynamics with growth factor-dependent apoptosis and proliferation regulations." *Mathematical and Computer Modelling*, **49**:2128–2137, 2009.

[ACG96] Janis L Abkowitz, Sandra N Catlin, and Peter Guttorp. "Evidence that hematopoiesis may be a stochastic process in vivo." *Nature Medicine*, **2**(2):190–197, 1996.

[ACM02] Janis L Abkowitz, Sandra N Catlin, Monica T McCallie, and Peter Guttorp. "Evidence that the number of hematopoietic stem cells per animal is conserved in mammals." *Blood*, **100**(7):2665–2667, 2002.

[All10] LJS Allen. *An Introduction to Stochastic Processes with Applications to Biology.* Taylor and Francis, 2010.

[Ao04] P. Ao. "Potential in stochastic differential equations: novel construction." *J. Phys. A: Math. Gen.*, **37**:L25–L30, 2004.

[Ao09] P. Ao. "Global view of bionetwork dynamics: adaptive landscape." *J. Genet. Genomics*, **36**:63–73, 2009.

[APJ01] S. J. Arnold, M. E. Pfrender, and A. G. Jones. "The adaptive landscape as a conceptual bridge between micro and macroevolution." *Genetica*, **112-113**:9–32, 2001.

[APS95] Janis L Abkowitz, Monica T Persik, Grady H Shelton, Richard L Ott, J Veronika Kiklevich, Sandra N Catlin, and Peter Guttorp. "Behavior of hematopoietic stem cells in a large animal." *Proceedings of the National Academy of Sciences*, **92**(6):2031–2035, 1995.

[BBM03] Samuel Bernard, Jacques Bélair, and Michael C Mackey. "Oscillations in cyclical neutropenia: new evidence based on mathematical modeling." *Journal of Theoretical Biology*, **223**(3):283–298, 2003.

[BBM07] Gareth J Baxter, Richard A Blythe, and Alan J McKane. "Exact solution of the multi-allelic diffusion model." *Mathematical biosciences*, **209**(1):124–170, 2007.

[BKB15] Katrin Busch, Kay Klapproth, Melania Barile, Michael Flossdorf, Tim Holland-Letz, Susan M Schlenner, Michael Reth, Thomas Höfer, and Hans-Reimer Rodewald. "Fundamental properties of unperturbed haematopoiesis from stem cells in vivo." *Nature*, **518**(7540):542–546, 2015.

[BM07] Richard A Blythe and Alan J McKane. "Stochastic models of evolution in genetics, ecology and linguistics." *Journal of Statistical Mechanics: Theory and Experiment*, **2007**(07):P07018, 2007.

[BMH17] Lisa B Boyette, Camila Macedo, Kevin Hadi, Beth D Elinoff, John T Walters, Bala Ramaswami, Geetha Chalasani, Juan M Taboas, Fadi G Lakkis, and Diana M Metes. "Phenotype, function, and differentiation potential of human monocyte subsets." *PloS one*, **12**(4):e0176460, 2017.

[BPS16] Luca Biasco, Danilo Pellin, Serena Scala, Francesca Dionisio, Luca Basso-Ricci, Lorena Leonardelli, Samantha Scaramuzza, Cristina Baricordi, Francesca Ferrua, Maria Pia Cicalese, et al. "In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases." *Cell Stem Cell*, 2016.

[BSI12] Lauren Bragg, Glenn Stone, Michael Imelfort, Philip Hugenholtz, and Gene W Tyson. "Fast, accurate error-correction of amplicon pyrosequences using Acacia." *Nature Methods*, **9**(5):425–426, 2012.

[BVZ12] Leonid V Bystrykh, Evgenia Verovskaya, Erik Zwart, Mathilde Broekhuis, and Gerald de Haan. "Counting stem cells: methodological constraints." *Nature Methods*, **9**(6):567–574, 2012.

[CBE12] Michael R Copley, Philip A Beer, and Connie J Eaves. "Hematopoietic stem cell heterogeneity takes center stage." *Cell Stem Cell*, **10**(6):690–697, 2012.

[CBG11] Sandra N Catlin, Lambert Busque, Rosemary E Gale, Peter Guttorp, and Janis L Abkowitz. "The replication rate of human hematopoietic stem cells in vivo." *Blood*, **117**(17):4460–4466, 2011.

[CG16] T. Chou and C. D. Greenman. "A hierarchical kinetic theory of birth, death and fission in age-structured interacting populations." *Journal of Statistical Physics*, **164**:49–76, 2016.

[CM17] George W. A. Constable and Alan J. McKane. "Mapping of the stochastic Lotka-Volterra model to models of population genetics and game theory." *Phys. Rev. E*, **96**:022416, Aug 2017.

[CN15] Thiparat Chotibut and David R Nelson. "Evolutionary dynamics with fluctuating population sizes and strong mutualism." *Physical Review E*, **92**(2):022718, 2015.

[CN17] Thiparat Chotibut and David R Nelson. "Population Genetics with Fluctuating Population Sizes." *Journal of Statistical Physics*, **167**(3-4):777–791, 2017.

[CPG08] F Crauste, L Pujo-Menjouet, S Génieys, C Molina, and Gandrillon O. "Mathematical model of hematopoiesis dynamics with growth factor-dependent apoptosis and proliferation regulations." *Journal of Theoretical Biology*, **250**:322–338, 2008.

[CQD09] Younan Chen, Shengfang Qin, Yang Ding, Lingling Wei, Jie Zhang, Hongxia Li, Hong Bu, Yanrong Lu, and Jingqiu Cheng. "Reference values of clinical chemistry and hematology parameters in rhesus monkeys (Macaca mulatta)." *Xenotransplantation*, **16**(6):496–501, 2009.

[CRM16] George WA Constable, Tim Rogers, Alan J McKane, and Corina E Tarnita. "Demographic noise can reverse the direction of deterministic selection." *Proceedings of the National Academy of Sciences*, p. 201603693, 2016.

[CVJ16] Jerry L Chen, Fabian F Voigt, Mitra Javadzadeh, Roland Krueppel, and Fritjof Helmchen. "Long-range population dynamics of anatomically defined neocortical networks." *Elife*, **5**, 2016.

[DDH76] J. T. Dancey, K. A. Deubelbeiss, L. A. Harker, and C. A. Finch. "Neutrophil kinetics in man." *Journal of Clinical Investigation*, **58**(3):705, 1976.

[DKC96] Nina J Drize, Jonathan R Keller, and Joseph L Chertkov. "Local clonal analysis of the hematopoietic system shows that multiple small short-living clones maintain life-long hematopoiesis in reconstituted mice." *Blood*, **88**(8):2927–2938, 1996.

[DMW16] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. "Fluctuating fitness shapes the clone-size distribution of immune repertoires." *Proceedings of the National Academy of Sciences*, **113**(2):274–279, 2016.

[DNL12] S Doulatov, F Notta, E Laurenti, and J. E. Dick. "Hematopoiesis: A human perspective." *Cell Stem Cell*, **10**(2):120–136, 2012.

[DP13] Rob J De Boer and Alan S Perelson. "Quantifying T lymphocyte turnover." *Journal of Theoretical Biology*, **327**:45–87, 2013.

[DSS05] Charles R Doering, Khachik V Sargsyan, and Leonard M Sander. "Extinction Times for Birth-Death Processes: Exact Results, Continuum Asymptotics, and the Failure of the Fokker–Planck Approximation." *Multiscale Modeling & Simulation*, **3**(2):283–299, 2005.

[EGC16] Raluca Eftimie, Joseph J Gillard, and Doreen A Cantrell. "Mathematical models for immunology: Current state of the art and future research directions." *Bulletin of mathematical biology*, **78**(10):2091–2134, 2016.

[EIL01] Leah Edelstein-Keshet, Aliza Israel, and Peter Lansdorp. "Modelling perspectives on aging: Can mathematics help us stay young?" *Journal of Theoretical Biology*, **213**(4):509–525, 2001.

[Ewe63] WJ Ewens. "Numerical results and diffusion approximations in a genetic process." *Biometrika*, **50**(3/4):241–249, 1963.

[Ewe12] Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media, 2012.

[FCW43] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. "The relation between the number of species and the number of individuals in a random sample of an animal population." *The Journal of Animal Ecology*, pp. 42–58, 1943.

[Gar85] Crispin W Gardiner. *Handbook of Stochastic Methods: For physics, chemistry, and natural sciences*. Springer, Berlin, 1985.

[GC16] C. D. Greenman and T. Chou. "Kinetic theory of age-structured stochastic birth-death processes." *Physical Review E*, **93**:012112, 2016.

[GKC15] Sidhartha Goyal, Sanggu Kim, Irvin SY Chen, and Tom Chou. "Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques." *BMC Biology*, **13**(1):85, 2015.

[GT05] Antoine Guisan and Wilfried Thuiller. "Predicting species distribution: offering more than simple habitat models." *Ecology letters*, **8**(9):993–1009, 2005.

[HBJ06] Sun-Hee Hong, John Bunge, Sun-Ok Jeon, and Slava S Epstein. "Predicting microbial species richness." *Proceedings of the National Academy of Sciences of the United States of America*, **103**(1):117–122, 2006.

[HBK16] T Höfer, K Busch, K Klapproth, and HR Rodewald. "Fate mapping and quantitation of hematopoiesis in vivo." *Annual Review of Immunology*, **34**(1):449–478, 2016.

[HHM07] Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, and David Mark Welch. "Accuracy and quality of massively parallel DNA pyrosequencing." *Genome biology*, **8**(7):1, 2007.

[HHV12] Martin Hartmann, Charles G Howes, David VanInsberghe, Hang Yu, Dipankar Bachar, Richard Christen, Rolf Henrik Nilsson, Steven J Hallam, and William W Mohn. "Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests." *The ISME journal*, **6**(12):2199–2218, 2012.

[HMJ15] MR Hoyem, F Maloy, P Jakobsen, and BO Brandsdal. "Stem cell regulation: Implications when differentiated cells regulate symmetric stem cell division." *Journal of Theoretical Biology*, **380**:203–219, 2015.

[Hod99] Richard J Hodes. "Telomere length, aging, and somatic cell turnover." *The Journal of Experimental Medicine*, **190**(2):153–156, 1999.

[HR16] T Höfer and HR Rodewald. "Output without input: the lifelong productivity of hematopoietic stem cells." *Current Opinion in Cell Biology*, **43**:69–77, 2016.

[Hub01] Stephen P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press, 2001.

[HWH03] Tom CJ Hill, Kerry A Walsh, James A Harris, and Bruce F Moffett. "Using ecological diversity measures with bacterial communities." *FEMS microbiology ecology*, **43**(1):1–11, 2003.

[JHH13] Qian Jin, Huilin Han, XiMin Hu, XinHai Li, ChaoDong Zhu, Simon YW Ho, Robert D Ward, and Ai-bing Zhang. "Quantifying species diversity with a DNA barcoding-based method: Tibetan moth species (Noctuidae) on the Qinghai-Tibetan Plateau." *PloS one*, **8**(5):e64428, 2013.

[JL90] Craig T Jordan and Ihor R Lemischka. "Clonal and systemic analysis of long-term hematopoiesis in the mouse." *Genes & Development*, **4**(2):220–232, 1990.

[Ken48a] David G Kendall. "On some modes of population growth leading to RA Fisher's logarithmic series distribution." *Biometrika*, **35**(1/2):6–15, 1948.

[Ken48b] David G Kendall. "On the generalized" birth-and-death" process." *The Annals of Mathematical Statistics*, pp. 1–15, 1948.

[KEW17] Samson J Koelle, Diego A Espinoza, Chuanfeng Wu, Jason Xu, Rong Lu, Brian Li, Robert E Donahue, and Cynthia E Dunbar. "Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants." *Blood*, **129**(11):1448–1457, 2017.

[Kim64] M. Kimura. "Diffusion Models in Population Genetics." *J. Appl. Probab.*, **1**:177–232, 1964.

[KKP10] Sanggu Kim, Namshin Kim, Angela P Presson, Dong Sung An, Si Hua Mao, Aylin C Bonifacino, Robert E Donahue, Samson A Chow, and Irvin SY Chen. "High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones." *Journal of Virology*, **84**(22):11771–11780, 2010.

[KKP14] S Kim, N Kim, AP Presson, ME Metzger, AC Bonifacino, M Sehl, SA Chow, GM Crooks, DS Dunbar, CE An, RE Donahue, and IS Chen. "Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study." *Cell Stem Cell*, **14**(4):473–485, 2014.

[KS07] David A Kessler and Nadav M Shnerb. "Extinction rates for fluctuation-induced metastabilities: a real-space WKB approach." *Journal of Statistical Physics*, **127**(5):861–886, 2007.

[LCH16] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-París. "How many TCR clonotypes does a body maintain?" *Journal of theoretical biology*, **389**:214–224, 2016.

[Led12] Roy Lederman. "Homopolymer Length Filters." Technical report, YALEU/DCS/TR1465, 2012.

[Lev66] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals." In *Soviet physics doklady*, volume 10, p. 707, 1966.

[LEZ16] Julio Lahoz-Beneytez, Marjet Elemans, Yan Zhang, Raya Ahmed, Arafa Salam, Michael Block, Christoph Niederalt, Becca Asquith, and Derek Macallan. "Human neutrophil kinetics: modeling of stable isotope labeling data supports short blood neutrophil half-lives." *Blood*, **127**(26):3431–3438, 2016.

[MCT02] CE Muller-Sieburg, RH Cho, M Thoman, B Adkins, and H Sieburg. "Deterministic regulation of hematopoietic stem cell self-renewal and differentiation." *Blood*, **100**(4):1302–1309, 2002.

[MEG07] Brian J McGill, Rampal S Etienne, John S Gray, David Alonso, Marti J Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J Enquist, Jessica L Green, Fangliang He, et al. "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework." *Ecology letters*, **10**(10):995–1015, 2007.

[MF14] Avital Mendelson and Paul S Frenette. "Hematopoietic stem cell niche maintenance during homeostasis and regeneration." *Nature Medicine*, **20**(8):833, 2014.

[MGD06] Joby L McKenzie, Olga I Gan, Monica Doedens, Jean CY Wang, and John E Dick. "Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment." *Nature Immunology*, **7**(11):1225–1233, 2006.

[Mil00] RA Miller. "Telomere diminution as a cause of immune failure in old age: an unfashionable demurral." *Biochemical Society Transactions*, **28**(2):241–245, 2000.

[Mot32] I Motomura. "A statistical treatment of ecological communities." *Zoological Magazine*, **44**:379–383, 1932.

[MS87] G Morchio and F Strocchi. "Mathematical structures for long-range dynamics and symmetry breaking." *Journal of mathematical physics*, **28**(3):622–635, 1987.

[MSB12] Christa E Muller-Sieburg, Hans B Sieburg, Jeff M Bernitz, and Giulio Cattarossi. "Stem cell heterogeneity: implications for aging and regenerative medicine." *Blood*, **119**(17):3900–3907, 2012.

[MSH09] Anna Marciniak-Czochra, Thomas Stiehl, Anthony D Ho, Willi Jäger, and Wolfgang Wagner. "Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation." *Stem Cells and Development*, **18**(3):377–386, 2009.

[MSW09] Anna Marciniak-Czochra, Thomas Stiehl, and Wolfgang Wagner. "Modeling of replicative senescence in hematopoietic development." *Aging (Albany NY)*, **1**(8):723, 2009.

[MTB13] Erica Manesso, José Teles, David Bryder, and Carsten Peterson. "Dynamical modelling of haematopoiesis: an integrated view over the system in homeostasis and under perturbation." *Journal of the Royal Society Interface*, **10**(80):20120817, 2013.

[NW70] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of molecular biology*, **48**(3):443–453, 1970.

[ORK03] Ivar Østby, Leiv S Rusten, Gunnar Kvalheim, and Per Grøttum. "A mathematical model for reconstitution of granulopoiesis after high dose chemotherapy with autologous stem cell transplantation." *Journal of Mathematical Biology*, **47**(2):101–136, 2003.

[Orr09] H. Allen Orr. "Fitness and its role in evolutionary genetics." *Nat. Rev. Genet.*, **10**:531–539, 2009.

[PBV10] Janesh Pillay, Ineke den Braber, Nienke Vrisekoop, Lydia M Kwast, Rob J de Boer, José AM Borghans, Kiki Tesselaar, and Leo Koenderman. "In vivo labeling with ^2H_2O reveals a human neutrophil lifespan of 5.4 days." *Blood*, **116**(4):625–627, 2010.

[PJ96] Jean Peccoud and Christine Jacob. "Theoretical uncertainty of measurements using quantitative polymerase chain reaction." *Biophysical Journal*, **71**(1):101–108, 1996.

[PPB96] Josef T Prchal, Jaroslav F Prchal, Monika Belickova, Shande Chen, Yongli Guan, G Larry Gartland, and Max D Cooper. "Clonal stability of blood cell lineages indicated by X-chromosomal transcriptional polymorphism." *Journal of Experimental Medicine*, **183**(2):561–567, 1996.

[PQP08] Todd L Parsons, Christopher Quince, and Joshua B Plotkin. "Absorption and fixation times for neutral and quasi-neutral populations with density dependence." *Theoretical Population Biology*, **74**(4):302–310, 2008.

[PZF17] Amit A Patel, Yan Zhang, James N Fullerton, Lies Boelen, Anthony Rongvaux, Alexander A Maini, Venetia Bigley, Richard A Flavell, Derek W Gilroy, Becca Asquith, et al. "The fate and lifespan of human monocyte subsets in steady state and systemic inflammation." *Journal of Experimental Medicine*, **214**(7):1913–1923, 2017.

[Qia06] H. Qian. "Open-system nonequilibrium steady state: statistical thermodynamics, fluctuations, and chemical oscillations." *J. Phys. Chem. B*, **110**:15063–15074, 2006.

[QLC14] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. "Diversity and clonal selection in the human T-cell repertoire." *Proceedings of the National Academy of Sciences*, **111**(36):13139–13144, 2014.

[QLD11] Christopher Quince, Anders Lanzen, Russell J Davenport, and Peter J Turnbaugh. "Removing noise from pyrosequenced amplicons." *BMC bioinformatics*, **12**(1):1, 2011.

[RBK99] Nathalie Rufer, Tim H Brümmendorf, Steen Kolvraa, Claus Bischoff, Kaare Christensen, Louis Wadsworth, Michael Schulzer, and Peter M Lansdorp. "Telomere fluorescence measurements in granulocytes and T lymphocyte subsets point to a high turnover of hematopoietic stem cells and memory T cells in early childhood." *The Journal of Experimental Medicine*, **190**(2):157–168, 1999.

[SBM14] T Székely, K Burrage, M Mangel, and MB Bonsall. "Stochastic dynamics of interacting haematopoietic stem cell niche lineages." *PLoS Computational Biology*, **10**:e1003794, 2014.

[Sel74] Peter H Sellers. "On the theory and computation of evolutionary distances." *SIAM Journal on Applied Mathematics*, **26**(4):787–793, 1974.

[She97] D. Sherrington. "Landscape paradigms in physics and biology: introduction and overview." *Physica D*, **107**:117–121, 1997.

[SHM14] T Stiehl, AD Ho, and A Marciniak-Czochra. "The impact of CD34+ cell dose on engraftment after SCTs: personalized estimates based on mathematical modeling." *Bone marrow transplantation*, **49**(1):30, 2014.

[SK12] Z Sun and NL Komarova. "Stochastic modeling of stem-cell dynamics with control." *Mathematical Biosciences*, **240**:231–240, 2012.

[SKL07] Bryan E Shepherd, Hans-Peter Kiem, Peter M Lansdorp, Cynthia E Dunbar, Geraldine Aubert, Andre LaRochelle, Ruth Seggewiss, Peter Guttorp, and Janis L Abkowitz. "Hematopoietic stem-cell behavior in nonhuman primates." *Blood*, **110**(6):1806–1813, 2007.

[SLJ12] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. "The long-range interaction landscape of gene promoters." *Nature*, **489**(7414):109, 2012.

[SM11] T Stiehl and A Marciniak-Czochra. "Characterization of stem cells using mathematical models of multistage cell lineages." *Mathematical and Computer Modelling*, **53**:1505–1517, 2011.

[SRC14] Jianlong Sun, Azucena Ramos, Brad Chapman, Jonathan B Johnnidis, Linda Le, Yu-Jui Ho, Allon Klein, Oliver Hofmann, and Fernando D Camargo. "Clonal dynamics of native haematopoiesis." *Nature*, **514**(7522):322–327, 2014.

[SRM11] Hans B Sieburg, Betsy D Rezner, and Christa E Muller-Sieburg. "Predicting clonal self-renewal and extinction of hematopoietic stem cells." *Proceedings of the National Academy of Sciences*, **108**(11):4370–4375, 2011.

[SW81] Temple F Smith and Michael S Waterman. "Identification of common molecular subsequences." *Journal of molecular biology*, **147**(1):195–197, 1981.

[SW10a] J. Seita and I. L. Weissman. "Hematopoietic stem cell: self-renewal versus differentiation." *Systems Biology and Medicine*, **2**(6):640–653, 2010.

[SW10b] Jun Seita and Irving L Weissman. "Hematopoietic stem cell: self-renewal versus differentiation." *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **2**(6):640–653, 2010.

[SXX17] Gajendra W Suryawanshi, Song Xu, Yiming Xie, Tom Chou, Namshin Kim, Irvin SY Chen, and Sanggu Kim. "Bidirectional Retroviral Integration Site PCR Methodology and Quantitative Data Analysis Workflow." *Journal of visualized experiments: JoVE*, p. e55812, 2017.

[Tak12] Hajime Takayama. *Cooperative Dynamics in Complex Physical Systems: Proceedings of the Second Yukawa International Symposium, Kyoto, Japan, August 24–27, 1988*, volume 43. Springer Science & Business Media, 2012.

[TEW10] Andreas Trumpp, Marieke Essers, and Anne Wilson. "Awakening dormant haematopoietic stem cells." *Nature Reviews. Immunology*, **10**(3):201, 2010.

[VBH05] Igor Volkov, Jayanth R Banavar, Fangliang He, Stephen P Hubbell, and Amos Maritan. "Density dependence explains tree species abundance and diversity in tropical forests." *Nature*, **438**(7068):658, 2005.

[VBZ13] Evgenia Verovskaya, Mathilde JC Broekhuis, Erik Zwart, Martha Ritsema, Ronald van Os, Gerald de Haan, and Leonid V Bystrykh. "Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding." *Blood*, **122**(4):523–532, 2013.

[Wad57] C. H. Waddington. *The Strategy of the Genes: A Discussion of Some Aspect of Theoretical Biology*. New York: The MacMillan Company, 1957.

[Wan05] M Wang. *Nonhomogeneous Birth-death Processes, M. S. Thesis: California State Polytechnic University, Pomona*. M. S. Thesis: California State Polytechnic University, Pomona, 2005.

[WHL14] Adrianto Wirawan, Robert S Harris, Yongchao Liu, Bertil Schmidt, and Jan Schröder. "HECTOR: a parallel multistage homopolymer spectrum based error corrector for 454 sequencing data." *BMC bioinformatics*, **15**(1):1, 2014.

[WLO08] Anne Wilson, Elisa Laurenti, Gabriela Oser, Richard C van der Wath, William Blanco-Bose, Maike Jaworski, Sandra Offner, Cyrille F Dunant, Leonid Eshkind, Ernesto Bockamp, et al. "Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair." *Cell*, **135**(6):1118–1129, 2008.

[WLT09] Anne Wilson, Elisa Laurenti, and Andreas Trumpp. "Balancing dormant and self-renewing hematopoietic stem cells." *Current Opinion in Genetics & Development*, **19**(5):461–468, 2009.

[Wri32] S. Wright. "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution." *Proc. Sixth Int. Cong. Genet.*, **1**:356–366, 1932.

[WXW08] J. Wang, L. Xu, and E. Wang. "Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations." *Proc. Natl. Acad. Sci. USA*, **105**:12271–12276, 2008.

[XJJ14] Song Xu, Shuyun Jiao, Pengyao Jiang, and Ping Ao. "Two-time-scale population evolution on a singular landscape." *Physical Review E*, **89**(1):012724, 2014.

[XKG16] Jason Xu, Samson Koelle, Peter Guttorp, Chuanfeng Wu, Cynthia E Dunbar, Janis L Abkowitz, and Vladimir N Minin. "Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis." *arXiv preprint arXiv:1610.07550*, 2016.

[YSK15] J Yang, Z Sun, and NL Komarova. "Analysis of stochastic stem cell models with control." *Mathematical Biosciences*, **266**:93–107, 2015.

[ZES13] Veronika I Zarnitsyna, Brian D Evavold, Louis N Schoettle, Joseph N Blattman, and Rustom Antia. "Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire." *Frontiers in immunology*, **4**, 2013.

[ZLM12] Changjing Zhuge, Jinzhi Lei, and Michael C Mackey. "Neutrophil dynamics in response to chemotherapy and G-CSF." *Journal of Theoretical Biology*, **293**:111–120, 2012.