

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Quantifying Dynamics of Microbial Evolution for Nascent Beneficial Mutations in Barcode Lineage Tracking Experiment

Permalink

<https://escholarship.org/uc/item/2p4196mr>

Author

Kuo, Huan-Yu

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Quantifying Dynamics of Microbial Evolution
for Nascent Beneficial Mutations in Barcode Lineage Tracking Experiment

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Physics with a Specialization in Quantitative Biology

by

Huan-Yu Kuo

Committee in charge:

Sergey Kryazhimskiy, Chair
Terence Hwa, Co-Chair
Vineet Bafna
Elena Koslover
Wouter-Jan Rappel

2023

Copyright

Huan-Yu Kuo, 2023

All rights reserved.

The Dissertation of Huan-Yu Kuo is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my beloved husband, Yi-Ping Wang, and my cherished cat, Angie.

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	xiii
Acknowledgements	xiv
Vita	xvi
Abstract of the Dissertation	xvii
Introduction	1
Chapter 1 Background	4
1.1 Introduction to Biological Fitness	4
1.1.1 Absolute fitness, relative fitness and fitness effect in discrete time	4
1.1.2 Absolute fitness and fitness effect in continuous time	6
1.1.3 Lineage trajectory and the mean fitness effect	11
1.1.4 Distribution of fitness effect (DFE)	13
1.2 Introduction to the Barcode Lineage Tracking Experiment	14
1.2.1 DNA barcode	15
1.2.2 Evolution experiment	15
1.2.3 Barcode measurement	16
1.3 Challenges, Objectives, and Approaches	16
1.3.1 Challenges to evolutionary dynamics analysis	17
1.3.2 Objectives and basic approaches of this dissertation	18
Chapter 2 Shortcomings of the Levy-Blundell Method	19
2.1 Introduction	19
2.2 Review of the Levy-Blundell method	19
2.3 Critical assumptions of the Levy-Blundell method	22
2.3.1 The “neutral decline” assumption	22
2.3.2 The no lineage extinction assumption	24
2.3.3 Probing the validity of the Levy-Blundell model assumptions with BLT simulations	25
2.4 Results	26
2.4.1 Neutral decline assumption is violated in both experimental and simulation data	26

2.4.2	Performance of the Levy-Blundell method is sensitive to the choice of reference lineages	28
2.4.3	Failure of lineage identification in the presence of strong genetic drift . .	33
2.5	Discussion	35
2.5.1	The working region for the mean-fitness estimation changes over time . .	35
2.5.2	An intuition for an offset in expected read count	37
2.6	Methods	39
2.6.1	BLT simulation	39
Chapter 3	Bayesian Filtering Method for Estimating Fitness of Nascent Adaptive Mutants in BLT Experiments	40
3.1	Introduction	40
3.2	The Bayesian filtering method	41
3.2.1	Overview of the Bayesian filtering method	41
3.2.2	Stochastic models for updating lineage dynamics	42
3.2.3	Estimating the mean fitness of the population	51
3.2.4	Identifying adapted lineages and estimating their fitness effects	53
3.3	Results	53
3.3.1	Receiver Operating Characteristic (ROC) curve for the Bayesian Filtering method	53
3.3.2	Accuracy of fitness estimates of the Bayesian Filtering method	56
3.3.3	Application to existing BLT data	62
3.4	Discussion	63
3.5	Materials and Methods	65
3.5.1	Experimental Estimation of Measurement Noise	65
3.5.2	Bayesian Filtering Algorithm	70
3.5.3	Simulation Method (Bayesian)	71
Chapter 4	A Method for Estimating the Arrival Rate for a Class of Adaptive Mutations	74
4.1	Introduction	74
4.2	Model of stochastic dynamics of a nascent mutant in an adapting population . .	76
4.2.1	Theory	76
4.2.2	Approximation for the analytical solution	80
4.3	Likelihood function of mutation rate	84
4.4	Results and Discussions	85
4.4.1	The validity of strong- and weak-mutation approximations	85
4.4.2	Rates of adaptive mutation in yeast	89
4.5	Discussion	91
Chapter 5	Estimating the Mutation Rate Spectrum from Lineage Tracking Data	96
5.1	Introduction	96
5.2	Estimating the mutation rate spectrum	100
5.3	Results	102

5.3.1	Analytical approach to the Wright-Fisher model for mutant population expanding in a background of increasing mean-fitness	102
5.3.2	The establishment probability in the Wright-Fisher model	105
5.3.3	Empirical calculation on the detection probability of established mutants	108
5.3.4	Inference of the mutational rate spectrum	112
5.4	Discussion	114
Appendix A	Barcode Frequency Dynamics	120
Appendix B	Prediction of mutant population dynamics in Wright-Fisher model	123
Bibliography		126

LIST OF FIGURES

Figure 1.1.	Schematic figure of pairwise competition assay experiment for measuring fitness (from Ref. [19])	8
Figure 1.2.	Graphic representation of typical bacterial growth curve in culture medium [104]	10
Figure 1.3.	Example of pairwise competition between parasite <i>P. falciparum</i> wildtype (NHP4026, red) and an artemisinin-sensitive mutation in <i>kelch13</i> (P3, green) in the flask (from Ref. [98]).	10
Figure 1.4.	The dynamics of sequence evolution in BYB1-G07 (from Ref. [57]).	11
Figure 1.5.	The distribution of fitness effects of random mutations in vesicular stomatitis virus (from Ref. [22])	13
Figure 1.6.	Overviwe of barcode lineage tracking (BLT) experiment	15
Figure 2.1.	Example of fitting the noise parameter κ and the mean fitness $\bar{s}_k$ (denoted as $\bar{x}$ in figure) from a histogram distribution of $P(r_k)$ constructed for barcodes with $r_{k-1} = 30$. Figure is obtained from Supplementary Figure 19 in Ref. [58].	21
Figure 2.2.	Baseline scenario simulation. (A) barcode frequency trajectory (B) DFE (C) mean-fitness trajectory	25
Figure 2.3.	Large-drift scenario simulation. (A) barcode frequency trajectory (B) DFE (C) mean-fitness trajectory	26
Figure 2.4.	Failure of the neutral decline assumption (black dash line) for experimental data (A) and simulated data (B).	27
Figure 2.5.	Read distribution of large drift scenario for lineages of read 81 – 100 and the inference of mean-fitness and noise based on Levy-Blundell method.	29
Figure 2.6.	Read distribution of baseline scenario for lineages of read 21 – 40 and the inference of mean-fitness and noise based on Levy-Blundell method.	30
Figure 2.7.	Performance of the Levy-Blundell method on simulated data in the baseline scenario, with reference lineages chosen to have read counts between 81 and 100.	31
Figure 2.8.	Lineage identification and fitness estimation by Levy et al. with reference chosen from 21 – 40 read count for the baseline simulation case.	32

Figure 2.9.	fraction of barcodes which are present in the previous time but have zero read count at the next time point for (black) baseline and (blue) strong drift scenario.	33
Figure 2.10.	Lineage identification and fitness estimation by Levy et al. with reference chosen from 81 – 100 read count for the large-drift simulation case.	34
Figure 2.11.	Read distribution of large drift scenario for lineages of read 81 – 100 and the inference of mean-fitness and noise based on Levy-Blundell method. .	36
Figure 2.12.	The working region for the unbiased inference of mean fitness shifts over time. Data from the baseline simulation scenario.	37
Figure 3.1.	S-model.	45
Figure 3.2.	N-model.	47
Figure 3.3.	Measured barcode frequency in the sequencing data (y axis) on average faithfully represents the cell lineage frequency in the population (x axis). Each point represents a unique barcode, blue line shows the 1-to-1 line. ..	49
Figure 3.4.	Variance in the reads count versus the mean read count. Same data as in Figure 3.3. Each point represents the data from barcodes of same initial frequency.	50
Figure 3.5.	Simulation of a BLT experiment under strong selection, strong genetic drift and small coverage for assessing the performance of the Bayesian Filtering method.	54
Figure 3.6.	Receiver operating characteristic (ROC) curve for the identification of adapted lineages for $K = 1$ (blue curve), $K = 2$ (black curve) and $K = 3$ (red curve).	55
Figure 3.7.	Accuracy of the Bayesian filtering method on simulated data.	57
Figure 3.8.	Accuracy of the Levy-Blundell method on simulated data. Notations are the same as in Figure 3.7.	58
Figure 3.9.	Selection coefficient $s(1/gen)$ to the inverse of maximum size at the bottleneck $1/f_{max}/N_b$ for true positive lineages (red circles) and false negative lineages (blue circles) with a one-to-one line (black dash).	59
Figure 3.10.	Lineage trajectories for (top left) 1196 true positive lineages, colored by the error of s estimates. Lineage trajectories for (bottom left) 640 false negative, (top right) 14 false positive and (bottom right) 98150 true negative lineages, colored by initial barcode read counts.	60

Figure 3.11.	Application of the Bayesian filtering method to the yeast BLT data in yeast published in Ref. [58].	61
Figure 4.1.	Strong and weak mutant approximations for function b_1 for a mutant with $s = 0.8$	86
Figure 4.2.	Strong and weak mutant approximations for function b_1 for a mutant with $s = 0.4$. Notations are the same as in Figure 4.1.	87
Figure 4.3.	Theory with the strong-mutation approximation captures the distribution of establishment times and survival probability of a nascent adaptive mutant	88
Figure 4.4.	Distribution of adaptive mutations across the most common driver loci. Driver loci with 5 or more mutations are shown. Figur is reproduced from Ref. [102].	89
Figure 4.5.	Estimation of the rate of beneficial mutations. The log-likelihood function for the estimation of the chrIII-3n beneficial mutation rate is shown as an example. Red point indicates the identified maximum value. Horizontal red bar shows the standard error calculated based on Fisher Information.	91
Figure 4.6.	Probabilities of observing adaptive mutations at the most common driver loci in the whole-genome sequencing data.	92
Figure 5.1.	Schematic plot for the observed DFE resulting from the arrival mutants, probability of mutant establishment $P_{est}(s)$ and probability of detection $P_{det}(s)$	100
Figure 5.2.	A population trajectory of mutant ($s = 5\%$, arising at $t_0 = 0$ generation) over simulating time (black line, y-axis: log population size, x-axis: time in generation), fitted by theoretical trajectory (red dash line) with best-fit of establishment time $\tau = 21.1$ generation.	103
Figure 5.3.	Survival probability for a mutant of $s = 5\%$ arising at $t_0 = 0$ in the Wright-Fisher model (blue circle) over time, compared to the theoretical prediction (black line).	104
Figure 5.4.	Survival probability at time $T = 160$ generation for a mutant arising at $T = 0$ in the Wright-Fisher simulation (blue circle), theoretical long-term stable survival probability with competitive background (red dot line), and theoretical survival probability at 160 generations (black dashed line).	105
Figure 5.5.	Probability distribution of established time for a mutant of $s = 5\%$ arising at $t_0 = 0$ in the Wright-Fisher model (blue histogram), compared to the theoretical distribution (black line) derived from the branching process.	106

Figure 5.6.	Establishment probability for mutants that randomly arise during the simulating time in the Wright-Fisher model (blue circle), compared to the theory (cyan dash line).	107
Figure 5.7.	Distribution of fitness effect (DFE) for the established mutants that randomly arise during the simulating time in the Wright-Fisher model (blue circle), compared to the theory (cyan circle).	108
Figure 5.8.	Arising beneficial mutations $NT\mu(s)$ from input function (black circle line) and simulation sampling results (blue circle line) with population size $N = 2 \times 10^6$, total time $T = 160$	109
Figure 5.9.	Schematic Plots (A) Obtaining the inferred mean-fitness trajectory and observed DFE (B) Estimating the probability $P_{det}(s)$ of detecting established mutation of s	110
Figure 5.10.	Probability of detecting established mutants with fitness effect s by empirical computation (black circle), fit by logistic function (black dash line) $P_{det}(s) = (0.840 - 0.09e^{-109(s-0.0235)}) / (1 + e^{-109(s-0.0235)})$	111
Figure 5.11.	Mean fitness trajectory for the Wright-Fisher simulation (black dash), Bayesian filtering program inference (green circle), and logistic function fit (green dotted-line) to the Bayesian inference.	112
Figure 5.12.	Observed distribution of fitness effect (DFE) by applying the Bayesian filtering program to the Wright-Fisher simulation data.	113
Figure 5.13.	Mutation rate spectrum: theory (black line) and the inference (red circle line) by back-calculating the observed DFE.	114
Figure 5.14.	Mutation rate spectrum: theory (black line) and the inference (red circle line) by back-calculating the shifted, observed DFE.	115
Figure 5.15.	Observed DFE (blue circle) and its backward-calculating established DFE (black circle)	116
Figure 5.16.	Established DFE in simulation (black square) and its forward-calculating observed DFE (blue square)	116
Figure 5.17.	Inaccurate estimated fitness effect by Bayesian filtering program compared to the simulation value	119
Figure 5.18.	The observed DFE (blue histogram) and the observed DFE left-shifted by 2.8% (orange histogram).	119

Figure B.1.	Wright-Fisher simulation (A) trajectory of mutant population (B) distribution of establishment time (C) survival probability over time (D) survival probability at the end time to the fitness effect.	124
-------------	---	-----

LIST OF TABLES

Table 3.1.	Constant parameters used in the Bayesian filtering model	42
Table 3.2.	Notation of probabilities and variables used in the Bayesian filtering model	43
Table 3.3.	Putative lineage class and initial belief distribution	52
Table 3.4.	Initial clone frequencies in the mixture culture	66
Table 3.5.	Total barcode reads for 10 replicated experiments	69
Table 4.1.	Solution of branching process with zero mean fitness (second column, all formula obtained in Ref. [17]) and with non-zero mean-fitness.	83
Table 4.2.	Numbers of isolates carrying the most common adaptive mutations and the estimated mutation rates.	95
Table 5.1.	The long-term survival probability for a mutant of s , arising at time $t = t_0$ and expanding in a zero mean-fitness ($\bar{s} = 0$) or an increasing mean-fitness background, compared between branching process and the Wright-Fisher model.	104

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Sergey Kryazhimskiy, for his invaluable guidance, exceptional mentorship, and continuous encouragement throughout my time as a PhD student. His support has been instrumental in shaping my overall development as a researcher. The regular meetings, insightful conversations, and constructive feedback provided by Dr. Kryazhimskiy have not only deepened my scientific understanding but also inspired me to bridge the disciplines of physics and biology in my research endeavors. Without his guidance, this research would not have been possible, and I am truly thankful for his contributions.

I would also like to extend my heartfelt gratitude to my esteemed co-chair, Terence Hwa, and my esteemed committee members, Elena Koslover, Vineet Bafna, and Wouter-Jan Rappel, for their invaluable guidance and unwavering support. I am profoundly thankful for their significant contributions in driving the progress of Chapter 5 during the last few months of my PhD journey.

I am deeply thankful to Sandeep Venkataram for his inspiring discussions and generous provision of the experimental data utilized in Chapter 4. His invaluable assistance over the years has played a pivotal role in my personal and professional growth, particularly in acquiring hands-on experience in the wet lab.

Furthermore, I would like to express my appreciation for the contributions of the members of the Kryazhimskiy lab. Their valuable insights, constructive suggestions, and mental supports have enriched my research adventure. I extend special thanks to Alena Martsul, Elissa Heredia, Michael Overton, Minkyu Kim, Sandeep Venkataram, Sarah Ardell, Sean Guy, and Shohreh H. Sikaroodi for their exceptional contributions.

Collectively, the guidance, support, and inspiration provided by my esteemed committee members, namely Sergey Kryazhimskiy, Terence Hwa, Elena Koslover, Vineet Bafna, Wouter-Jan Rappel, as well as all the exceptional members of the Kryazhimskiy lab, have played a fundamental role in the resounding success of this research endeavor. I am sincerely grateful.

Chapter 2, in full is currently being prepared for submission for publication of the material. Huan-Yu, Kuo; Sergey, Kryazhimskiy. The dissertation author was the primary researcher and author of this material.

Chapter 3, in full is currently being prepared for submission for publication of the material. Huan-Yu, Kuo; Sergey, Kryazhimskiy. The dissertation author was also the primary researcher and author of this material.

Chapter 4, in its entirety, has been published in *Nature Ecology & Evolution* (2023): 1-12 (Ref. [102]) in the title of “Mutualism-enhancing mutations dominate early adaptation in a two-species microbial community.” Sandeep, Venkataram; Huan-Yu, Kuo; Erik F. Y., Hom; Sergey, Kryazhimskiy. The dissertation author was the researcher and author of this publication.

Chapter 5 is a work in progress that contains unpublished material coauthored with Sergey, Kryazhimskiy. The dissertation author was the primary author of this chapter.

VITA

- 2013 Bachelor of Science in Physics, National Tsing Hua University, Taiwan
- 2014 Master of Science in Physics, National Tsing Hua University, Taiwan
- 2015 Research Assistant, Taipei Veterans General Hospital & National Central University, Taiwan
- 2023 Doctor of Philosophy in Physics with a Specialization in Quantitative Biology, University of California San Diego, U.S.A.

ABSTRACT OF THE DISSERTATION

Quantifying Dynamics of Microbial Evolution
for Nascent Beneficial Mutations in Barcode Lineage Tracking Experiment

by

Huan-Yu Kuo

Doctor of Philosophy in Physics with a Specialization in Quantitative Biology

University of California San Diego, 2023

Sergey Kryazhimskiy, Chair
Terence Hwa, Co-Chair

Quantifying the evolutionary dynamics of beneficial mutations in microbial populations relies on a massive sampling of nascent mutations, which has recently become possible by innovative barcode techniques. Despite barcode lineage tracking (BLT) experiments offer a straightforward means of capturing mutation frequency trajectories, accurately quantifying their dynamics remains a challenge. This dissertation addresses the critical issue of precisely and accurately estimating the fitness effects, mutation rates, and the distribution of fitness effects (DFE) associated with beneficial mutations identified in BLT experiments. First, I investigate the limitations of existing methods used to infer fitness effects and identify potential

reasons for failure. To address this issue, I develop a novel Bayesian filtering method for fitness inference, which significantly enhances the accuracy of estimation compared to current approaches. Secondly, to estimate the rate of beneficial mutations within specific classes, I develop a stochastic model of mutant population dynamics. By applying this model, I infer the rate of beneficial mutations in the BLT experiment, providing valuable insights into mutation dynamics. Lastly, I highlight an inherent bias in our observations of beneficial mutations through adaptive evolution relative to the true underlying DFE. To address this disparity, I propose a method for inferring the mutation rate spectrum by back-calculating the observed DFE. The work presented in this dissertation establishes a statistically solid foundation for the analysis of the BLT experiment.

Introduction

Every new adaptation originates in a single individual. Most vanish soon thereafter. The few that survive may encounter fierce competition with other contending adapted lineages [4, 5, 7, 17, 29, 33, 34, 42, 46, 55, 57, 58, 69, 73, 89]. This competition, termed “clonal interference”, is particularly severe in organisms with limited recombination that have access to a large supply of beneficial mutations, such as microbes and cancers evolving in the presence of drugs [68, 82, 105] or viruses evading the host’s immune system [41, 96]. In the clonal interference regime, also known as the “concurrent mutation” regime [17], the chance of fixation of an allele depends not only on its own fitness benefit but also on the benefits provided by competing alleles [29, 34, 90]. Thus, to understand the dynamics of rapid adaptation, we must know the statistics of fitness effects of adaptive mutations, in particular the distribution of fitness effect (DFE). However, our knowledge of DFEs is extremely limited, even in well studied laboratory systems.

The main challenge in measuring the DFE is that it is difficult to collect large and unbiased samples of beneficial mutations. The most statistically sound strategy would be to isolate them from random samples of spontaneous mutations. However, doing so is prohibitively inefficient because in most organisms mutations are rare events [92] and the vast majority of them are neutral or deleterious [22]. The fraction of beneficial mutations could be as low as $\sim 2\%$, with stronger beneficial mutations being even more rare [58]. As a result, obtaining accurate statistics of beneficial mutations from samples of random mutations would be extremely burdensome. Instead, researchers leverage the power of natural selection, which over time elevates the frequencies of beneficial mutations in the population. Indeed, it is relatively easy

to identify and isolate beneficial mutations in recently adapted populations [13, 25, 48, 53]. However, this strategy preferentially captures mutations with the strongest fitness benefits because they are more likely to escape genetic drift when rare and survive clonal interference. As such, the information about the diversity of adaptive mutations gained with this approach is not only incomplete but also biased in a way that depends on idiosyncratic and often unknown details of the evolutionary dynamics in the population from which the mutations are sampled.

A recently developed barcode lineage tracking (BLT) method provides an alternative experimental strategy for learning the statistics of beneficial mutations in asexual microbes [58, 101]. The key idea is to introduce many (typically, $\sim 10^5$) random neutral DNA barcodes into an otherwise isogenic population and track barcode frequencies over time using deep sequencing. This approach allows one to measure the fitness benefits of many simultaneously segregating driver mutations, including weak ones, in a single short evolution experiment, while avoiding laborious genetic reconstructions and screens. As a result, measuring DFE of beneficial mutations becomes accessible experimentally.

While the BLT approach shows great potential, its success is currently limited by several challenging statistical problems. How do we identify the few adapted lineages (those that have already acquired a beneficial mutation) among many thousands of neutral ones (those that have not yet acquired any beneficial mutations) based on barcode frequency dynamics? How do we estimate the fitness advantages for those adapted lineages from frequency dynamics? How do we estimate the rate at which different adaptive mutations arise? My dissertation focuses on the development of analytical methods for quantifying the evolutionary dynamics for the mutations in BLT experiments and answering these important applied questions.

In chapter 1, I provide some essential background information that is necessary to understand the rest of the work. In particular, I describe a typical setup for BLT experiments, define biological fitness in the context of these experiments, and more precisely outline the problems to solve in this dissertation.

In chapter 2, I demonstrate that the key assumptions made in current BLT analysis fail

under certain scenarios, which can lead to inaccurate identification of adapted lineages and their fitness.

In chapter 3, I develop a novel Bayesian framework to solve problems addressed in the previous chapter, and implement this method on both simulation and experimental studies.

In chapter 4, I model the population dynamics of mutant growth in a competitive background. I apply this model to estimate the rate of beneficial mutation in the same class for BLT experiment.

In chapter 5, I develop a method to infer the mutation rate spectrum of fitness effect as the unbiased underlying evolutionary dynamics of beneficial mutations.

Chapter 1

Background

1.1 Introduction to Biological Fitness

The ability of an organism to survive and reproduce is crucial for its long-term evolutionary success, and it is measured by a quantity called fitness. Conceptually, fitness is the reproductive success of a genotype in its habitat. The precise definitions of fitness vary across systems. The two most common definitions are the “Wrightian fitness” in models with discrete time and the “Malthusian fitness” in models with continuous time, which are defined below. In these standard models of evolution, the change of a genotype’s abundance in the population depends on its absolute frequency in the population and on the absolute frequencies of other genotypes. It is therefore more convenient to use relative fitness than absolute fitness and, correspondingly, of the fitness effect of a mutation. In this chapter, I introduce the definitions of biological fitness, following Refs. [15, 28, 77, 108], and briefly review the method of fitness inference from microbial experiments, following Refs. [12, 19, 98].

1.1.1 Absolute fitness, relative fitness and fitness effect in discrete time

In models with discrete time, absolute fitness is the average per capita lifetime contribution of individuals of that genotype to the population after one generation. It can be considered as the average number of offspring born in one generation. The simplest reproduction model is the Wright-Fisher where all parents are entirely replaced by their offspring every generation.

In this model, assuming that all individuals in the population have the same genotype, the Wrightian absolute fitness W_w (equation (1.1)) is simply the ratio of population size n between two generations k and $k - 1$:

$$W_w = n_k / n_{k-1}, \quad (1.1)$$

and its population size (equation (1.2)) would increase in power law if $W_w > 1$, or decay in power law for $W_w < 1$ from a initial size n_0

$$n_k = W_w n_{k-1} = (W_w)^k n_0. \quad (1.2)$$

The emergence of random mutations in a population creates genetic variations such that the ancestor and those genotypes arising by mutation would compete with each other. Due to the competition, the presence of a genotype in the population would increase or decrease depending on its absolute fitness and the absolute fitness of other genotypes. Because of the nature of competition, one cannot simply realize a genotype's fraction from its absolute fitness. Instead, it's more common to use relative fitness and selection coefficients in evolutionary biology studies. Relative fitness w_w (equation (1.3)) is the absolute fitness normalized to that of another reference (usually, the “wildtype”) genotype in the competition. By definition, the relative fitness of the reference genotypes is 1 ($w_w^{Ref} = 1$).

$$w_w \equiv W_w / W_w^{Ref} = 1 + s_w \quad (1.3)$$

Fitness effect s_w (equation (1.4)), or called selection coefficient, is defined as the difference between w_w and w_w^{Ref} .

$$s_w \equiv w_w - w_w^{Ref} = w_w - 1 \quad (1.4)$$

Note that the fitness effect can mean fitness advantage or disadvantage depending on its sign. In some studies, the most beneficial variant is chosen as the reference, in which case it is common

to define the fitness effect as $s = w_w^{Ref} - w_w$ such that s means reduction in fitness. In the study of adaptation, it is natural to choose the ancestor (wildtype) as the reference genotype. In this thesis, the fitness effect is defined with ancestor as reference genotype. With that, the beneficial mutations would result in $s_w > 0$ since they have greater fitness than the ancestor. Deleterious mutations have a negative selection coefficient $s_w < 0$.

The utility of the selection coefficient is that it links our theoretical understanding of how genotype's abundances change in competition to measurable quantities. Consider a relative size $\phi = n/n^{Ref}$ as the ratio of population size of one focal genotype and the reference. Experimentally, ϕ is easy to measure. One can measure the genotype's frequencies (fraction) x in population, using, for example, drug or fluorescent markers, and calculate $\phi = x/x^{Ref}$ from the measurements. Based on equation (1.2) and (1.3), the relative size ϕ can be written as

$$\phi_k = w_w \phi_{k-1}, \quad (1.5)$$

and the change of ϕ over one generation is

$$\Delta\phi = \phi_k - \phi_{k-1} = s_w \phi_{k-1}. \quad (1.6)$$

Equation (1.6) suggests a way to measure selection coefficients. However, in microbial populations, which are the focus of this thesis, the assumption of discrete non-overlapping generations is usually not met. It is therefore convenient to consider a model of evolution with continuous time and the corresponding definition of the selection coefficient, which I discuss next.

1.1.2 Absolute fitness and fitness effect in continuous time

In continuous time, absolute fitness is defined as per capita growth rate of population

$$\dot{n}(t) = W_m n(t), \quad (1.7)$$

where $n(t)$ denotes the genotype's population size at time t and W_m refers to the Malthusian fitness. For a simple Malthusian (exponential) growth model, the solution of population size $n(t)$ is

$$n(t) = e^{W_m t} n(0) \quad (1.8)$$

with growth rate W_m . Here the Malthusian fitness (continuous time) and the Wrightian fitness (discrete time) is not the same. Based on the definitions (equation (1.8) & (1.2)), the relation between two absolute fitness is given by

$$W_m = \ln W_w. \quad (1.9)$$

The fitness effect (selection coefficient) in continuous time is defined similarity to the s_w in discrete time:

$$\dot{\phi}(t) = s_m \phi(t). \quad (1.10)$$

Combined equations (1.10) & (1.7), the fitness effect s means the difference in Malthusian fitness between the focal genotype and the reference genotype,

$$s_m = W_m - W_m^{Ref}, \quad (1.11)$$

whereas in discrete time, as discussed above, fitness effect is given by

$$s_w = w_w - w_w^{Ref}. \quad (1.12)$$

Note that both relative fitness and fitness effect means to the biological fitness normalized in certain ways: relative fitness w is a normalized fitness such that the reference is one ($w_w^{Ref} = 1$), where the fitness effect is a normalized fitness such that the reference is zero ($s_w^{Ref} = 0, s_m^{Ref} = 0$). The biological meaning of fitness effect does not change in discrete time or continuous time. It is easy to show that the expression of the fitness effect is equivalent in slow grow (such that

we can approximate discrete time to continuous time). Consider in slow growth that a linear approximation of equation (1.9) for $W_w \approx 1$ leads to $W_m \approx W_w - 1$. By substituting the Wrightian fitness W_w by Malthusian fitness W_m , the fitness effect s_w in discrete time can be expressed by the Malthusian fitness, as shown in equation (1.13)

$$\begin{aligned}
 s_w &= w_w - w_w^{Ref} \\
 &= \frac{W_w}{W_w^{Ref}} - 1 \\
 &\approx \frac{W_m + 1}{W_m^{Ref} + 1} - 1 \\
 &\approx (1 + W_m)(1 - W_m^{Ref}) - 1 \\
 &\approx W_m - W_m^{Ref} = s_m
 \end{aligned} \tag{1.13}$$

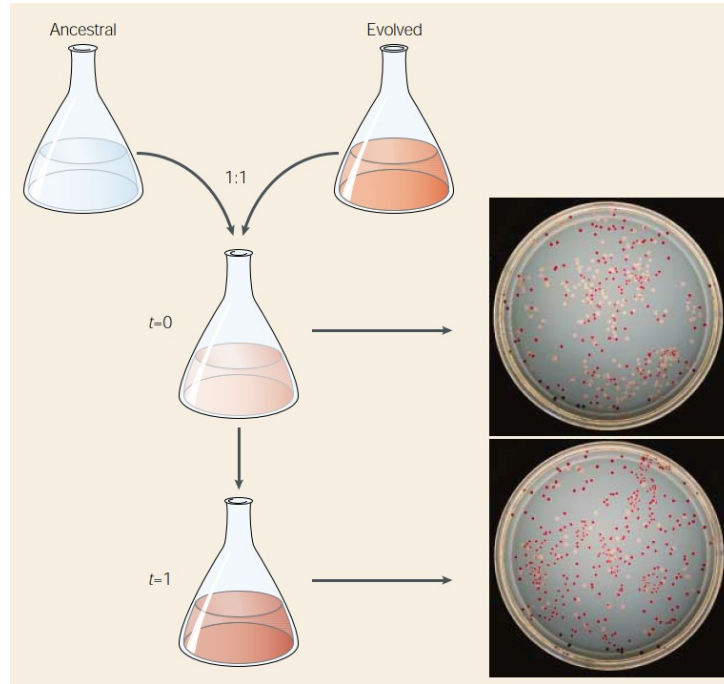


Figure 1.1. Schematic figure of pairwise competition assay experiment for measuring fitness (from Ref. [19])

The fitness effect is measured experimentally by pairwise growth competition experiment. A pairwise competition assay introduces two strains (reference genotype and target genotype)

mixed 1:1 and diluted in a competition environment. As shown in the figure 1.1, the initial densities and final densities of two strains are estimated. The “growth rate” $\tilde{W}_m$ of each competitor is calculated as the natural logarithm of the ratio of its final density $\rho(t + \Delta t)$ to its initial density $\rho(t)$:

$$\tilde{W}_m = \frac{1}{\Delta t} \ln \left(\frac{\rho(t + \Delta t)}{\rho(t)} \right), \quad (1.14)$$

and the fitness effect of target genotype is calculated as the difference in $\tilde{W}_m$:

$$\tilde{s}_m = \tilde{W}_m - \tilde{W}_m^{Ref} \quad (1.15)$$

In fact, the growth curve does not follow the Malthusian growth if a microbial organism is transferred in a fresh media. Experimentally, population growth usually includes four phases (figure 1.2): lag phase, log phase (exponential growth), stationary phase, and dead phase. Therefore, measuring populations at different growth phases would result in different values of the estimated growth rate $\tilde{W}_m$. However, by definition, fitness is the *overall* reproductive success of an organism across all conditions that it encounters during its lifetime, which may be difficult to measure in practice. Measuring fitness differences is typically easier.

Equation (1.16) shows that $\tilde{s}_m$ is equivalent to the fitness effect defined in equation (1.10). Therefore, $\tilde{s}_m$ can be estimated from the direct measurement of relative frequency $\phi(t)$ in a growth competition assay. One can sample the entire population over two time points in the competition growth and calculate the ratio of each strain (target strain and reference strain) using, for example, drug or fluorescent markers. For example, figure 1.3 shows the ratio (frequency) of two competition strains over several days of growth. The relative frequency trajectory $\phi(t)$ is calculated as the ratio between the frequency of target strain (green trajectory in figure 1.3) and the frequency of reference strain (red trajectory in figure 1.3). The fitness effect s of target strain is estimated by the $\phi(t)$ based on equation (1.10).

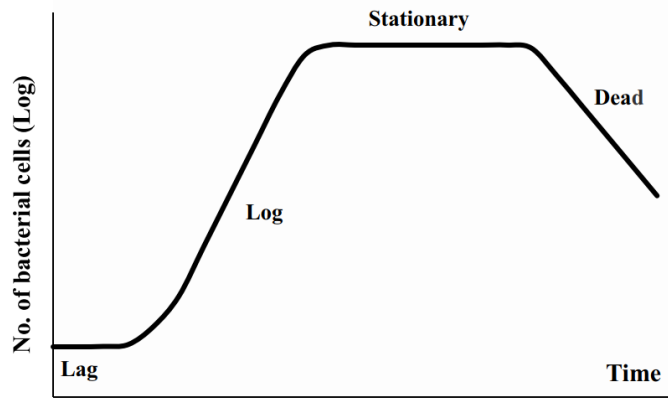


Figure 1.2. Graphic representation of typical bacterial growth curve in culture medium [104]

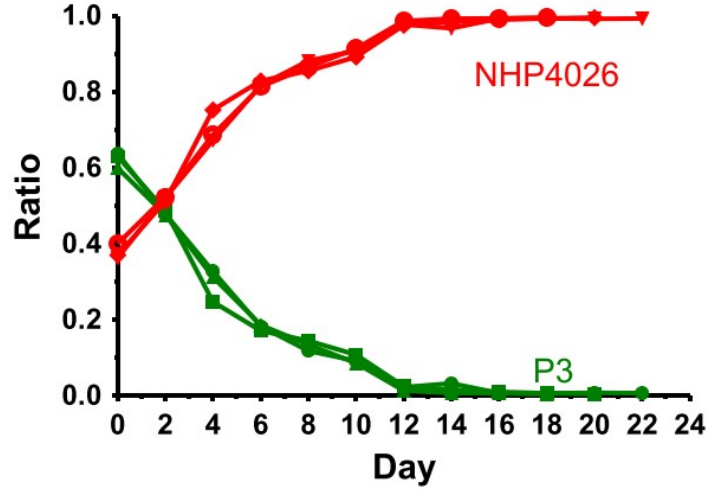


Figure 1.3. Example of pairwise competition between parasite *P. falciparum* wildtype (NHP4026, red) and an artemisinin-sensitive mutation in *kelch13* (P3, green) in the flask (from Ref. [98]).

$$\begin{aligned}
 \tilde{s}_m &= \tilde{W}_m - \tilde{W}_m^{Ref} \\
 &= \frac{1}{\Delta t} \ln \left(\frac{\rho(t + \Delta t)}{\rho(t = 0)} \right) - \frac{1}{\Delta t} \ln \left(\frac{\rho^{Ref}(t + \Delta t)}{\rho^{Ref}(t = 0)} \right) \\
 &= \frac{1}{\Delta t} \ln \left(\frac{\rho(t + \Delta t)}{\rho^{Ref}(t + \Delta t)} \right) - \frac{1}{\Delta t} \ln \left(\frac{\rho(t)}{\rho^{Ref}(t)} \right) \\
 &= \frac{1}{\Delta t} \ln(\phi(t + \Delta t)) - \frac{1}{\Delta t} \ln(\phi(t)) = \frac{1}{\phi(t)} \frac{\Delta \phi(t)}{\Delta t}
 \end{aligned} \tag{1.16}$$

The “**fitness effect**” used in the dissertation is $\tilde{s}_m$, the difference of overall fitness between the mutant strain and the ancestor strain. I term it as s to simplify the notation.

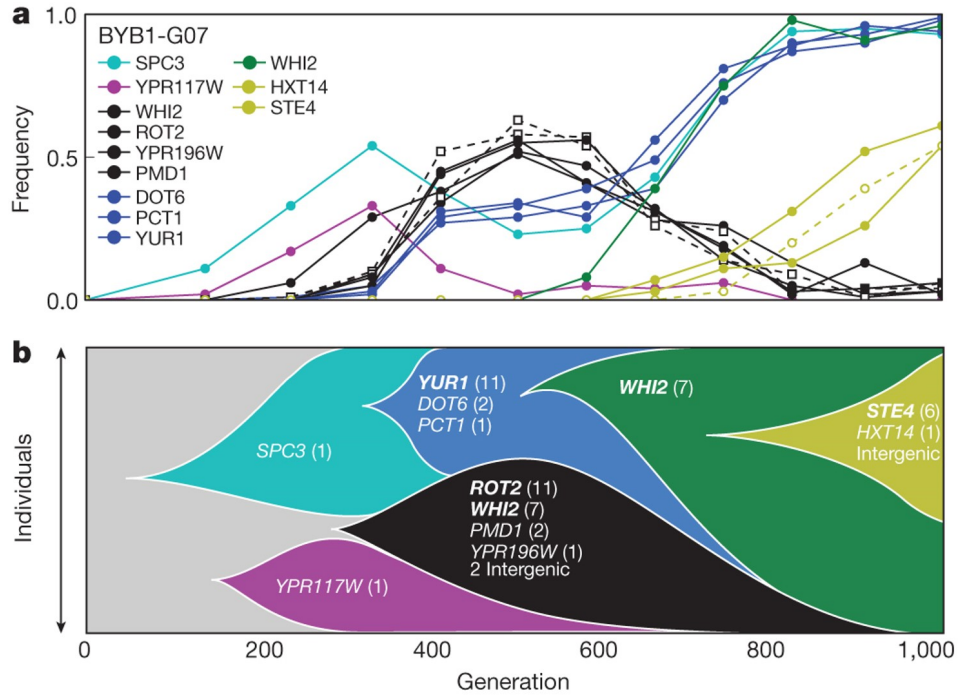


Figure 1.4. The dynamics of sequence evolution in BYB1-G07 (from Ref. [57]).

1.1.3 Lineage trajectory and the mean fitness effect

A lineage trajectory, the frequency of a small subpopulation of cells over time, can be used to discover beneficial mutations and estimate their fitness effect. The simplest example of the lineage trajectories is the frequency trajectories (red & green in figure 1.3) in a pairwise competition assay. However, a competition between two marked genotypes often masks the underlying diversity of new adaptive mutations that are being continuously generated [27, 38, 46]. When many adaptive mutations are present in a population at the same time, their frequencies would increase or decrease depending on their own fitness effects and the frequencies of all other lineages (figure 1.4 as example).

Here, I introduce the mathematical expression of lineage trajectory for any target mutation presented in an evolving population and introduce the concept of the population's mean fitness effect. Consider a large microbial population with genotypes (lineages) $i = 1, 2, \dots, m$, and each lineage has a fitness effect s^i compared to the reference lineage. For each lineage i , its frequency

trajectory is denoted as $x^i(t)$ and its relative frequency is denoted as $\phi^i(t) = x^i(t)/x^{Ref}(t)$. The instant change of a lineage's frequency is given as

$$\begin{aligned}\dot{x}^i(t) &= \frac{d}{dt} \left(\phi^i(t) x^{Ref}(t) \right) \\ &= \dot{\phi}^i(t) x^{Ref}(t) + \phi^i(t) \dot{x}^{Ref}(t).\end{aligned}\tag{1.17}$$

Substituting $\dot{\phi}^i$ in the left part of equation (1.17) by equation (1.10), we have

$$\begin{aligned}\dot{x}^i(t) &= s^i \phi^i(t) x^{Ref}(t) + \phi^i(t) \dot{x}^{Ref}(t) \\ &= s^i x^i(t) + \phi^i(t) \dot{x}^{Ref}(t).\end{aligned}\tag{1.18}$$

The reference frequency $x^{Ref}(t) = n^{Ref}(t)/N(t)$ is the fraction of reference size $n^{Ref}(t)$ to the total population size $N(t) = \sum_i n^i(t)$. The time derivative of $x^{Ref}(t)$ is given as

$$\begin{aligned}\dot{x}^{Ref}(t) &= \frac{1}{N(t)} \dot{n}^{Ref}(t) - \frac{n^{Ref}(t)}{N(t)^2} \dot{N}(t) \\ &= \frac{1}{N(t)} W_m^{Ref} n^{Ref}(t) - \frac{n^{Ref}(t)}{N(t)^2} \sum_i \dot{n}^i(t) \\ &= x^{Ref}(t) \left[W_m^{Ref} - \sum_i W_m^i x^i(t) \right] \\ &= x^{Ref}(t) \sum_i \left(W_m^{Ref} - W_m^i \right) x^i(t) \\ &= x^{Ref}(t) \sum_i \left(-s^i \right) x^i(t) \\ &= -x^{Ref}(t) \bar{s}(t).\end{aligned}\tag{1.19}$$

Here, $\bar{s}(t) = \sum_i s^i x^i(t)$, called mean-fitness effect, is the average of fitness effect s for current population composition. Combining equation (1.18) & (1.19), we derive the frequency change of lineage i as

$$\dot{x}^i(t) = \left(s^i - \bar{s}(t) \right) x^i(t).\tag{1.20}$$

The frequency change of a lineage i depends on three quantities: its current frequency $x^i(t)$, its fitness effect s^i , and current mean fitness effect $\bar{s}(t)$. Taking mean-fitness as a constant in a short time limit, we can integrate $\dot{x}^i(t)$ over a small time frame Δt and obtain

$$x^i(t + \Delta t) = x^i(t) e^{(s^i - \bar{s}(t))\Delta t}. \quad (1.21)$$

If the lineage's fitness effect is greater than the current population-mean fitness effect (i.e., $s^i > \bar{s}(t)$), it will increase in frequency; otherwise, the frequency decreases.

This expression of **lineage's frequency dynamic (equation 1.21)** is going to be central in the subsequent chapters.

1.1.4 Distribution of fitness effect (DFE)

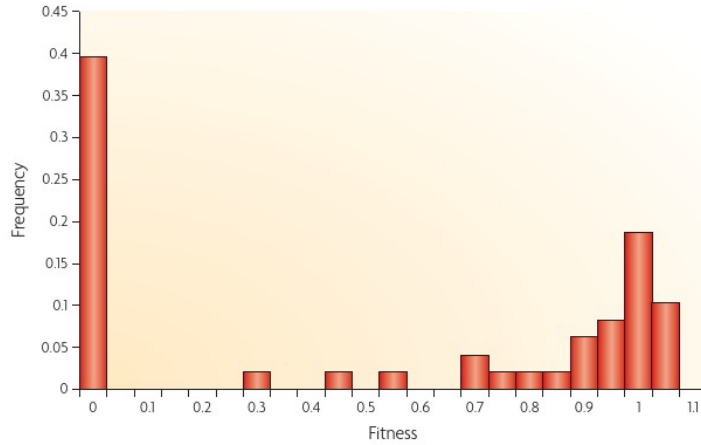


Figure 1.5. The distribution of fitness effects of random mutations in vesicular stomatitis virus (from Ref. [22])

What fitness effect does a new mutation have as it arises in an evolving population? Answering this question with certainty is of course not possible because mutations are random events. But it is possible to characterize the statistics of new mutations. The quantity that describes the probability that a new mutation will have a certain effect on fitness is called the distribution of fitness effects of new mutations and is usually abbreviated as “DFE”. The DFE is

a fundamental quantity in population genetics because it allows us to quantitatively understand many aspects of the evolutionary process. For example, knowing the DFE, we can predict how fitness would change over time [32, 34, 54, 106], how mutations would accumulate in the population [32, 54]. It is also relevant for understanding many evolutionary phenomena, such as the molecular clock [74, 83], the rate of genomic decay [64], the maintenance of genetic variation [11], and the evolution of recombination[51, 81]. Despite its importance, direct measurements of the DFE remain challenging. In most organisms, beneficial mutations are rare events. Isolating beneficial mutations from the adapted population in an unbiased manner is also challenging because weakly beneficial mutations are less likely to survive and be sampled. As such, the main challenge is that measuring the DFE requires observing many independent adaptive mutations with as great diversity of fitness effects as possible. Recently, an experimental approach termed “barcode lineage tracking” (BLT) has been developed for learning the statistics of fitness of many adaptive mutations simultaneously.

1.2 Introduction to the Barcode Lineage Tracking Experiment

Barcode lineage tracking (BLT) approach has been developed to observe many independent and isolated adaptive mutations with a single experiment [1, 2, 7, 14, 42, 43, 58, 60, 62, 100, 101]. This approach introduces a massive number of unique DNA barcodes into a large (ancestor) population. By deep-sequencing the barcodes, barcode frequencies can be tracked to a very low limit ($10^{-5} - 10^{-7}$). As a new adaptive mutation arises during evolution, it is permanently linked to one barcode and leads to an observable sustained increase in its barcode frequency. Therefore, the fitness effects (selection coefficients) of many beneficial mutations, including weak ones, can be inferred from these rising trajectories.

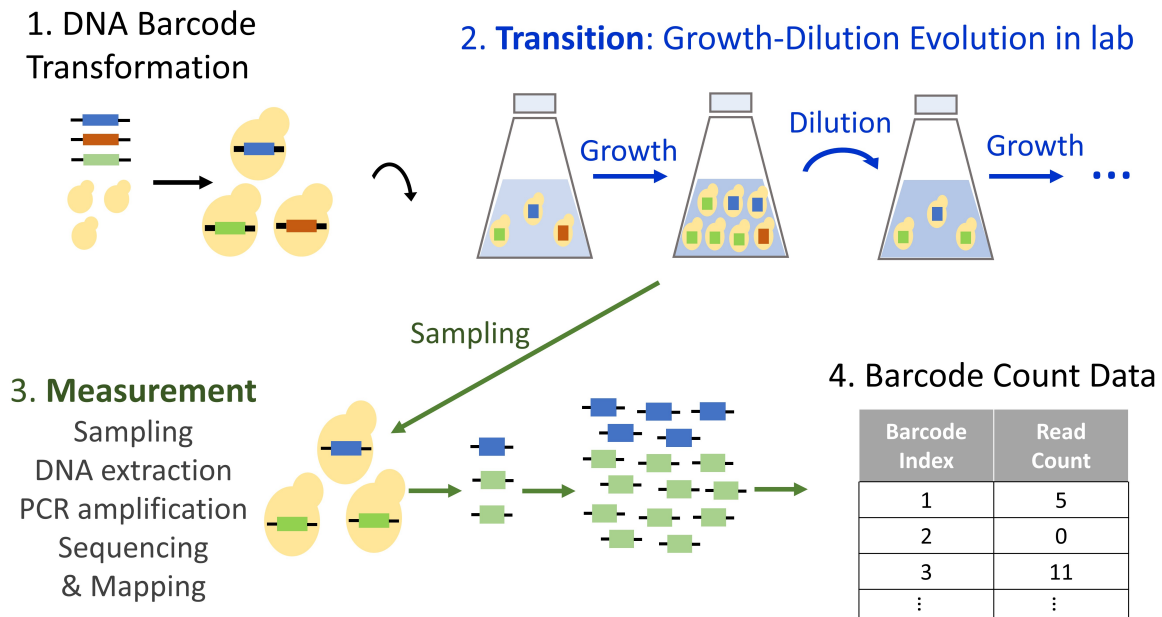


Figure 1.6. Overview of barcode lineage tracking (BLT) experiment

1.2.1 DNA barcode

DNA barcode is a random sequence of nucleotides in length of 20-30 bp. The length of barcode is designed to create enough information to distinguish many (typically, 10^4 to 10^5) lineages. The barcode libraries are usually preserved in synthetic oligonucleotides (oligo) libraries, that are integrated into plasmids and typically stored in *Escherichia coli*. To create a barcode-tagged strain, the barcode plasmid is transformed and integrated into the target organism either on a chromosome or an extra-chromosomal plasmid.

1.2.2 Evolution experiment

To observe the appearance and spread of adaptive mutations, a barcode-tagged microbial population is usually evolved for hundreds of generations under laboratory conditions. In a typical batch-culture setup, the barcoded population grows for a few generations until it exhausts the nutrients present in the medium after which point, the culture is diluted into fresh medium. The choice of bottleneck size during the transfer is determined by a balance between various practical considerations. For example, a large bottleneck size would decrease the number of

generation in each growth-diluting cycle thereby potentially slowing the experiment. On the other hand, a small bottleneck size would reduce lineage diversity decreasing the chances of finding adaptive mutations. For example, for 10^5 lineages growing in fresh media with a carrying capacity of 10^{11} cells, the bottleneck size can be 10^7 cells to ensure that each lineage carries 100 cells on average; therefore, the evolved time of one batch-growth is $\log_2(10^{11}/10^7) \approx 13$ generations.

1.2.3 Barcode measurement

A measurement of frequencies of barcoded lineages requires several steps [44]. First, the barcode carrying DNA is extracted from the sampled population. The DNA template is clean up and the barcode is amplified with the Polymerase Chain Reaction (PCR). During PCR, overhangs that contain sequences necessary for multiplexing and sequencing are attached to the barcode-containing fragments. Finally, the barcode amplicon DNA is purified and sequenced, typically on an Illumina platform.

1.3 Challenges, Objectives, and Approaches

Barcode lineage tracking (BLT) experiments provide a simple, rapid, and quantitative approach to measuring the distribution of fitness effect (DFE) of beneficial mutations for asexual microbes. While a statistical method for DFE inference from BLT experiments has been developed by Levy et al (Ref. [58]), its precision, accuracy and the domain of applicability have not been rigorously evaluated. In particular, it remains unclear how to optimally model genetic drift and the increase of mean fitness effect during evolution. As I will show in the next chapter, the previously developed BLT analysis method is limited to conditions where selection is weak. However, the strength of selection is usually unknown a priori. Thus, it is necessary to develop a more robust method for the analysis of BLT experimental data.

1.3.1 Challenges to evolutionary dynamics analysis

The main problem addressed in this thesis (Chapter 3) is how to precisely and accurately estimate the fitness effect of a nascent beneficial mutation from the BLT data. While the fitness effect of a lineage can be inferred from its frequency trajectory, such inference relies on an accurate estimate of the mean fitness effect. Theoretically, the mean fitness effect can be inferred from the frequency trajectory of any single reference (ancestor) lineage, but such estimate would be very noisy because frequencies of individual barcoded lineages are very low ($10^{-5} \sim 10^{-7}$). Since errors in the estimate of mean fitness translate into biases in the estimates of fitness effects of individual mutations, it is important that mean fitness is estimated precisely. The Levy-Blundell method for BLT data analysis attempts to overcome the noisy problem by using a “bulk” reference consisting of a subset of “small” (i.e., low-frequency) lineages (20 – 50 barcode read count). However, this approach is limited to evolutionary conditions with weak selection, as will be shown below.

The second problem, addressed in Chapter 4, is to estimate the rate at which different types of adaptive mutations arise. How many different adaptive routes are available to evolution and how likely each route is to be taken are long-standing questions in evolutionary biology [8, 19, 21, 75, 94]. While the beneficial mutations that drive the frequency increases in adapted barcode lineages can be identified by whole genomics sequencing, there has been no robust method for inferring the rate at which different types of adaptive mutations arise. As a beneficial mutation randomly arises in a single cell in the population, the mutant has to survive genetic drift and sufficiently expand in size before it can be observed. Thus, an inference of the mutation rate from BLT observations hinges on a model of this stochastic process.

The last challenge relates to the distribution of the fitness effect. The DFE of new beneficial mutations (DFE) is a fundamental evolutionary quantity, and it has been measured in evolutionary studies [1, 58] through the barcode lineage tracking (BLT) experiments. However, the measured DFE is not exactly the true underlying DFE. It is because lineages with weakest

fitness effects are less likely to reach a large size and be detected before they begin to decline due to the competition. How do we estimate the true DFE from the measured DFE? It is surprising that this question is never been investigated systematically.

1.3.2 Objectives and basic approaches of this dissertation

The overarching aim of this dissertation is to develop novel statistical methods for the analysis of BLT experiments.

The first objective, addressed in Chapter 2, is to investigate the Levy-Blundell method for the inference of fitness effects of adapted lineages, delineate its domain of validity and find possible reasons for its failure outside of this domain.

The second objective, addressed in Chapter 3, is to develop a new more robust method for fitness inference that overcomes the problems identified in Chapter 2. I refer to this novel approach as the Bayesian Filtering method.

The third objective, addressed in Chapter 4, is to develop a method for estimating the rate of beneficial mutations that belong to a given class.

The last objective, addressed in Chapter 5, is to identify the extent to which the DFE that can be directly calculated from the adapted barcoded lineages is biased relative to the true underlying DFE and to outline an approach for how this bias can be corrected.

Overall, the work presented in this dissertation forms a statistically solid foundation for the analysis of BLT experiments.

Chapter 2

Shortcomings of the Levy-Blundell Method

2.1 Introduction

As mentioned above, [58] were the first to carry out a BLT experiment in an evolutionary context. They also developed a method for identifying adapted lineages and inferring their fitness, which I refer to as the “Levy-Blundell” method. Since most the subsequent BLT work [1, 2, 7, 59, 61, 63, 73, 91, 101] applies methods that is based on the same approach, I focus here on the analysis of this method. In particular, in this Chapter, I briefly review the Levy-Blundell method and discuss its key concept and assumptions. I then use simulations to demonstrate that the key assumptions fail under certain scenarios, which can lead to inaccurate identification of adapted lineages and their fitness. In the next Chapter, I propose a new method that overcomes these problems.

2.2 Review of the Levy-Blundell method

The Levy-Blundell method establishes a stochastic model to characterize lineage dynamics. The model is based on a simple birth-and-death process as a mutant population starting from an initial number of cells n_0 randomly grows and/or shrinks with a given birth rate and death rate. They derived the approximate expression for the probability $P_t(n)$ that the mutant population

size is n after time t ,

$$P_t(n) = \sqrt{\frac{\mu_t^{1/2}}{4\pi\kappa_t n^{3/2}}} \exp\left[-\frac{(\sqrt{n} - \sqrt{\mu_t})^2}{\kappa_t}\right] \quad (2.1)$$

The probability $P_t(n)$ depend on two parameters: the expected lineage size μ_t at time t and the noise parameter κ_t . The mean $\mu_t = n_0 e^{\lambda t}$ depends on the growth rate λ , which is the difference between the birth and death rates. $\kappa_t = \frac{1}{2\lambda}(e^{\lambda t} - 1)$ equals half of the variance to mean ratio.

Equation (2.1) represents the underlying dynamics of a lineage in the real biological population. However, the actual number of cells that belong to a given barcoded lineages is not observed. Instead, one samples cells from the population at each sampling time k , sequences their barcodes (after a number of technical processing steps) with a total read depth R_t (total reads of sequence) and observes that a certain number r_k of the resulting sequencing reads belong to a given lineage. Here I refer to r_k as the barcode count at sampling time k . Levy et al. use an expression analogous to equation (2.1) to model the observed dynamics of lineage barcode counts. Specifically, they posit that the probability of observing read count r_k at the sampling time k for a lineage with read count r_{k-1} at the previous sampling time is given by

$$P(r_k) = \sqrt{\frac{\mu^{1/2}}{4\pi\kappa r_k^{3/2}}} \exp\left[-\frac{(\sqrt{r_k} - \sqrt{\mu})^2}{\kappa}\right], \quad (2.2)$$

with μ defined by

$$\frac{\mu}{R_k} = \frac{r_{k-1}}{R_{k-1}} e^{(s - \bar{s}_k)\Delta t_k}, \quad (2.3)$$

where Δt_k is time between two sampling events, s and $\bar{s}_k$ are the selection coefficient of lineages and the mean population fitness effect. Unlike in the birth-death process, the variance parameter κ is a priori unknown since it characterizes all systematic noise in the BLT experiments, including growth, dilution, DNA extraction, DNA amplification, etc. Levy et al. estimated $\kappa \approx 3$ in their experiment. To determine μ , they model the expected read μ (Equation 2.3) similarly to the birth-death process. The barcode frequency would increase or decrease from the previous cycle

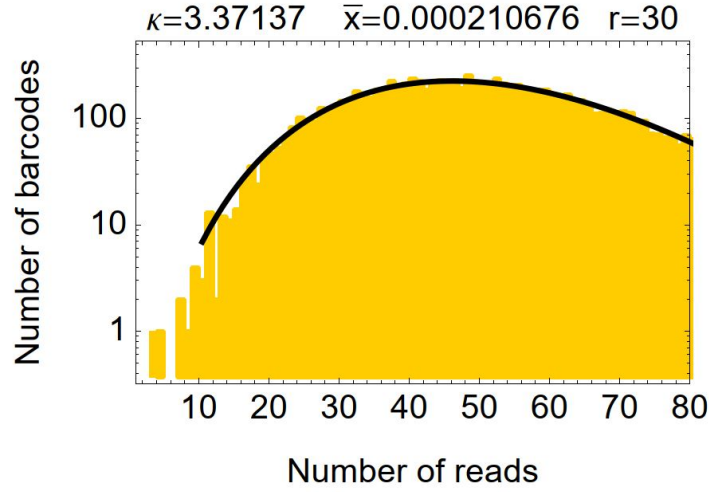


Figure 2.1. Example of fitting the noise parameter κ and the mean fitness $\bar{s}_k$ (denoted as $\bar{x}$ in figure) from a histogram distribution of $P(r_k)$ constructed for barcodes with $r_{k-1} = 30$. Figure is obtained from Supplementary Figure 19 in Ref. [58].

if the lineage has s greater or less than the mean of population $\bar{s}_k$.

The first step in the Levy-Blundell method is to use the equation 2.3 to infer the mean fitness $\bar{s}_k$. Since the chance that an adaptive mutation occurs in on a given barcode background depends linearly on the lineage size, most low-frequency lineages fail to acquire an adaptive mutation during the relatively short duration of the BLT experiment and thus remain neutral, i.e., their $s = 0$. They exploit this fact to infer $\bar{s}_k$. In particular, they select lineages with read count $r_{k-1} = 20 - 40$ at each time point $k - 1$ as a “neutral reference”. They then obtain the distribution of read counts r_k at the next time point k for these neutral reference lineages. Levy et al. then apply fit Equation 2.2 (with $s = 0$) to this distribution to infer the mean-fitness $\bar{s}_k$ and the noise parameter κ . See figure 2.1 as an example.

Once the mean-fitness trajectory $\bar{s}_k$ is known, the Levy-Blundell method takes the entire barcode frequency trajectory of each lineage, and infers the most likely selection coefficient s and the establishment time of a putative adaptive mutation. Finally, it identifies which lineages likely acquired an adaptive mutation during the BLT experiment by comparing the likelihoods of the observed trajectory under a neutral model (equation 2.3 with $s = 0$) and the adaptive model

(equation 2.3 with the most likely s).

2.3 Critical assumptions of the Levy-Blundell method

2.3.1 The “neutral decline” assumption

Estimating the fitness effect of a beneficial mutant in BLT analysis requires knowing population’s mean fitness trajectory (see equation (1.21)). If the mutant’s fitness is greater than the current mean fitness of the population, the frequency of the mutant increases; otherwise, the frequency would decline. As mentioned above, the mean fitness trajectory can be inferred from the trajectories of ancestral (neutral) lineages, i.e., those that have not acquired an adaptive mutation. Since the neutral lineage have, by definition, no fitness advantage over the ancestor (i.e., $s = 0$), equation (1.21) for the frequency $x(t)$ of a neutral lineage becomes

$$x(t + dt) = x(t)e^{-\bar{s}(t)dt}. \quad (2.4)$$

In principle, equation (2.4) can be easily solved for the mean fitness $\bar{s}(t)$, provided that the neutral lineage trajectory is known. However, in practice, the true frequency of neutral (or any other) lineages in the population is not known. Instead, the Levy-Blundell method approximates the true barcode frequency in the population by the frequency of the corresponding barcode in the sequencing data, i.e.,

$$\frac{\mu}{R_k} = \frac{r_{k-1}}{R_{k-1}} e^{-\bar{s}_k \Delta t_k}, \quad (2.5)$$

where μ is the expected count of the focal barcode at the sampling time k and R_k is the total barcode count at time k . I refer to this approach as the “**neutral decline**” method. The key assumption of the neutral decline method is that the observed sampled barcode frequency is an accurate and unbiased estimator of the true population frequency of a lineage. At first glance, this assumption seems to be correct because the sample frequency (asymptotically) equals the true population frequency. However, the expected barcode count μ in equation (2.5) represents

the expected barcode count at the next sampling time point of all lineages whose barcode count at the previous sampling time was r_{k-1} , and such lineages may (and in general do) have different true population frequencies. It is thus not a priori clear whether such conditional expected value in fact changes according to equation (2.5). Indeed, I next demonstrate that this not the case in general.

To this end, I derive the barcode dynamics directly from cell dynamics (equation (1.21)). I first denote $\mu = \mathbb{E}(r_k | r_{k-1})$ as the expected barcode read count for lineages with read count r_{k-1} at the previous sampling time $k - 1$. The expected value can be broken down to the integration over read count and cell number. See Appendix A for calculations. I work out the expected barcode read count for neutral lineages ($s = 0$), which is given as

$$\mathbb{E}(r_k | r_{k-1}) = \frac{R_k}{N_{k-1}} e^{-\bar{s}_k \Delta t_k} \mathbb{E}(n_{k-1} | r_{k-1}), \quad (2.6)$$

where n_{k-1} and N_{k-1} are the lineage (cell) size and the total population size at sampling time point $k - 1$. Note that the expected size $\mathbb{E}(n_{k-1} | r_{k-1})$ of lineage depends on the distribution of lineage sizes $P(n_{k-1})$ and the probability $P(r_{k-1} | n_{k-1})$ of observing r_{k-1} reads for a lineage whose true cell count is n_{k-1} (the “measurement process”),

$$\mathbb{E}(n_{k-1} | r_{k-1}) = \int dn_{k-1} n_{k-1} \frac{P(r_{k-1} | n_{k-1}) P(n_{k-1})}{P(r_{k-1})}. \quad (2.7)$$

To learn the properties of $\mathbb{E}(n_{k-1} | r_{k-1})$, I consider the simplest case where the $P(n_{k-1})$ is an exponential distribution, and the measurement process $P(r_{k-1} | n_{k-1})$ is a negative binomial (NB) distribution with mean of $R_{k-1} n_{k-1} / N_{k-1}$ and variance of $2\kappa \times \text{mean}$), in which case I obtain

$$\mu(r_{k-1}) \equiv \mathbb{E}(r_k | r_{k-1}) \approx \frac{R_k}{R_{k-1}} (2\kappa + r_{k-1}) e^{-\bar{s}_k \Delta t_k}. \quad (2.8)$$

Equation (2.8) shows that the expected read count μ has a nonzero offset, i.e., $\mu \neq 0$ when $r_{k-1} = 0$, in contrast to the neutral decline equation (2.5). This offset is determined by

measurement noise κ .

Although the neutral decline equation (2.5) does not hold in general, it can still be used for estimating $\bar{s}$ for certain values of r_{k-1} . Specifically, in the limit of large read $r_{k-1} \gg 2\kappa$, the offset can be ignored, and equation (2.8) is well approximated by the neutral decline equation (2.5). However, lineages with large read counts are more likely to be non-neutral (i.e., have acquired an adaptive mutation), especially at later time points, whereas neutral lineages are more likely to be in the regime $r_{k-1} \sim \kappa$. It may be possible to find a range of values r_{k-1} that are large enough that the neutral decline assumption holds but not so large that they have acquired an adaptive mutation. But what where such range of r_{k-1} values lies is a priori unknown. In the absence of such knowledge, it's expected that the inference of mean fitness would be sensitive to the size of reference lineages chosen for the inference.

2.3.2 The no lineage extinction assumption

The mean-fitness inference in Levy-Blundell method is based on fitting the observed distribution of read counts to the functional form given by equation (2.2). This probability distribution is undefined for $r_k \rightarrow 0$. Thus, it excludes the possibility of sampling zero reads ($r_k = 0$) from the reference lineages. In other words, the model assumes that the reference lineages represented in the sequencing data at one sampling time point are still represented at the next sampling time point. This assumption appears to hold in the Levy et al. dataset. The majority of their reference lineages are still present at the next sampling time (see Figure 2.1). However, in general, the low-frequency lineages could disappear due to genetic drift (extinction) or sequencing noise (pseudo-extinction). Importantly, the strength of genetic drift, which would depend on abiotic environmental factors (e.g., high stress) and biotic competition (many beneficial mutations, with higher fitness advantage), may be very difficult or impossible to control experimentally. When a microbial population evolves under a larger drift, this assumption is expected to fail.

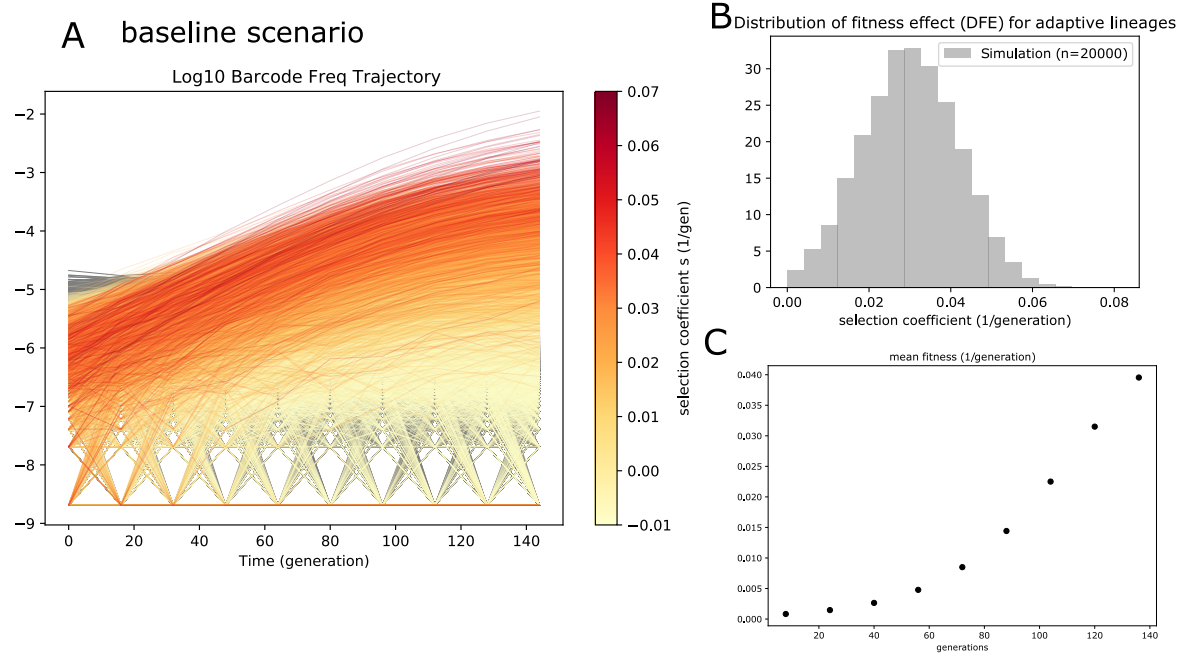


Figure 2.2. Baseline scenario simulation. (A) barcode frequency trajectory (B) DFE (C) mean-fitness trajectory

2.3.3 Probing the validity of the Levy-Blundell model assumptions with BLT simulations

To test the validity of the assumptions on which the Levy-Blundell method is based, I simulated two batch-culture BLT experiment scenarios. In the so-called **baseline scenario**, which is shown in Figure 2.2, selection is weak (such that adaptive mutations provide fitness gains of 1-5%), genetic drift is weak (such that a typical lineage is initially represented by 100 individuals at the bottleneck, and population is measured every 2 transfers) and sequencing coverage is high (such that each individual at the bottleneck is represented by 100 reads). These parameters are similar to those for the BLT experiment published in Ref. [58]. I use this simulation to investigate the validity of the neutral decline assumption and possible consequences of violating it.

In the second scenario, shown in Figure 2.3, which I refer to as the **strong drift scenario**, genetic drift is strong in the sense that a typical lineage is initially represented by 10 individuals at the bottleneck, and population is sampled every 6 transfers. I use this simulation to investigate the validity of the no lineage extinction assumption and possible consequences of violating it.

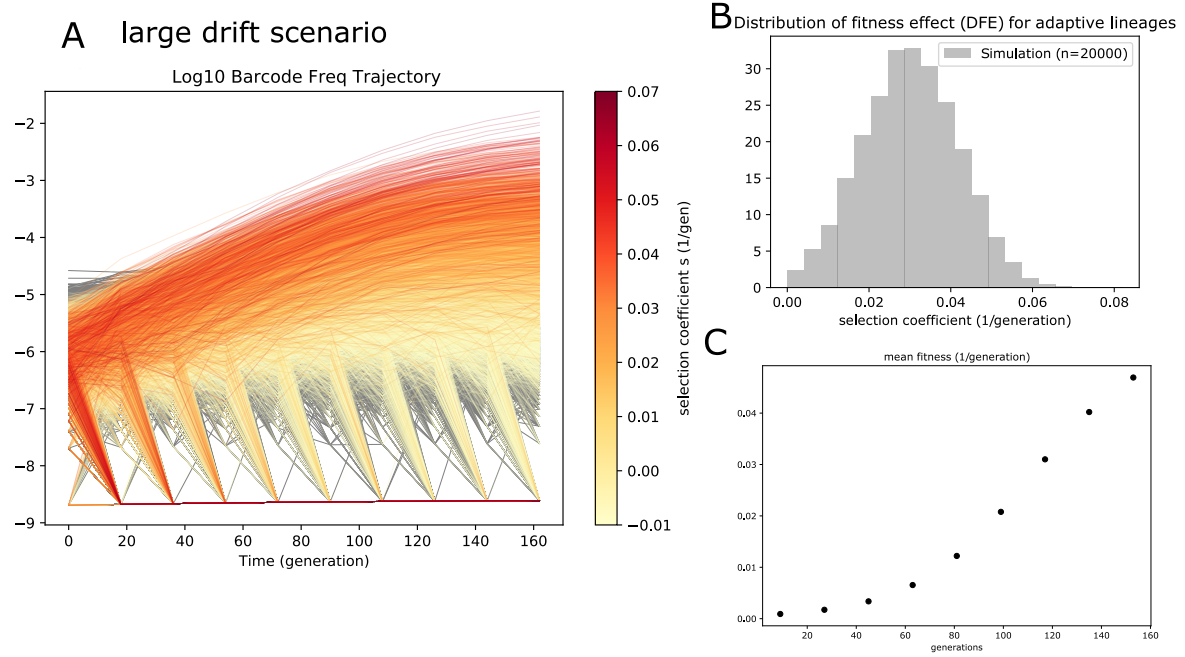


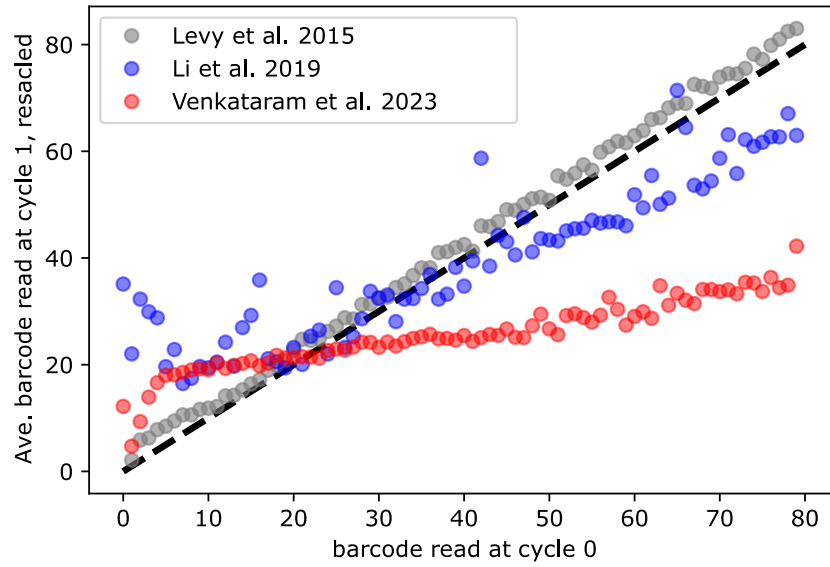
Figure 2.3. Large-drift scenario simulation. (A) barcode frequency trajectory (B) DFE (C) mean-fitness trajectory

2.4 Results

2.4.1 Neutral decline assumption is violated in both experimental and simulation data

I demonstrate the failure of neutral decline principle by both simulated and experimental data. The experimental data is derived from the published yeast BLT experiments published in Refs.[58, 61, 102]. For each experimental dataset, I select “small” lineages, i.e., those whose barcode read count was within 0 to 80 at the initial time sampling point (cycle 0). Those barcodes are grouped with the same read count and calculate their average read count at the next cycle (cycle 1). I plot the average of read count at the cycle 1, rescaled by the ratio of total read counts between two cycles ($y = \mathbb{E}(r_1) \times R_0/R_1$), against the read count at cycle 0. While the neutral decline (equation (2.5)) predicts that expected barcode is simply linear proportional to the barcode read at the previous cycle (black dash line in Figure 2.4 A and B), our calculation of re-scaled read has a nonzero offset. The offset values are 1.47 for Levy et al dataset, 12.05 for

A



B

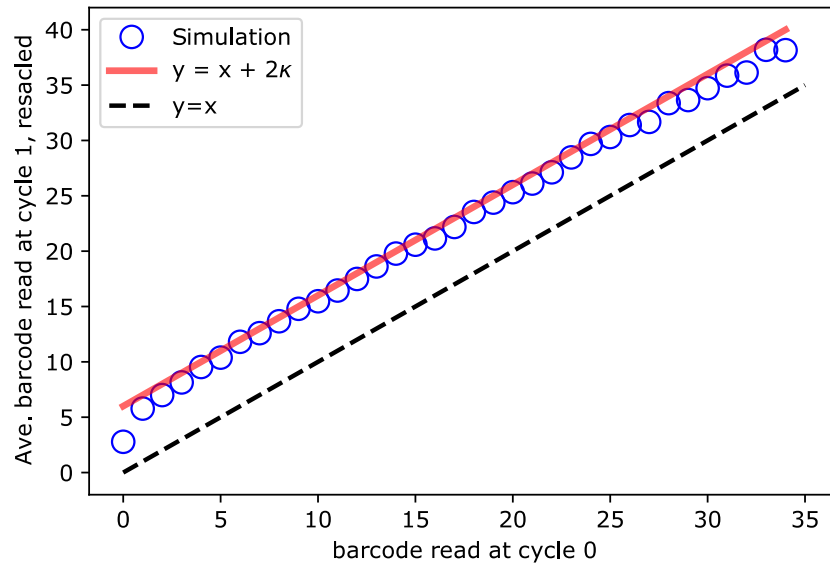


Figure 2.4. Failure of the neutral decline assumption (black dash line) for experimental data (A) and simulated data (B).

Li et al., 16.13 for Venkataram et al. (Figure 2.4 A). Note that the slope of the line is expected to be not equal to 1 if the mean fitness is not zero.

This calculation is also applied to our simulated data in the baseline scenario. Since the prior of lineage size in our simulation is exponential distributed, it is expected that the offset value is noise parameter 2κ . Figure 2.4B shows that the simulation results (blue open circle) agree with the theoretical prediction (red dash line) of barcode dynamics.

2.4.2 Performance of the Levy-Blundell method is sensitive to the choice of reference lineages

I next use our simulated data to demonstrate that the accuracy of fitness inference by the Levy-Blundell method is sensitive to the choice of reference lineages. To this end, I compare the fitness effect inferences that are based on two groups of lineages. In the first group, reference lineages have read counts between 81 and 100, which is close to the typical read count per barcode at the beginning of the BLT simulation. In the second group, reference are selected to have read counts between 21 and 40, the range used by Levy et al. [58]. For each time interval and for each group of lineages with the same read count at the previous time point, I construct the histograms of the read counts at the next time point. I infer the mean fitness of the population and the noise parameter κ by fitting the probability distribution given by equation (2.2) with the mean read count given by the neutral decline equation (2.5) to these histograms (Figures 2.5 and 2.6). I then calculate the “Levy method” mean fitness at each sampling time as an average of inferred mean fitness across all read counts within the given range. Thus, the adapted lineages are identified and their fitness effect s are estimated following the Levy-Blundell method (see Ref. [58] for details).

In addition, I calculate the “reference decline” mean fitness $\bar{s}_k^{\text{Ref}}$ by adding all read counts of selected lineages i together as a big reference. The “reference decline” mean fitness $\bar{s}_k^{\text{Ref}}$ is calculated based on

$$\frac{\sum_i r_k^i}{R_k} = \frac{\sum_i r_{k-1}^i}{R_{k-1}} e^{-\bar{s}_k^{\text{Ref}} \Delta t_k}, \quad (2.9)$$

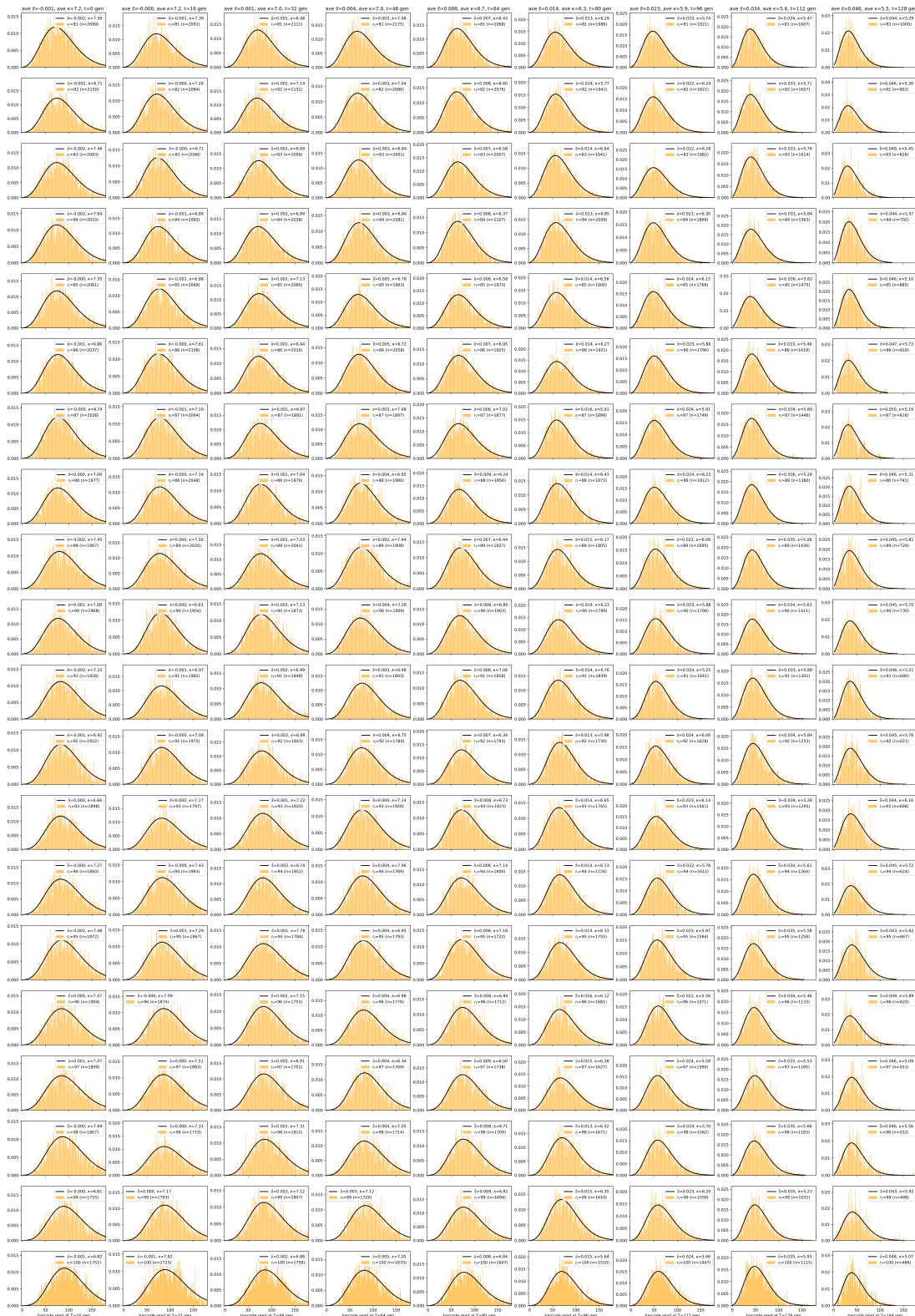


Figure 2.5. Read distribution of large drift scenario for lineages of read 81 – 100 and the inference of mean-fitness and noise based on Levy-Blundell method.



Figure 2.6. Read distribution of baseline scenario for lineages of read 21 – 40 and the inference of mean-fitness and noise based on Levy-Blundell method.

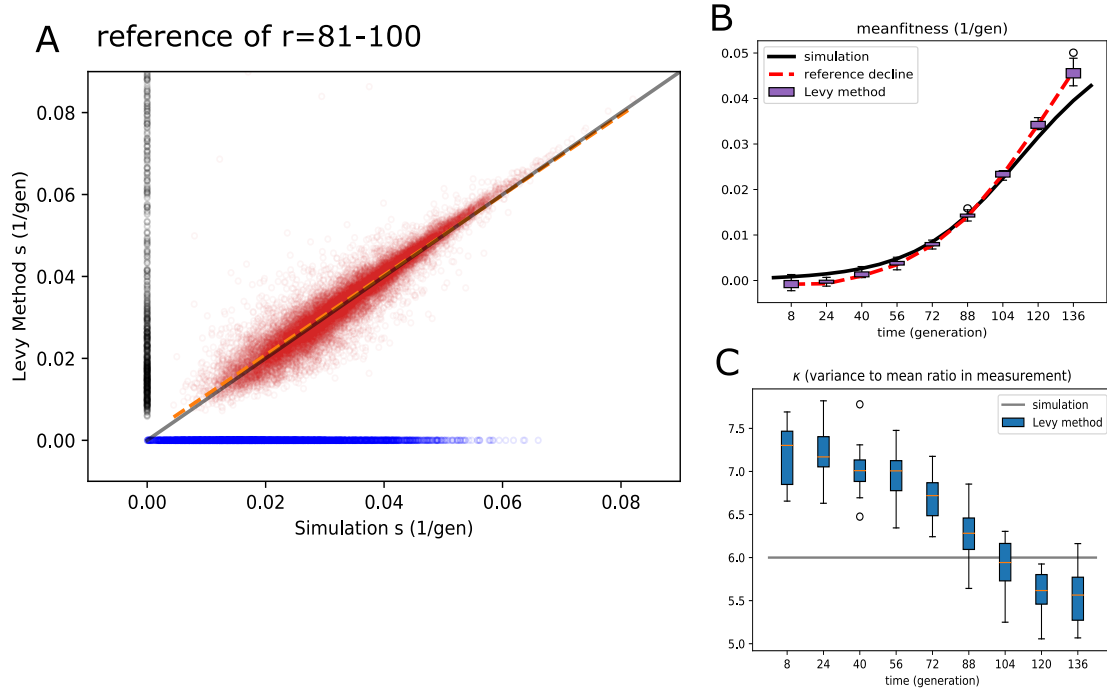


Figure 2.7. Performance of the Levy-Blundell method on simulated data in the baseline scenario, with reference lineages chosen to have read counts between 81 and 100. (A) Estimated fitness effects for true positives, i.e., adapted lineages (red), false positives, i.e., neutral lineages identified as adapted (black), and false negative adapted lineages not identified as such (blue). (B) mean fitness trajectory inferred by the Levy-Blundell method (purple bars), by bulk neutral decline, compared to the true simulated mean-fitness trajectory (black dash line). (C) Noise parameter κ inferred by the Levy-Blundell method (blue bars) compared to the true noise parameter in the simulation (black line).

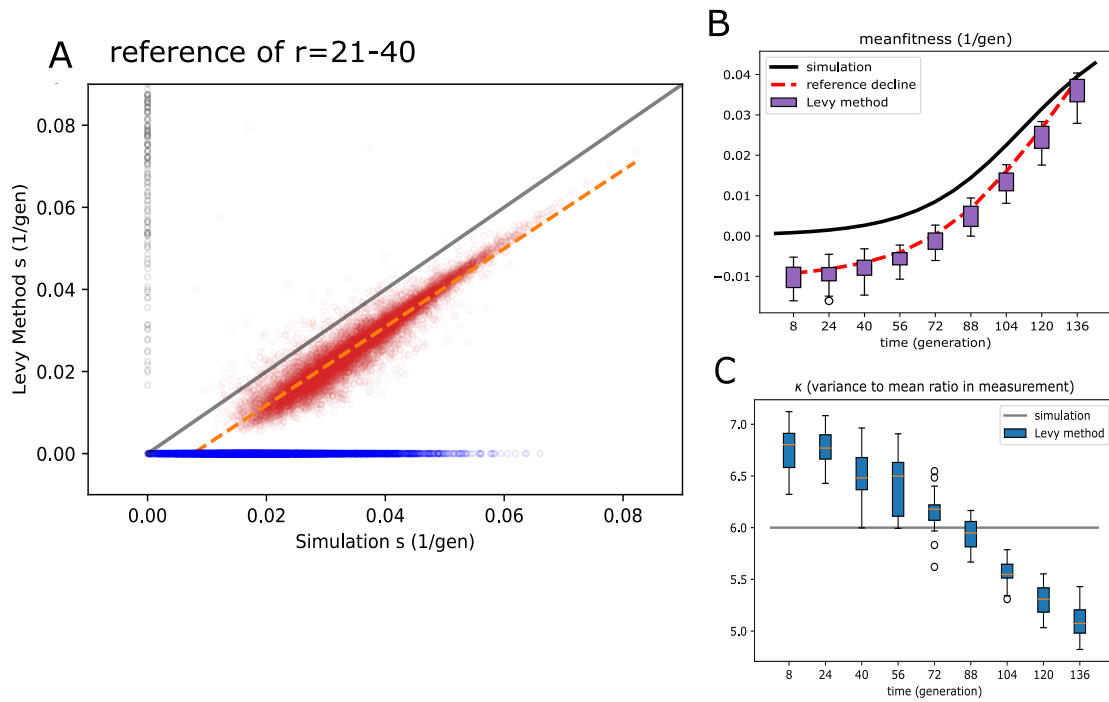


Figure 2.8. Lineage identification and fitness estimation by Levy et al. with reference chosen from 21 – 40 read count for the baseline simulation case. (A) Fitness effect estimation for true positive (red), false positive (black) and false negative (blue) lineages. (B) mean fitness trajectory inferred by Levy et al average (purple bar), by bulk neutral decline (red dash line) compared to the simulation mean-fitness trajectory (black dash line). (C) noise parameter inferred by Levy et al. average (blue bars) compared to the simulated noise (black line)

where r_{k-1} is from 81 and 100 for the first group, and r_{k-1} is from 21 and 40 for the second group.

I find that the Levy-Blundell method performs well for the first reference group ($80 < r \leq 100$) with 79% true positive rate, 93% precision of adapted calls, and 0.4% error in the inferred selection coefficients (Figure 2.7A). In contrast, the method's performance is significantly worse for the second reference groups ($20 < r \leq 40$). While the method correctly identifies many lineages as adapted (69% true positive rate and 96% precision of adapted calls), it systematically underestimates the selection coefficients of adapted lineages by 24% (Figure 2.8A). This bias is caused by the violation of the neutral decline assumption, which leads to underestimate of mean fitness inference (compare Figures 2.7 B and 2.8 B), which in turn propagates to the bias in the estimates of selection coefficients of all lineages.

2.4.3 Failure of lineage identification in the presence of strong genetic drift

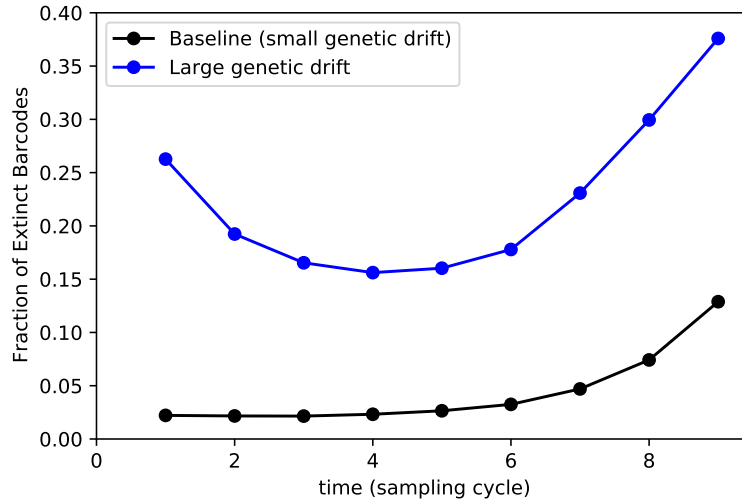


Figure 2.9. fraction of barcodes which are present in the previous time but have zero read count at the next time point for (black) baseline and (blue) strong drift scenario.

I next demonstrate that the Levy-Blundell method, which ignores lineage extinctions and pseudo-extinctions, can inaccurately identify adapted lineage in the presence of strong

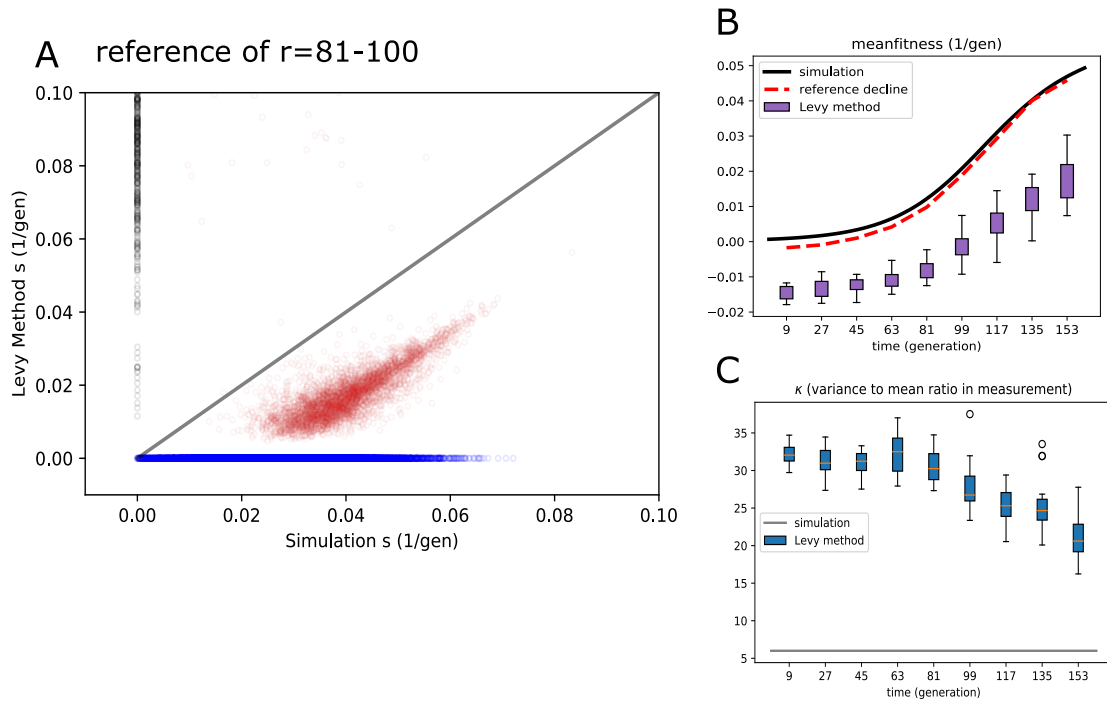


Figure 2.10. Lineage identification and fitness estimation by Levy et al. with reference chosen from 81 – 100 read count for the large-drift simulation case. (A) Fitness effect estimation for true positive (red), false positive (black) and false negative (blue) lineages. (B) mean fitness trajectory inferred by Levy et al average (purple bar), by bulk neutral decline (red dash line) compared to the simulation mean-fitness trajectory (black dash line). (C) noise parameter inferred by Levy et al. average (blue bars) compared to the simulated noise (black line)

genetic drift. Genetic drift refers to random changes in lineage frequencies due to demographic fluctuations. If genetic drift is strong, many lineages that are present and observed at one time point go extinct or drop below the detection threshold at the next sampling time point. In particular, in our strong genetic drift simulations, the probability of observing zero reads at the next time point for lineages that were observed at the previous time point is 15% – 40% (Figure 2.9 blue line), compared to just 2% – 13% in the baseline scenario (Figure 2.9 black line). To determine the consequences of ignoring these lineage extinctions and pseudo-extinctions, I selected lineages with $80 < r \leq 100$ reads as reference, inferred mean fitness and identified adapted lineages following the Levy-Blundell procedures. I find that the neutral decline assumption works well for these reference lineages, such that the mean fitness estimated from equation (2.9) is close to the true mean fitness (Figure 2.10 B, red dashed line). However, the Levy-Blundell method, infers mean fitness not from this equation directly, but rather by fitting the histograms of read counts at the next sampling time point while excluding lineages that are not observed (see Figure 2.11). As a result, it under-estimates mean fitness (Figure 2.10 B, purple bars), which results in much worse performance compared to the baseline scenario. Specifically, in this regime, the method has low 23% true positive rate, 77% precision of adapted calls, and a large 29% systematic bias in the estimates of selection coefficients of individual adapted lineages (Figure 2.10 A).

2.5 Discussion

2.5.1 The working region for the mean-fitness estimation changes over time

I have pointed out that the failure of the neutral decline assumptions due to the existence of measurement noise. It might appear that one it is easy to fix this problem by inferring mean fitness from equation (2.8) that accounts for measurement noise instead of the neutral decline equation (2.5). However, the conditional distribution $P(r_k | r_{k-1})$ depends on the the distribution of lineage sizes $P(n_{k-1})$ which is unknown and changes over time. Moreover, it is possible to



Figure 2.11. Read distribution of large drift scenario for lineages of read 81 – 100 and the inference of mean-fitness and noise based on Levy-Blundell method.

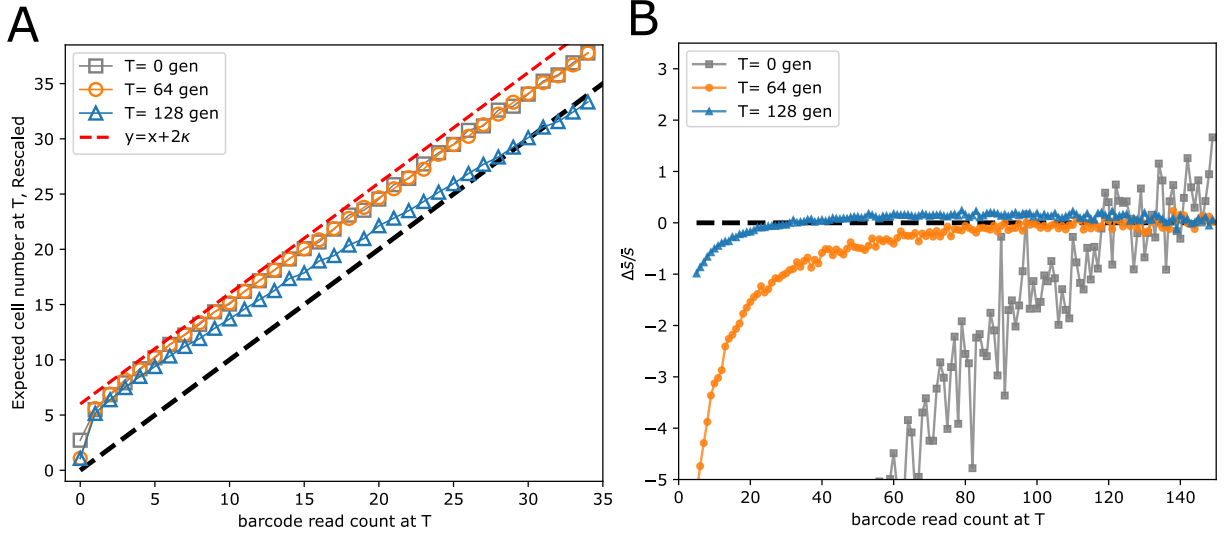


Figure 2.12. The working region for the unbiased inference of mean fitness shifts over time. Data from the baseline simulation scenario. (A) Expected cell number at T from lineages of the same barcode read count at T . (B) Error in mean-fitness estimate based on neutral decline for different simulation times. $\Delta\bar{s}$ is the absolute difference between inferred fitness effect and the true simulated fitness effect. Error is the $\Delta\bar{s}$ normalized by the simulated s .

obtain an analytical expression for $P(r_k|r_{k-1})$ only for the simplest forms of the distribution of lineage sizes $P(n_{k-1})$. Figure 2.12A shows that the expected cell number $\mathbb{E}(n_k|r_{k-1})$ follows the prediction (given by equation (A.7), red dash line) only at the beginning of the simulation but deviates from it at later times. Since the distribution of lineage sizes $P(n_{k-1})$ changes over time due to natural selection, it is in general impossible to define a fixed working region of reference lineage frequencies where the neutral decline equation (2.5) would produce consistently unbiased estimates of mean fitness (see Figure 2.12B).

2.5.2 An intuition for an offset in expected read count

I have shown that the expected read count (equation (2.8), Fig.2.4) has a nonzero offset. Here, I provide an intuitive explanation for this somewhat puzzling observation. First, note that there are no reasons to believe that there are any systematic biases in the measurement procedure. This implies that the expected frequency of a given barcode in the sequencing data equals the population frequency of the corresponding lineage, as one intuitively expects. However, the

converse is in general not true. That is, the expected population frequency of lineages with a given observed barcode frequency is in general not equal to the observed barcode frequency. This statement is obviously true for barcodes that are represented by zero reads and hence have frequency zero in the sequencing data. Clearly, the expected frequency of lineages represented by zero reads is not zero because some lineages were not sequenced by chance. The statement above is true also for barcodes represented by a non-zero number of reads. The key idea is that the conditioning on the observed barcode frequency (or, equivalently, on the read count) implies that we are calculating the expectation over multiple lineages that are present at different frequencies in the population. And this conditional expected population frequency, given the read count, depends on the distribution of lineage frequencies in the population in a non-intuitive way.

Consider the following example that illustrates this dependence. Imagine that the sequencing measurement process is such that a lineage present at frequency $x\%$ in the population is represented in the sequencing data by $x - 1$, x or $x + 1$ reads with probabilities 0.25, 0.5 and 0.25, respectively. Suppose that we would like to know the expected frequency of lineages with the observed barcode count of 1 read. This frequency clearly depends on the distribution of lineage frequencies in the population. Consider a simple scenario where we have prior knowledge that there are a total of 100 lineages in the population, and each of them is present at frequency 1%. In this case, the expected frequency of lineages represented by 1 read is of course 1%. Now consider a more complex scenario where we have prior knowledge that there are a total of 75 lineages in the population, with 50 of them (i.e., two thirds) being present at frequency 1% each and 25 of them (i.e., one third) being present at frequency 2% each. Then, we expect that there will be 31.25 lineages represented by one read, 25 of which have true population frequency of 1% and 6.25 of which have true population frequency of 2%. Thus, the expected frequency of lineages represented by 1 read is $25/31.25 \times 1\% + 6.25/31.25 \times 2\% = 1.2\%$.

2.6 Methods

2.6.1 BLT simulation

The simulation starts with $N_L = 5 \times 10^5$ barcoded lineages of which 20,000 are adapted. The selection coefficient of these adapted lineages were randomly drawn from a normal distribution with mean 3% and width ($2 \times$ standard deviation) of 2%. These numbers are chosen based on the BLT experiments in Ref. [58]. For simplicity, we do not allow any new adaptive mutations to arise in the simulation. At the initial time point, each barcode is represented by n cells with n being randomly drawn from an exponential distribution with mean of N/N_L , where the total population size is $N = 128 \times 10^8 = 25,600N_L$ for baseline scenario and $N = 4 \times 10^7 = 80N_L$ for large-drift scenario. At each cycle, we dilute the total population by the factor D ($D = 256$ for baseline scenario and $D = 8$ for strong drift scenario) to the bottleneck size N_b ($N_b = 100N_L$ for baseline scenario and $N_b = 10N_L$ for strong-drift scenario). After dilution, each lineage increases or decreases deterministically based on its s and the instant mean-fitness $\bar{s}_t$. The dilution-growth cycles are repeated for ~ 160 generations. In our simulation, the mean fitness increases to 4% at the end (which is comparable to 5% at 120 generation in Ref. [58]). Approximately every 16 generations, the barcodes are randomly sampled from each lineage. We model the sampling (barcode measurement) using the negative binomial distribution with $2\kappa = 6$, consistent with the value presented in Ref. [58].

Acknowledgments

Chapter 2, in full is currently being prepared for submission for publication of the material. Huan-Yu, Kuo; Sergey, Kryazhimskiy. The dissertation author was the primary researcher and author of this material.

Chapter 3

Bayesian Filtering Method for Estimating Fitness of Nascent Adaptive Mutants in BLT Experiments

3.1 Introduction

As discussed in the previous chapter, the only existing (Levy-Blundell) method for the BLT data analysis has two serious limitations. One limitation arises from the fact that the Levy-Blundell method ignores lineage extinctions. The source of its second limitation is that it approximates the dynamics of lineage frequencies in the population of barcoded cells with the dynamics of measured barcode frequencies in the sequencing data. I have shown that these limitations can decrease the ability to identify adapted lineages and cause systematic biases in the estimates of selection coefficients of adapted lineages. My analysis of these limitations suggests that it may be possible to improve the mean-fitness estimate (and, consequently, the estimates of fitness effects of all lineages) by (a) taking lineage extinctions into account, and (b) explicitly modeling the true lineage frequency dynamics. However, implementing the second idea is challenging because the true cell lineage frequencies are unknown. To overcome this challenge, I construct a Bayesian statistical model, which include cell lineage sizes and their fitness effects as unknown variables and the corresponding barcode read counts as observables. I develop a Bayesian framework to estimate the unknown variables for each lineages. The

posterior distributions for the unknown variables for all lineages are updated over time, resulting in what's called the Bayesian filtering method.

3.2 The Bayesian filtering method

3.2.1 Overview of the Bayesian filtering method

I set out to develop a more robust statistical method, the Bayesian filtering method, for identifying adapted lineages and inferring the fitness effects of nascent adaptive mutations. The key object of the method is the belief probability for the selection coefficient and the lineage size of each lineage. For each lineage, I model how the observable variable, its barcode read count in the sequencing data, changes over time based on its hidden variables (selection coefficient and cell lineage size). I cast this model within the Bayesian framework and compute the posterior belief probability for the fitness effect of each lineage. Starting with an uninformed prior belief probability, this method incorporates the barcode data from each subsequent sampling time point to continuously update the belief distribution for the fitness of each lineage and its size (see Section 3.2.2).

To identify which lineages are adapted (i.e., have acquired an adaptive mutation), I develop two models. The N-model models the case where the focal lineage has not yet acquired a beneficial mutation, so that it has no selective advantage over the ancestor. The S-model models the case where the focal lineage has acquired an adaptive mutation, so that the lineage has a selective advantage over the ancestor. I then evaluate which model is better supported by the data by computing the Bayes factor. Within each time interval, I classify lineages as putatively adapted or neutral. Lineages from both classes are randomly selected to estimate the mean-fitness of the next time interval (see Section 3.2.3). I take the inferred mean-fitness as a constant parameter in the Bayesian model to continue updating the Bayesian probabilities of individual lineages. After analyzing the entire time series in this way, I call a lineage as adapted if it has been classified as putative adapted for at least two time points. For lineages called as

Table 3.1. Constant parameters used in the Bayesian filtering model

T_k	time (in generation) of sampling time point k
R_k	total read depth at sampling time point k
$\bar{s}_k$	the mean selection coefficient in the population (“mean fitness” for short) in the time interval $[T_{k-1}, T_k]$
ε	measurement error parameter in the barcode measuring process (equation (3.9))
D	dilution ratio
N_b	bottleneck population size (after dilution)
N	population size at the end of growth cycle (before dilution) $= DN_b$

adapted, I then estimate their selection coefficients and the uncertainty around them based on the current belief distributions (see Section 3.2.4).

3.2.2 Stochastic models for updating lineage dynamics

I build up two stochastic models, the **S-model** and the **N-model**, to classify adapted lineages and to update the understanding of lineage dynamics (Table 3.2) over time. I consider three important variables for each lineage: its selection coefficient, its population size, and the barcode read counts within two consecutive time steps $k - 1$ and k . I model the relationship between those variables with the S- and N-models. The only difference between the N- and S-models is that the selection coefficient s is a constant and equal to zero in the N-model, whereas in the S-model s is a random variable. I build up the N-model for the purpose of lineage classification: if a lineage behaves neutrally (as if $s = 0$), the dynamics of its barcode read counts can be explained by both models equally well; if a lineage behaves as adapted (as if $s > 0$), the dynamics of its barcode read counts will be explained much better by the S-model than by the N-model. Table 3.1 lists the constant parameters used in the two stochastic models.

Table 3.2. Notation of probabilities and variables used in the Bayesian filtering model

Notation	Description	Eqn.
s_i	selection coefficient of lineage i	
n_{ik}	population size of lineage i at T_k after growth	
r_{ik}	barcode count of lineage i at T_k	
$\mathbf{r}_{ik}$	vector of barcode counts of lineage i between T_0 and T_k , $\mathbf{r}_{ik} = (r_{ik}, r_{ik-1}, \dots, r_{i0})$	
Variable indices i, k are reduced inside the probability bracket and moved outside.		
$P_{ik}^{\text{pr}}(n, s \mathbf{r}_{k-1})$	joint belief probability for n_{ik}, s_i for lineage i prior to observation at time T_k within the S-model	[3.1]
$P_{ik}^{\text{pr}}(n \mathbf{r}_{k-1})$	belief probability for n_{ik} for lineage i prior to observation at time T_k within the N-model	[3.4]
$P_{ik}^{\text{b}}(n, s \mathbf{r}_{\mathbf{k}})$	posterior belief probability for n_{ik}, s_i for lineage i after observation at time T_k within the S-model	[3.2]
$P_{ik}^{\text{b}}(n \mathbf{r}_{\mathbf{k}})$	posterior belief probability for n_{ik} for lineage i after observation at time T_k within the N-model	[3.5]
$Z_{ik}^{\text{S}}(r \bar{s})$	marginal likelihood of data for lineage i at time T_k under the S-model, i.e., the probability of r_{ik} within the S-model	[3.3]
$Z_{ik}^{\text{N}}(r \bar{s})$	marginal likelihood of data for lineage i at time T_k under the N-model, i.e., the probability of r_{ik} within the N-model	[3.6]
$P_k^{\text{tr}}(n_k n_{k-1}; s_k, \bar{s})$	Transitional probability , i.e., the probability for a lineage with selection coefficient s_k whose size was n_{k-1} at T_{k-1} to have size n_k at time T_k	[3.8]
$P_k^{\text{m}}(r n; \varepsilon, R)$	Measurement probability , i.e., the probability of observed barcode read count r for a lineage with size n given the total read depth R_k	[3.9]
$L_k(\bar{s})$	log-likelihood function of mean-fitness $\bar{s}_k$	

S-model

Consider a lineage of size n_{k-1} and selection coefficient s at time T_{k-1} . I do not have direct access to n_{k-1} . Instead, I sample the population and obtain the barcode read count r_{k-1} for this lineage. I describe this stochastic **measurement process** in Section 3.2.2. At the same time, the population is diluted to a bottleneck size N_b , which I model by randomly sampling N_b individuals from the total population of size N . After dilution, the focal lineage deterministically grows to size n_k at time T_k . The stochastic process that maps n_{k-1} onto n_k , which includes dilution and growth, is called the **transition process** and is described in Section 3.2.2. The S-model is schematically depicted in Figure 3.1A, where oval nodes represent hidden variables (the selection coefficient s and the population size n), rectangular nodes represent the observables (barcode read counts), and the links represent the stochastic processes that map between the variables.

As mentioned above, I consider the belief distribution for hidden variables, and update it after each observation as follows. I begin with a belief distribution $P_{k-1}^b(n_{k-1}, s | \mathbf{r}_{k-1})$ which describes our knowledge of hidden variables at time T_{k-1} , where $\mathbf{r}_{k-1} = \{r_{k-1}, r_{k-1}, \dots, r_0\}$ is a set observed data up to T_{k-1} (see Table 3.2 for a list of notations). With the belief at T_{k-1} , I can make a guess of hidden variables at the next time T_k using the transitional process. I refer to this guess as the prior belief distribution $P_k^{\text{pr}}(n_k, s | \mathbf{r}_{k-1})$ at T_k which is defined as

$$P_k^{\text{pr}}(n_k, s | \mathbf{r}_{k-1}) = \int dn_{k-1} P_k^{\text{tr}}(n_k | n_{k-1}; s, \bar{s}) P_{k-1}^b(n_{k-1}, s | \mathbf{r}_{k-1}), \quad (3.1)$$

where $P_k^{\text{tr}}(n_k | n_{k-1}; s, \bar{s}_k)$ is the transitional probability distribution (see Table 3.2), and $\bar{s}_k$ is the mean fitness, which is for now assumed as known. Note that our belief of the selection coefficient of the lineage is not changed during the transition, i.e., $P_k^{\text{pr}}(s | \mathbf{r}_{k-1}) = P_{k-1}^b(s | \mathbf{r}_{k-1})$. The transition only updates the population size from the previous time $k-1$ to the next time k . Once the prior distribution is obtained, I can update the knowledge of hidden variables using the new observed data, r_k . The updating from the prior to the posterior belief $P_k^b(n_k, s | \mathbf{r}_k)$ is given

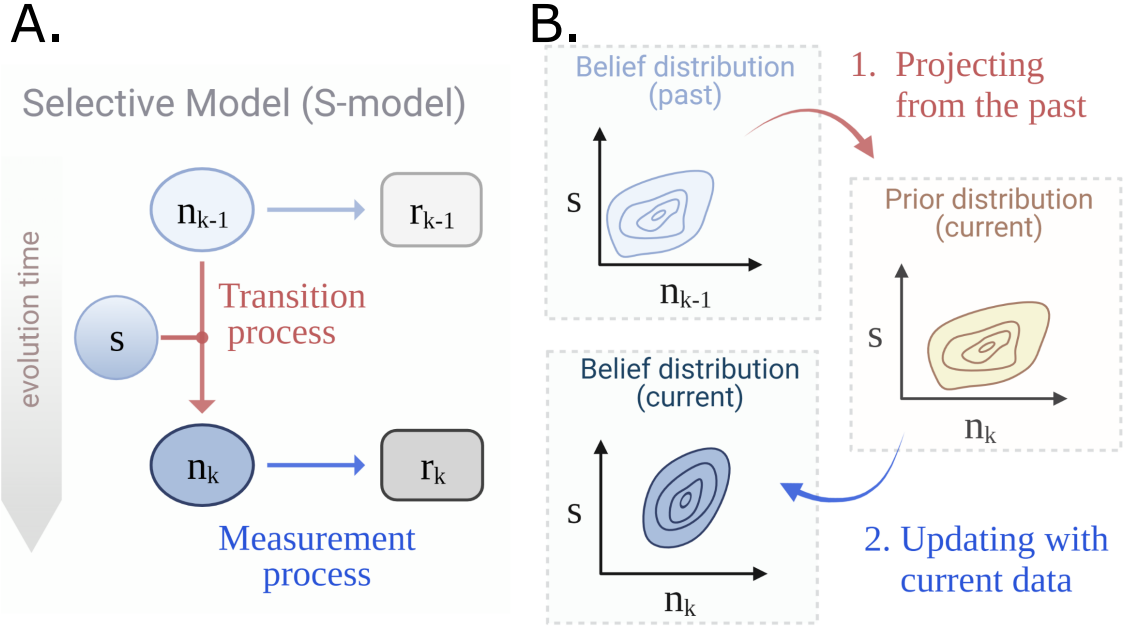


Figure 3.1. S-model. (A) Structure of the model for an individual lineage. Ovals represent lineage's hidden variables, selection coefficient s and the lineage population sizes n_{k-1} , n_k at sampling time points T_{k-1} and T_k , respectively. Rounded squares represent the observable variables r_{k-1} , r_k , which are the barcode read counts at sampling time points T_{k-1} and T_k , respectively. Arrows represent dependencies between the variables. The transitional process is a Markovian process that models how lineage size changes over time. Measurement process models how the observable read count depends on the lineage population size. (B) Model iteration. Each lineage at each time is characterized by the current belief probability distribution for the hidden variables. The belief distribution at the next time point is generated by a two-step process from the belief distribution at the previous time point. At the first step, the past belief distribution is projected forward using the transition process to obtain the prior distribution at the next time point. At the second step, the prior distribution is updated using the Bayes rule with the current observable data to produce the current belief distribution.

by

$$P_k^b(n_k, s | \mathbf{r}_k) = \frac{P_k^{\text{pr}}(n_k, s | \mathbf{r}_{k-1}) P_k^{\text{m}}(r_k | n_k)}{Z_k^S(r_k)}, \quad (3.2)$$

where $P_k^{\text{m}}(r_k | n_k)$ is the measurement probability (see Table 3.2). The $P_k^{\text{m}}(r_k | n_k)$ is also a likelihood function of n_k in Equation 3.2. The denominator $Z_k^S(r_k)$ is a normalization constant.

$$Z_k^S(r_k) = \int ds \int dn_k P_k^{\text{pr}}(n_k, s | \mathbf{r}_{k-1}) P_k^{\text{m}}(r_k | n_k). \quad (3.3)$$

I label this constant with the super-index S , which stands for the S-model, because $Z_k^S(r_k)$ is also the total probability for a lineage to generate the observed read count r_k under the S-model. Importantly, our knowledge of both the selection coefficient and lineage size are updated at this step. Thus, I started with the past belief distribution $P_{k-1}^b(n_{k-1}, s | \mathbf{r}_{k-1})$ and updated it to the new belief distribution $P_k^b(n_k, s | \mathbf{r}_k)$. I then move to the next time interval and repeat this procedure. Therefore, the knowledge of a lineage's selection coefficient and size are iteratively updated at each observation.

N-model

The N-model is identical to the S-model with the single exception that the selection coefficient of the lineage is fixed to be zero. As a result, there is only one hidden variable, lineage population size n_k (see Figure 3.2). Thus, equations (3.1)–(3.3) become

$$P_k^{\text{pr}}(n_k | \mathbf{r}_{k-1}) = \int dn_{k-1} P_k^{\text{tr}}(n_k | n_{k-1}; s = 0, \bar{s}_k) P_{k-1}^b(n_{k-1} | \mathbf{r}_{k-1}), \quad (3.4)$$

$$P_k^b(n_k | \mathbf{r}_k) = \frac{P_k^{\text{pr}}(n_k | \mathbf{r}_{k-1}) P_k^{\text{m}}(r_k | n_k)}{Z_k^N(r_k)}, \quad (3.5)$$

$$Z_k^N(r_k) = \int dn_k P_k^{\text{pr}}(n_k | \mathbf{r}_{k-1}) P_k^{\text{m}}(r_k | n_k), \quad (3.6)$$

respectively (see Table 3.2 for notations).

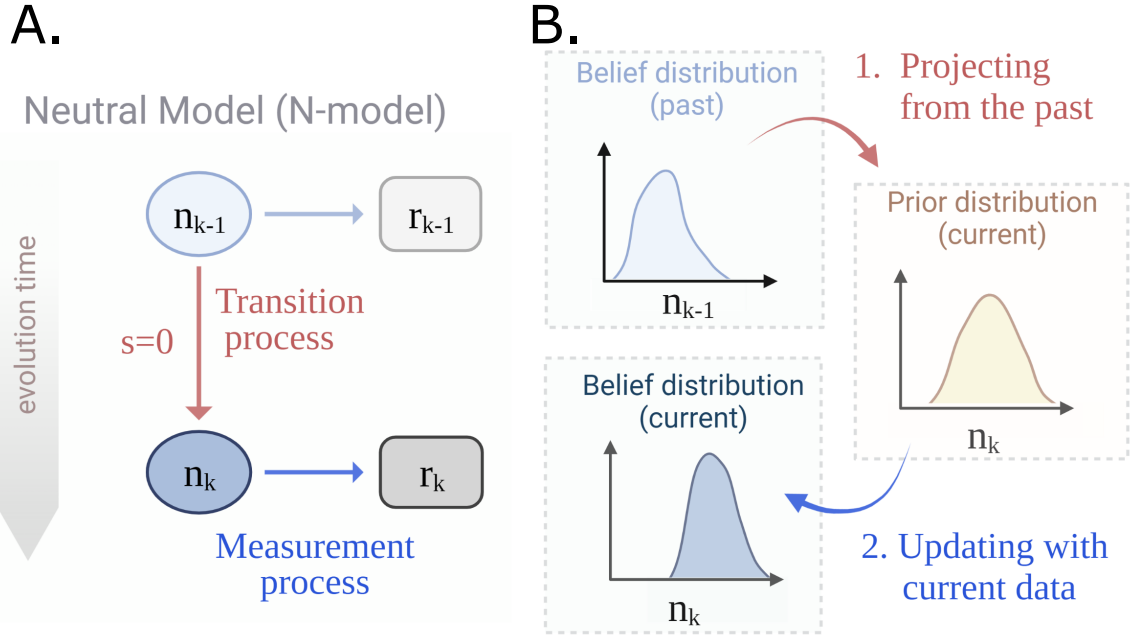


Figure 3.2. N-model. Notations are as in Figure 3.1. The N-model differs from the S-model only in that, in the former, the selection coefficient is fixed and equal to zero, so that there is only one hidden variable, lineage population size.

Transition process

Consider a barcoded lineage that has fitness s relative to the ancestor and that is represented by n_{k-1} individuals in the population at time T_{k-1} . On average, such lineage will expand or contract depending on whether s is larger or smaller than the mean fitness $\bar{s}$. Demographic noise will also cause its size to fluctuate randomly around this expectation, such that the size n_k of the lineage at the next measurement time T_k is described by the transition probability $P_k^{\text{tr}}(n_k | n_{k-1}; s, \bar{s})$. I model this transition probability in a batch-culture regime as follows. As it is common in such experiments, I assume that sampling happens immediately prior to a dilution step. During the dilution, a fraction $1/D$ of the population is transferred into fresh media, and the rest is discarded, where D is the dilution factor so that the size of the focal lineage becomes

n'_{k-1} , which I draw from the Poisson distribution with mean n_{k-1}/D :

$$P(n'_{k-1}|n_{k-1}) = \frac{1}{\Gamma(1+n'_{k-1})} \left(\frac{n_{k-1}}{D}\right)^{n'_{k-1}} e^{-\frac{n_{k-1}}{D}}. \quad (3.7)$$

I assume that after the dilution, the lineage grows or shrinks exponentially and deterministically for the cycle duration Δt , such that by the end of the cycle, its size becomes $n''_{k-1} = D n'_{k-1} e^{(s-\bar{s}_k)\Delta t}$. The transition probability of n''_{k-1} after one cycle of growth becomes

$$\begin{aligned} P(n''_{k-1}|n_{k-1}) &= P(n'_{k-1}|n_{k-1}) \left| \frac{dn'_{k-1}}{dn''_{k-1}} \right| \\ &= \frac{1}{\Gamma(1+n'_{k-1})} \left(\frac{n_{k-1}}{D}\right)^{n'_{k-1}} e^{-\frac{n_{k-1}}{D}} \left| \frac{dn'_{k-1}}{dn''_{k-1}} \right| \\ &= D \gamma(n_{k-1}; \tilde{k}, \theta) \left| \frac{dn'_{k-1}}{dn''_{k-1}} \right| \\ &= e^{-(s-\bar{s}_k)\Delta t} \gamma(n_{k-1}; \tilde{k}, \theta), \end{aligned} \quad (3.8)$$

where the one-cycle transition probability is also the likelihood function of n_{k-1} which is proportional to a gamma distribution $\gamma(n_{k-1}; \tilde{k}, \theta)$ with shape parameter $\tilde{k} = 1 + \frac{n''_{k-1}}{D} e^{-(s-\bar{s}_k)\Delta t}$ and scale parameter $\theta = D$. If the next sampling point is the end of this cycle, then $P_k^{\text{tr}}(n_k | n_{k-1}; s, \bar{s}) = P(n''_{k-1}|n_{k-1})$. Otherwise, I repeat the one-cycle transition until the next sampling time point T_k is reached. Importantly, this transition process model allows for the possibility of lineage extinction during dilution.

Measurement process

The actual barcode lineage sizes are not observable. Instead, the population is sampled at discrete time points. After undergoing a series of processing steps, these samples are sequenced at the barcode locus, and the reads carrying each barcode are counted. This measurement process can be described by a probability distribution $P_k^{\text{m}}(r | n; \varepsilon, R)$ of observing r reads for a barcode lineage that is present at size n in the population at time point k , where R is total reads and ε is a

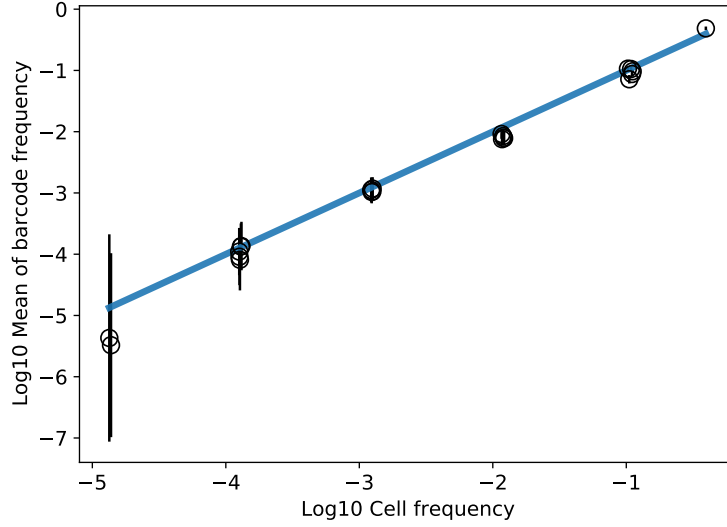


Figure 3.3. Measured barcode frequency in the sequencing data (y axis) on average faithfully represents the cell lineage frequency in the population (x axis). Each point represents a unique barcode, blue line shows the 1-to-1 line.

measurement noise parameter defined below.

The simplest model for such a measurement process is the Poisson distribution with mean $\mu_r = Rn/N$, where N denotes the total population size. Importantly, it captures the possibility of a lineage pseudo-extinction, i.e., failure to observe a lineage even if it is present in the population. However, sequencing read counts are often overdispersed with respect to the Poisson distribution [9], likely because the sample processing steps involve exponential amplification of template molecules [70, 87]. Thus, I model the measurement process with a negative binomial (NB) distribution with mean $\mu_r = Rn/N$ and variance σ_r^2 . The NB distribution is mathematically convenient, it is widely used to model read counts [10, 67, 84], and it contains the Poisson distribution as a special case when the variance equals the mean, $\sigma_r^2 = \mu_r$.

To estimate the variance σ_r^2 of the NB distribution, I performed a calibration experiment. Briefly, I assembled a population of *Saccharomyces cerevisiae* with a total size of 1.6×10^7 individuals that consisted of 26 barcoded sub-populations present at five different frequencies (from 10^{-1} to 10^{-5}) and sequenced it with 10 replicates to an average of 2×10^5 reads per

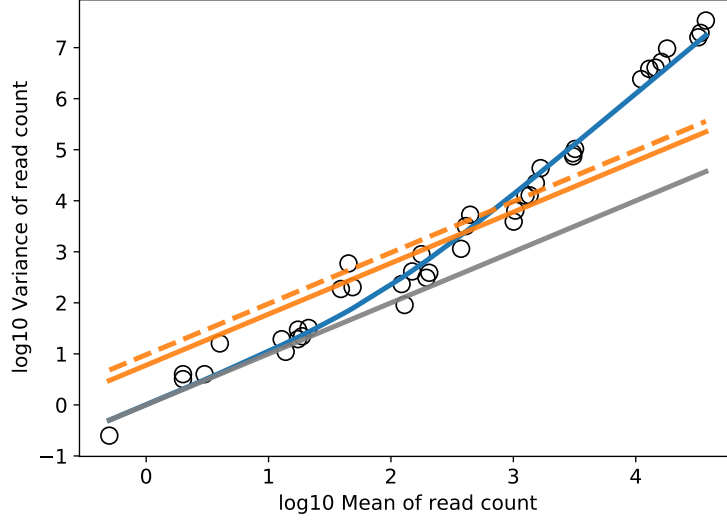


Figure 3.4. Variance in the reads count versus the mean read count. Same data as in Figure 3.3. Each point represents the data form barcodes of same initial frequency. Blue line shows the best-fit of quadratic function ($\sigma^2 = \mu(1 + \varepsilon\mu)$ with $\varepsilon = 0.012$, $R^2 = 0.98$) to the data. Grey line is the expectation ($\sigma^2 = \mu$, $R^2 = 0.61$) from Poisson distributed measurement process. Solid orange line is the Levy-Blundell ($\sigma^2 = 2\kappa\mu$, $\kappa = 3$, $R^2 = 0.81$). Dash orange line is the best-fit of linear function ($\sigma^2 = 2\kappa\mu$, $\kappa = 4.8$, $R^2 = 0.82$).

replicate (see section 3.5.1 for details).

As expected, the mean of the barcode frequency estimated from sequencing data equals the true frequency of each lineage (see Figure 3.3). To test the extent of over-dispersion, I plotted the variance of the measured barcode frequency against the mean. I find that the variance is a non-linear, convex function of the mean (Figure 3.4). I capture this nonlinearity with a model $\sigma_r^2 = \mu_r + \varepsilon\mu_r^2$, where ε is a measurement noise parameter that controls the degree of over-dispersion of the NB distribution. I estimate $\varepsilon \sim 0.01$. Finally, I specify the measurement process as a NB distribution with this parameter, which is given by

$$\begin{aligned}
 P_k^m(r|n; \varepsilon, R) &= \text{NB}(r; \mu_r = Rn/N, \sigma_r^2 = \mu_r + \varepsilon\mu_r^2) \\
 &= \frac{\Gamma(r+1/\varepsilon)}{\Gamma(r+1)\Gamma(1/\varepsilon)} \left(\frac{1}{1 + \varepsilon\mu_r} \right)^{1/\varepsilon} \left(\frac{\varepsilon\mu_r}{1 + \varepsilon\mu_r} \right)^r.
 \end{aligned} \tag{3.9}$$

3.2.3 Estimating the mean fitness of the population

Putative lineage class and initial belief distribution

I now consider the statistical problem of calling the focal lineage as adapted, based on its change in frequency during a single sampling interval $[T_{k-1}, T_k]$, $k = 1, 2, \dots$. I approach this problem within the Bayesian framework. I consider two models, the N-model, where a given lineage has no selective advantage over the ancestor, i.e., $s = 0$, and the S-model, where the lineage can have selective advantage $s > 0$ (see Section 3.2.2), and evaluate which model is better supported by the data obtained at time point T_k . To do so, I evaluate the Bayes factor B_k for a lineage at time point T_k ($k \geq 1$) as

$$B_k = \frac{Z_k^S(r_k)}{Z_k^N(r_k)}, \quad (3.10)$$

where $Z_k^S(r_k)$ (Equation 3.3) and $Z_k^N(r_k)$ (Equation 3.6) are the probabilities that r_k reads are observed at the k th sampling time within the N- and S-models, respectively. The Bayes factor is expected to be 1 for neutral lineages and greater than 1 for adapted lineages. Thus, I classify lineages into three categories, putatively adapted (ADP), putatively neutral (NEU) and unknown (UNK) at each time point.

Starting at time point T_1 , I mark all i lineages putatively neutral (NEU) and define the past belief distribution $P_{i0}^b(n, s) = P_{i0}(n)P_0(s)$, where I choose $P_{i0}(n)$ to be a gamma distribution with the shape parameter $r_0^i / (1 + r_0^i \times 0.01)$ and scale parameter $N_0(1 + r_0^i \times 0.01) / R_0$, such that its mean is $N_0 r_0^i / R_0$, and I choose the prior distribution $P_0(s) = U(s)$ to be a continuous uniform distribution bounded by (s_{min}, s_{max}) , where the boundaries are reasonable s value (per cycle) for adapted mutations that are set by users.

For each putatively neutral lineage i , I calculate its belief probabilities within both N and S models. Since the lineage is assumed to be neutral, I assign its past belief of selection coefficient to be uniform $P_{ik}^b(s) = U(s)$ for all $k \geq 1$. A currently putatively neutral lineage

Table 3.3. Putative lineage class and initial belief distribution

Putative lineage class at time k	Condition if	Putative lineage class at time k+1	Run Bayesian model at k+1
UNK	$r_{k+1} \leq r_c$	UNK	N Model
	$r_{k+1} > r_c$	NEU	N Model, S Model ($P_{k+1}^{\text{pr}}(s) = U(s)$)
NEU	$r_{k+1} \leq r_c$	UNK	N Model
	$r_{k+1} > r_c \ \& \ B_k \leq 1$	NEU	N Model, S Model ($P_{k+1}^{\text{pr}}(s) = U(s)$)
	$r_{k+1} > r_c \ \& \ B_k > 1$	ADP	N Model, S Model ($P_{k+1}^{\text{pr}}(s) = U(s)$)
ADP	$r_{k+1} \leq r_c$	UNK	N Model
	$r_{k+1} > r_c \ \& \ B_k \leq 1$	NEU	N Model, S Model ($P_{k+1}^{\text{pr}}(s) = U(s)$)
	$r_{k+1} > r_c \ \& \ B_k > 1$	ADP	N Model, S Model ($P_{k+1}^{\text{pr}}(s) = P_k^{\text{b}}(s)$)

can be classified as unknown if $r_{k+1} \leq r_c$ at the next time point T_{k+1} , or as putatively adapted if $r_{k+1} > r_c$ and $B_{ik} > 1$, or as putatively neutral if $r_{k+1} > r_c$ and $B_{ik} < 1$ (see Table 3.3).

For each lineage marked as putatively adapted, I also calculate its belief probabilities within both N and S models. For such a lineage, I use the current belief distribution with respect to its selection coefficient $P_{ik}^{\text{b}}(s)$ as the prior for the next time point, i.e. $P_{k+1}^{\text{pr}}(s) = P_k^{\text{b}}(s)$. A lineage currently classified as putatively adapted can be then classified as unknown if $r_{k+1} < r_c$ at T_{k+1} , or as putatively adapted if $r_{k+1} > r_c$ and $B_{ik} > 1$, or as putatively neutral if $r_{k+1} > r_c \ \& \ B_{ik} < 1$ (see Table 3.3).

Finally, a lineage currently classified as unknown can be classified as either unknown $r_{k+1} \leq r_c$ or as putatively neutral if $r_{k+1} > r_c$ (see Table 3.3).

Estimating the mean fitness effect

To estimate mean fitness $\bar{s}_k$ in the interval $[T_{k-1}, T_k]$ for $k \geq 1$, I randomly select a subset N_k of lineages classified as putatively neutral and a subset A_k of lineages classified as putatively adapted. I found that choosing lineages from both groups ensures a more robust estimation of mean fitness even when the majority of neutral lineages go extinct. I then estimate s by

maximizing the log-likelihood function $L(\bar{s}_k)$.

$$L(\bar{s}_k) = \sum_{i \in N_k} \log Z_{ik}^N(r_k^i; \bar{s}_k) + \sum_{i \in A_k} \log Z_{ik}^S(r_k^i; \bar{s}_k). \quad (3.11)$$

3.2.4 Identifying adapted lineages and estimating their fitness effects

As iteratively applying the procedure described above for all lineages i and all time points k , I record the vector of Bayes factors B_{ik} for all lineages and use this information to carry out the final identification of adapted lineages and their fitness effects. I identify a lineage i as adapted if its B_{ik} exceeds a threshold value B_c for at least K time points. I determine the optimal values of B_c and K using simulated BLT data that are described in Section 3.5.3. To determine the fitness effects of an adapted lineage i , I harvest all belief distributions $P_{ik}^b(s)$ for all time point T_k when $B_{ik} > B_c$. I then estimate the fitness effect of this lineage using the belief distribution $P_{ik}^b(s)$ for the time point with the smallest variance.

3.3 Results

3.3.1 Receiver Operating Characteristic (ROC) curve for the Bayesian Filtering method

I first assessed the performance of this method using simulated data where the ground truth (the number of adapted lineages and their selection coefficients) is known. Specifically, I simulated a BLT experiment under strong selection (such that mean fitness of the population reached 12% after 140 generations of evolution, compared to 5% after 120 generations in the BLT experiment described in Ref.[58], strong genetic drift (bottleneck size = 10 per barcode, compared to 140 in the BLT experiment described in Ref.[58]) and small coverage (coverage 10x per individual, compared to 100x in the BLT experiment described in Ref.[58]). The lineage frequency trajectories, as well as the ground truth distribution of fitness effects of adapted lineages and the ground truth mean fitness of the population are shown in Figure 3.5. As discussed in

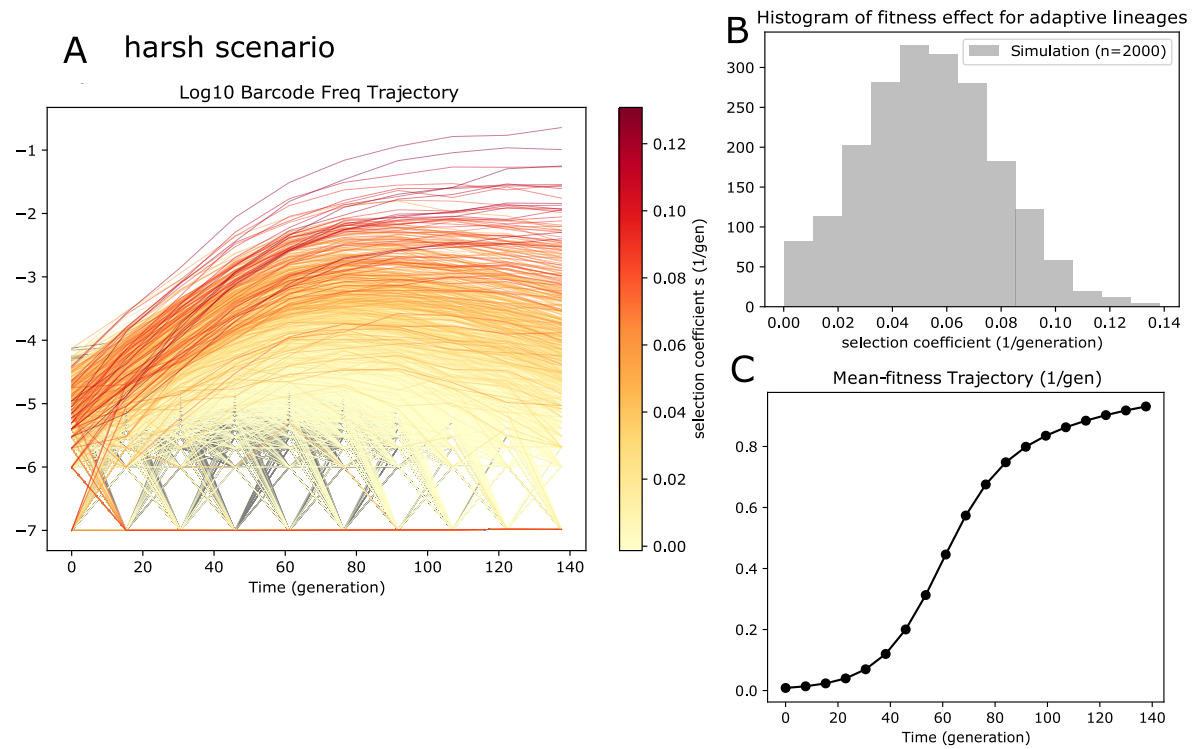


Figure 3.5. Simulation of a BLT experiment under strong selection, strong genetic drift and small coverage for assessing the performance of the Bayesian Filtering method. (A) Barcode frequency trajectories (B) The ground truth distribution of fitness effects of adapted lineages. (C) The ground truth mean-fitness trajectory.

Chapter 2, I expect that the Levy-Blundell method would perform poorly under these conditions.

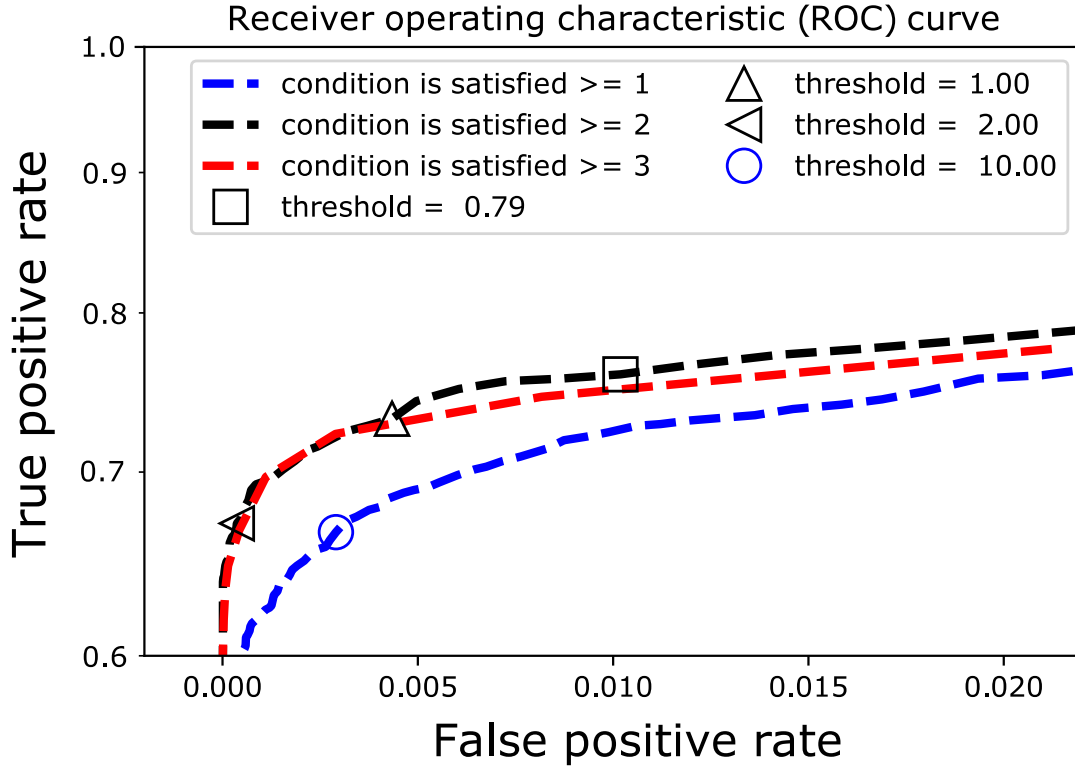


Figure 3.6. Receiver operating characteristic (ROC) curve for the identification of adapted lineages for $K = 1$ (blue curve), $K = 2$ (black curve) and $K = 3$ (red curve). K is the minimum number of time intervals at which a lineage's Bayes factor exceeds the specified threshold B_c . The following values of B_c are indicated by symbols: $B_c = 0.79$ (square), $B_c = 1$ (upright triangle), $B_c = 2$ (sideways triangle), $B_c = 10$ (circle).

The Bayesian filtering method has two free parameters for calling a lineage as adapted (see Section 3.2.4). Recall that I call a lineage adapted if the Bayes factor in favor of the S-model exceeds a threshold B_c at K or more time intervals. Thus, the first parameter B_c controls the strength of evidence in favor of the S-model at any one time interval. The second parameter K controls how consistent the evidence is across time points. Higher values of B_c and K should reduce the number of false positives (i.e., neutral lineages falsely identified as adapted), albeit at the expense of increasing the number of false negatives (i.e., truly adapted lineages not identified as such). To determine the optimal value of these parameters, I constructed three receiver

operating characteristic (ROC) curves, one for each value of $K = 1, 2, 3$, by varying B_c between 0.6 and 100. Each ROC curve shows how the fraction of true positives (i.e., adapted lineages that were identified as such) and the fraction of false positives (i.e., neutral lineages falsely identified as adaptive by my procedure) vary depending on the choice of the threshold B_c for a given fixed value of K (Figure 3.6).

I find that the value of $K = 2$ produces a ROC curve that is closer to optimal than either of the values of $K = 1$ or $K = 3$ in the sense that it results in higher rate of true positives for the same rate of false negatives. To determine the optimal value of B_c , notice that in general false positives represent a more severe problem than false negatives. This is because it's expected that the vast majority of lineages ($> 95\%$) will likely not gain any adaptive mutation in realistic BLT experiments [58]. In simulations, it had 1,836 adapted lineages and 98,000 non-adapted lineages. In this scenario, even if all adapted lineages were identified, the number of true positives would exceed the number of false positives (i.e., false discovery rate $\text{FDR} \geq 50\%$) only when the false positive rate is below $\sim 2\%$. In practice, I find that $\text{FDR} = 2\%$ is achieved at $B_c = 2.4$ for $K = 2$ (Figure 3.6,). Balancing the desire to maintain a high true positive rate and a low false positive rate, it's suggested using the Bayes factor threshold $B_c = 2$ which in my simulations achieves a true positive rate of 65% (I identify 1196 out of 1836 adapted lineages) and false positive rate of 0.014% (I falsely identify 14 neutral lineages as adapted), resulting in $\text{FDR} = 1.16\%$). Under these parameters, I identify a total of 1836 lineages as adapted, of which 1196 are in fact adapted lineages and 640 are in fact neutral.

3.3.2 Accuracy of fitness estimates of the Bayesian Filtering method

I next analyzed how accurately and precisely my method estimates the selection coefficients of lineages that were identified as adapted. It's found that, overall, the selection coefficient estimated by this method accurately represent the underlying true values, for the true positive lineages (Figure 3.7A). Specifically, it's found that my method produces 0.3% errors in the inferred selection coefficients) When I apply the Levy-Blundell method to the same simulated

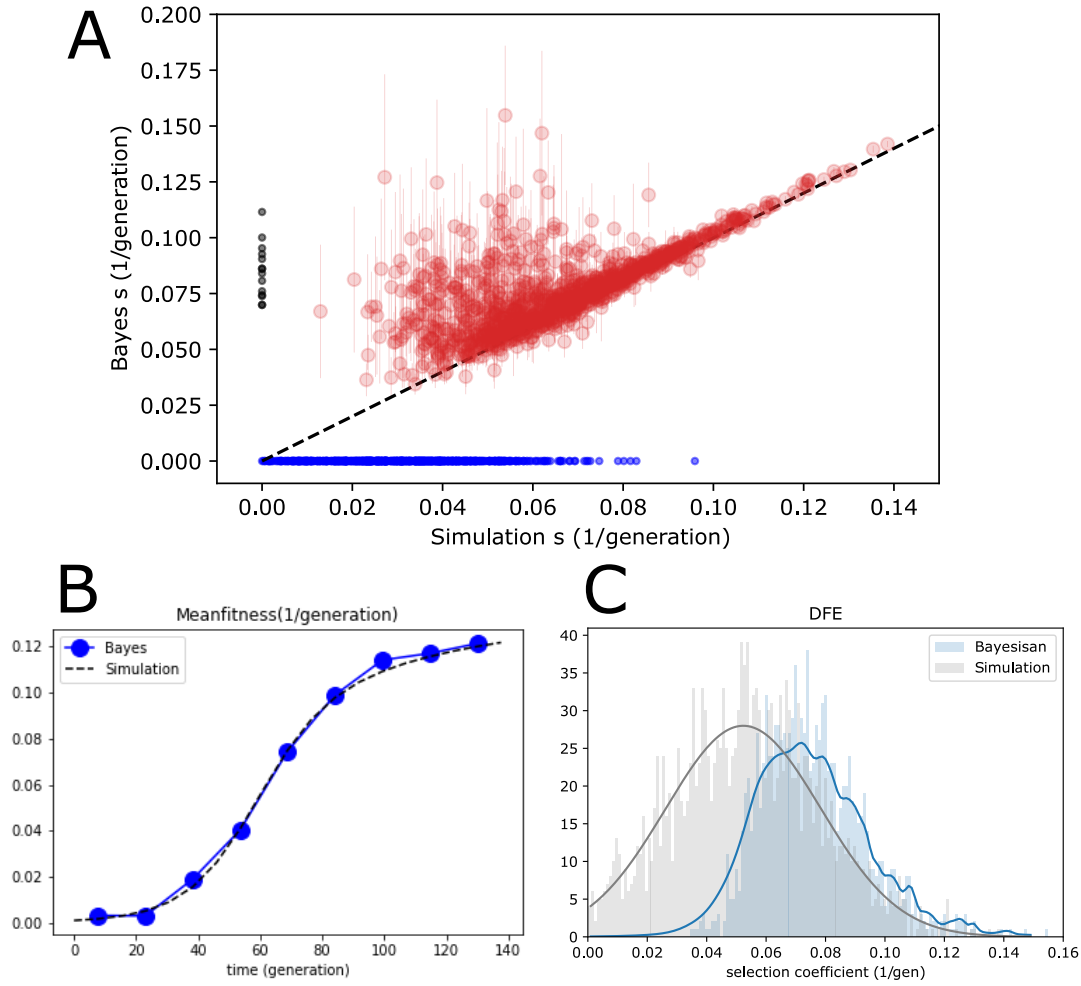


Figure 3.7. Accuracy of the Bayesian filtering method on simulated data. (A) Estimated fitness effects. True positives are shown with red circle; false positives are shown with black circles; and false negative are shown with blue circles. (B) True (black curve) and inferred (blue curve) mean-fitness trajectory (C) Inferred DFE (blue bars and curve) compared to the true DFE in the simulation (grey bars and line).

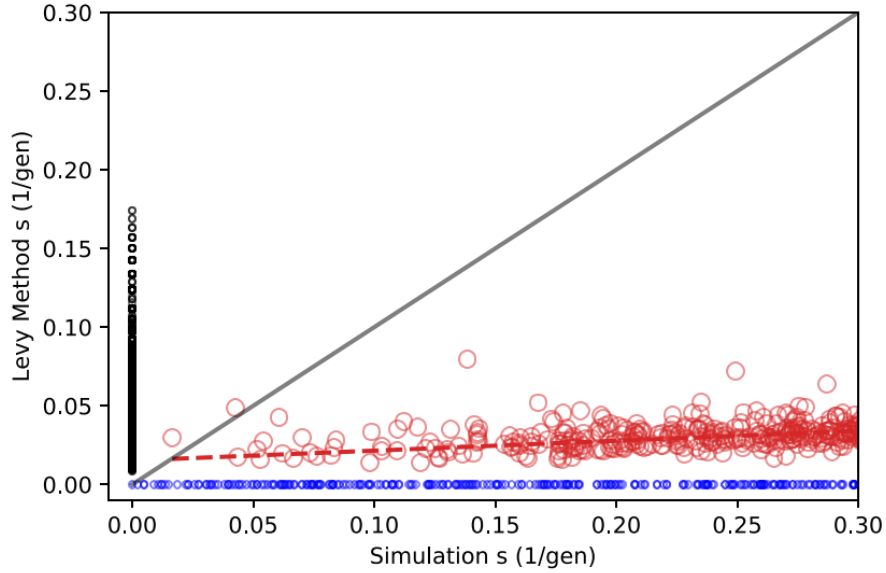


Figure 3.8. Accuracy of the Levy-Blundell method on simulated data. Notations are the same as in Figure 3.7.

data, and obtain low 15% precision in identifying adapted lineage and a strong bias in the estimates of selection coefficients, with a typical selection coefficient being underestimated by 18.7% (Figure 3.8).

Next, I compared the distribution of selection coefficients of adapted lineages (what I refer to as the “measured DFE”) to the true underlying DFE. I find that the measured DFE is right-shifted relative to the DFE. This is expected because lineages with small fitness effects are less likely to reach large size and be detected before they begin to decline due to clonal interference, i.e., $sf_{\max}N < 1$, where $f_{\max}$ and N are maximum frequency and effective population size. Indeed, I find that $sf_{\max}N_b \ll 1$ for the majority of false negative lineages (see Figure 3.9 where most blue dots are above the diagonal), while $sf_{\max}N_b \gg 1$ for the majority of true positive lineages (most red dots are below the diagonal). In other words, the trajectories of false negative (adapted lineages not identified as adapted) are similar to the trajectories of true negatives (neutral lineages identified as such), while true positives (adapted lineages identified as such) survive longer, reach higher frequencies and begin to decline later (Figure 3.10).

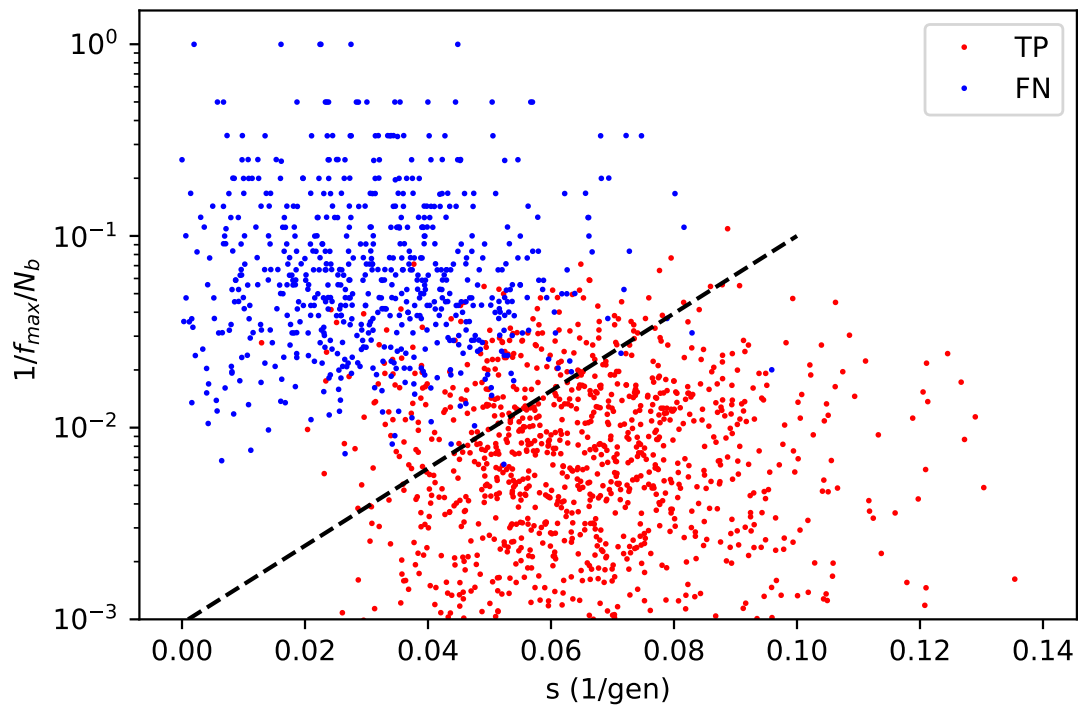


Figure 3.9. Selection coefficient $s(1/\text{gen})$ to the inverse of maximum size at the bottleneck $1/f_{\max}/N_b$ for true positive lineages (red circles) and false negative lineages (blue circles) with a one-to-one line (black dash).

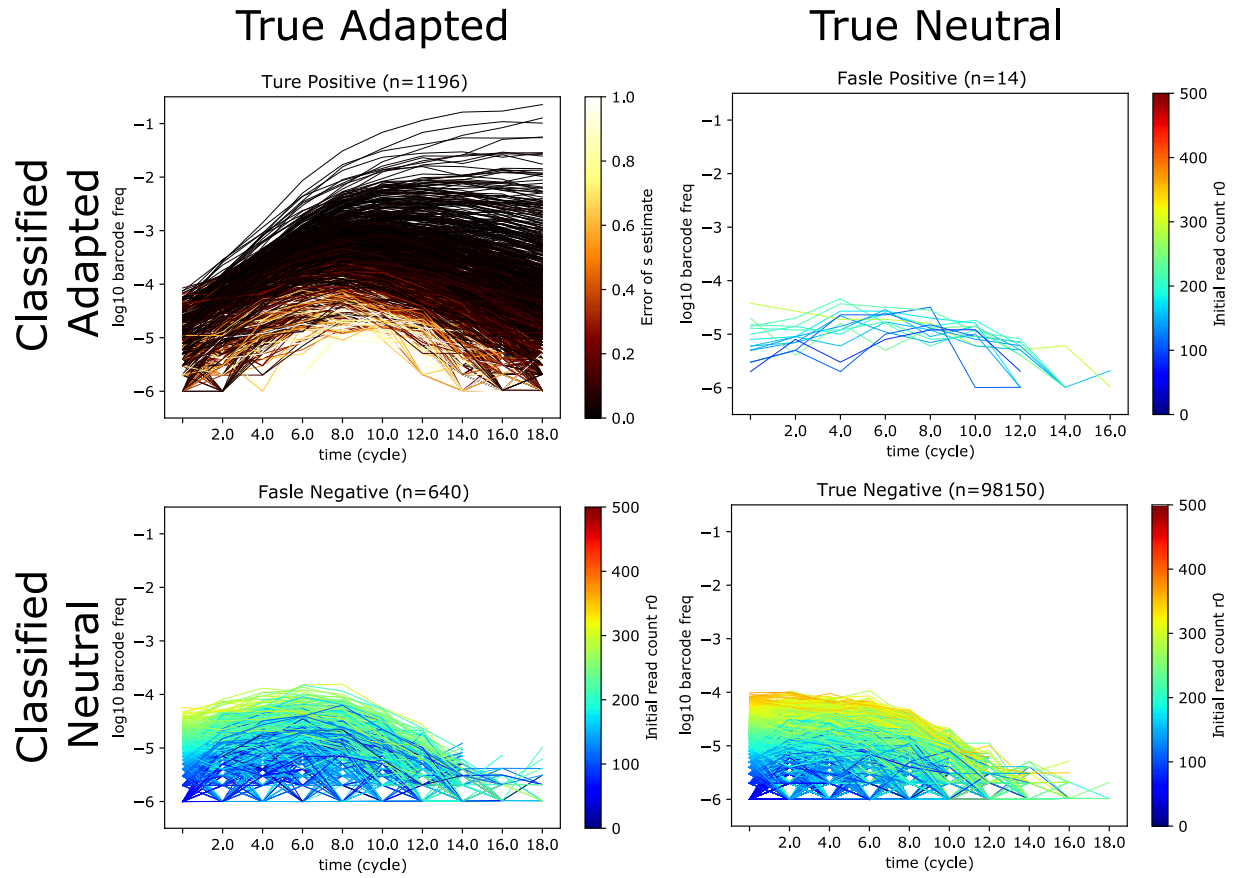


Figure 3.10. Lineage trajectories for (top left) 1196 true positive lineages, colored by the error of s estimates. Lineage trajectories for (bottom left) 640 false negative, (top right) 14 false positive and (bottom right) 98150 true negative lineages, colored by initial barcode read counts.

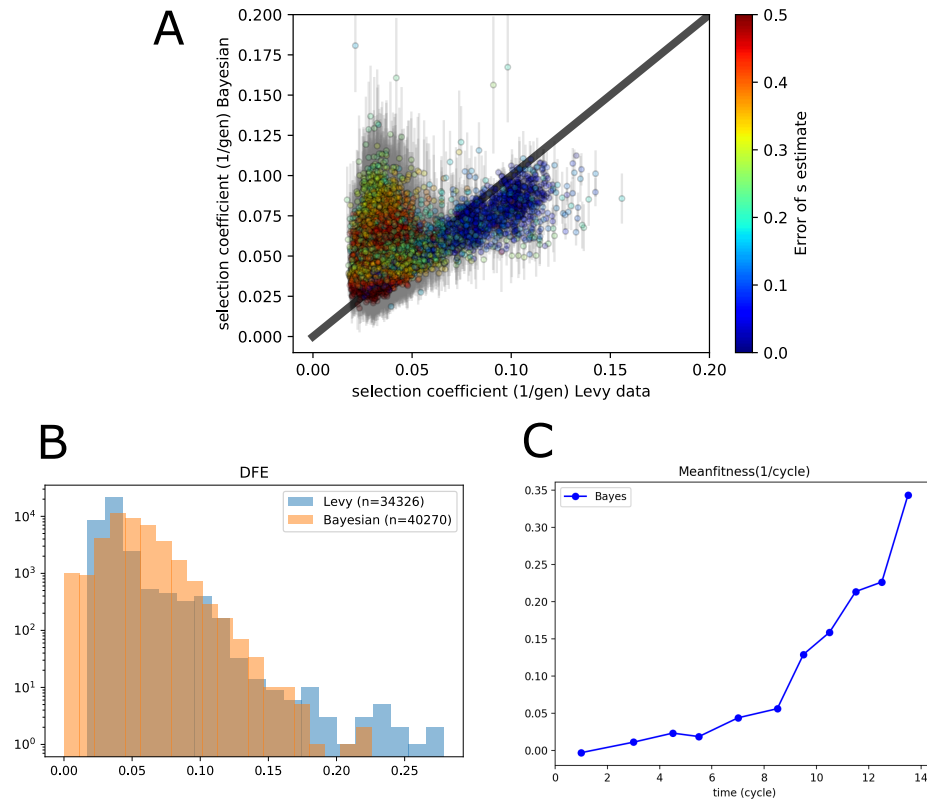


Figure 3.11. Application of the Bayesian filtering method to the yeast BLT data in yeast published in Ref. [58]. (A) Fitness effect of adapted lineages estimated by the Levy-Blundell method (x -axis) and by the Bayesian Filtering method (y -axis). Uncertainty in the estimates is indicated by color (B) DFE estimated by the Levy-Blundell method (blue) and by the Bayesian Filtering method (orange) (C) Mean-fitness trajectory, estimated by the Bayesian Filtering method.

3.3.3 Application to existing BLT data

Finally, I applied this approach to the BLT experimental data by Levy et al [58]. In this experiment, 5×10^5 barcodes were introduced into a yeast population and the population was propagated for 160 generations in batch culture, where on average $n = 140$ cells per barcoded lineages survived during each bottleneck, and each of these surviving individuals was represented by on average $c = 60$ reads in the sequencing data. The inferred population mean fitness has increased in this experiment to $\sim 5\%$ per generation at the 120 generations in the experiment (Figure 3.11)

My method identified 40,270 adapted lineages, compared to 34326 lineages reported by Levy et al. There are 23,386 lineages called as adapted by both methods. I find good correspondence between the two methods for those lineages whose selection coefficients estimated by both methods exceed 5% per generation ($R = 0.8$, Figure 3.11A). Incidentally, the uncertainty in my estimates for these lineages is also relatively small, with the coefficient of variation typically less than 20%. In contrast, the correspondence between the two methods is much worse for lineages that are reported by Levy et al to have small selection coefficients ($s < 5\%$). My estimates for these lineages are also less certain, with the coefficient of variation typically exceeding 20%, but it's observed not only more variation between the two estimates ($R = 0.024$, Figure 3.11A)), but also a systematic difference. Specifically, my method provides an estimate that is on average 0.17% per generation higher than the estimate provided by Levy et al. (Figure 3.11A). Since I do not have access to the ground truth in these data, I do not know what causes this discrepancy. However, my simulations discussed in the previous section suggest that the Bayesian method tends to overestimate the selection coefficients of lineages when the uncertainty is high (see Figure 3.7A) whereas the Levy-Blundell method generally tends to underestimate selection coefficients (see Figure 3.8). Thus, both biases lead to the discrepancies observed in Figure 3.11 for the Levy et al data.

3.4 Discussion

Overall, I have developed two Bayesian statistical models, namely the S-model and N-model, which incorporate hidden variables such as the selection coefficient and cell size at the individual lineage level. These models form the foundation of a Bayesian framework used to estimate hidden variables as posterior belief distributions, updating them over time based on a series of barcode read observations. Additionally, I conducted an additional experiment to determine the measurement noise utilized in the Bayesian models. Surprisingly, I discovered that the variance of the measured barcode read count follows a quadratic function of the mean, contrary to the linear function suggested in previous studies.

With this Bayesian framework in place, I proceeded to identify adaptive lineages using a threshold determined by optimal performance in simulation. The simulation was designed to represent a harsh evolving condition characterized by strong selection, significant genetic drift, and limited coverage, in which the previous Levy-Blundell method failed. The results demonstrate that the Bayesian method accurately identifies adapted lineages and provides precise estimations of their fitness effects.

Furthermore, I applied this novel approach to existing barcode lineage tracking (BLT) data collected under a gentle evolving condition and found a strong correspondence in identifying adapted lineages compared to the Levy-Blundell method.

One notable advantage of the Bayesian framework is its ability to overcome the challenges associated with inaccurate mean-fitness estimates. In this approach, mean-fitness is inferred directly from cell size rather than barcode read count. This enables the estimation of cell size for all individual barcoded lineages and the derivation of mean fitness from random lineages of varying sizes. As a result, the choice of "reference" lineages is not limited to small or neutral lineages. This novel approach shows great potential for applying BLT analysis to harsher evolving conditions characterized by strong selection coefficients (leading to the extinction of neutral lineages at later times), significant genetic drift (resulting in the inability of small lineages

to survive), and lower coverage (causing insufficient representation of small lineages).

In comparison to existing methods, my Bayesian filtering method exhibits two main strengths. First, it infers mean-fitness directly from cell size, eliminating the reliance on the assumption of neutral lineage decline. Consequently, my inference is not sensitive to the choice of reference lineages or their barcode read counts. Second, the Bayesian filtering method shows promise in identifying beneficial mutations and estimating their fitness effects in harsher environments. This expands the applicability of BLT experiments to study microbial evolution across a wider range of environments, beyond the gentle conditions typically explored. Additionally, since the distribution of fitness effects is unknown a priori in BLT experiments, the speed of adaptation is also unknown. The Bayesian analysis demonstrates potential in capturing beneficial mutations that evolve rapidly.

However, there are two main weaknesses associated with the Bayesian method. Firstly, it requires extensive computational resources due to the massive amount of computations involved, particularly in the computation of posterior distributions relying on extensive Markov Chain Monte Carlo (MCMC) sampling. To effectively analyze fitness, the Bayesian pipeline necessitates the use of multiple processors (16-32). Secondly, the Bayesian method tends to overestimate the fitness effect of weak beneficial lineages while not exhibiting the same bias for strong beneficial lineages. This discrepancy arises because the Bayesian method incorporates a "prior" guess of fitness effects for all lineages, and accurate fitness estimation relies on several observations from lineages. If a lineage lacks sufficient data, the fitness estimate becomes biased by the prior belief. As observed in the analysis, strong beneficial fitness is less affected by bias since the lineages persist at high frequencies even at later times. For many small beneficial mutations, data insufficiency prevents accurate estimations, either because the weak lineages go extinct at later times or because the mutations arise at later stages of the experiment.

3.5 Materials and Methods

3.5.1 Experimental Estimation of Measurement Noise

Strains

All strains used here were obtained from Dr. Sandeep Venkataram. Each of the strains carries a unique barcode determined in the previous study[102]. The strain is diploid yeast *Saccharomyces cerevisiae* derived from strain GSY6699, which carry the integrated DNA barcode (30 bp) at locus YBR209W [58].

Culture preparation

All yeast clones were stored at -70°C . I transfer 1 to 3 μl of the 26 clones into a 15 ml test tube with 3 ml of YPD (10 g yeast extract, 20 g peptone, 20 g dextrose, 1L Milli-Q water). After an overnight growth (shaken at 200 rpm at 30 degree) the yeast cultures reach saturation density at about 1.5×10^8 cells/ml. The saturated density of each clone is measured by Coulter Counter.

The clone mixture was designed to comprise 1 clone at 44.4% population, 5 clones at 10% each, 5 clones at 1% each, 5 clones at 0.1% each, 5 clones at 0.01% each, and 5 clones at 0.001% each. For clones with targeted frequencies of 1%, 0.1%, 0.01% and 0.001%, I dilute the saturated cultures 1:10, 1:100, 1:100 and 1:10000 in PBS respectively. To make the clone mixture, 200 μl of each diluted culture was pooled, followed by adding 5 of 200 μl culture (targeted frequency 10%) and 1 of 600 μl culture (targeted frequency 44%). The precise frequency of each clone was determined from the dilution ratio, pooled volume and its saturated density. The initial cell frequency of each clone are shown in Table 3.4.

To prepare the cell stocks, I washed the clone mixture and resuspended it in 20 ml of PBS. The mixture density was 1.57×10^7 cell/ml after resuspension. I aliquoted the clone mixture for 20 portions of 1 ml into 2 ml cryogenic tubes. Each 1 ml culture was vortex-mixed with 500 μl of 80% glycerol (400 μl 100% glycerol in 100 μl 1XPBS), and stored at -70°C . I picked 10

Table 3.4. Initial clone frequencies in the mixture culture

Clone	Barcode Sequence	Saturated Density (10 ⁸ cells/ml)	Dilution Ratio	Volume (μl) added to mixture	Clone Frequency
1	CGGCCGGATAAACTGTTTAAATGTTTGTTT	1.527	1	600	39.4584%
2	TTTCTGGCAAACCGATTACTTGTATGTA	1.196	1	200	10.2961%
3	AGTAATAATAATGAAATTCGTAATTCAATT	1.223	1	200	10.5312%
4	GTCCACCAAACCTCATTTACCTATTAAATT	1.262	1	200	10.8671%
5	TTTTGGTTAAAAATAGTTTGCCTTTTAAATT	1.282	1	200	11.0419%
6	ATTACAACCAAACCGTTTATCGGTTGCATG	1.296	1	200	11.1599%
7	TTTTTTCTTAACTGAGTTCCGCTTTTTAGT	1.351	10	200	1.1634%
8	ATTAGCGACAACATCATTGCTTGTTCACGG	1.363	10	200	1.1735%
9	TTTTCAAAAAATAGTTTGCCGTTTGCCGA	1.388	10	200	1.1952%
10	ATTAGTAATAATAGCCTTAGACCTTCATTT	1.399	10	200	1.2052%
11	AATGTTACAAAACCTTATTTCTTTTAAACA	1.411	10	200	1.2153%
12	GTCCACCAAACCTCTGTTTACCGTTGGAGG	1.426	100	200	0.1228%
13	-	1.438	100	200	0.1238%
14	AATGACGGTAACCTCATTCCTTTTGGACC	1.438	100	200	0.1238%
15	TTTTTAAGTAACGCACTTGCCCTTTAGCCG	1.455	100	200	0.1253%
16	TTTTCGGTCAAGCTCATTACGGATTTGAG	1.461	100	200	0.1258%
17	TTTTGTATAAGGTGCTTAAGCGTTATCTT	1.463	1000	200	0.0126%
18	TTTTTTTACAAGACCTTTCTTACTTTAATT	1.470	1000	200	0.0127%
19	TTTTACCTAAATAAACTTAAGTGTTTCATGG	1.481	1000	200	0.0128%
20	GTCCAAGTGAATCAGATTATACGTTTCCTG	1.506	1000	200	0.0130%
21	ATTACACAAAACGTTCTTCCTTGTGTAAA	1.520	1000	200	0.0131%
22	ATTAGTCATAATCCCTTAGGGTTTAGCCG	1.553	10000	200	0.0013%
23	TTTTATTATAATCGAATTAAGTTTAAAGTT	1.600	10000	200	0.0014%
24	TTTTCTTCTAAAATGTTTTATGTTAAGGA	1.605	10000	200	0.0014%
25	-	1.685	10000	200	0.0015%
26	GTCCTTATAAACACATTGAGTATTATCGA	1.745	10000	200	0.0015%

cell stocks for further barcode measurement experiments.

Barcode measurement process

I follow the protocols published in our previous work (Ref. [102]) for DNA isolation and library preparation. This procedure is repeated for 10 cell stocks which are prepared from the same population.

DNA isolation

The cell stock was thawed, spun down and the supernatant was removed. The pellet was resuspended in 300μl 3% SET buffer (3% SDS, 10mM EDTA, 30mM Tris) and incubated at 65°C for 15 minutes. The tube was cooled to room temperature by immersing in room temperature water, then 2.5 μg of RNase A was added. After vortexing, the mixture was

incubated at 37°C for 1 hour, after which it was cooled on ice. 150 µl of 3M Sodium Acetate was then added and mixed by inversion, after which it was cooled on ice for further 5 minutes before centrifuging at maximum speed for 10 minutes on a tabletop centrifuge. The supernatant was transferred to a new tube and DNA was precipitated by the addition of 500 µl of isopropanol, which was then mixed by inversion and pelleted by centrifugation for 1 minute at maximum speed. The supernatant was removed and the pellet was washed with 200 µl cold 70% ethanol without vortexing before being allowed to dry inverted for 30 minutes at room temperature. After that DNA was resuspended in 50 µl molecular biology grade water. The DNA template was stored at −20°C.

Library preparation

The barcode locus was amplified through a 2-step PCR protocol. The first amplification added inline indices for sample multiplexing, universal molecular identifiers (UMI) for removing PCR duplicates during analysis and Illumina-compatible adapter sequences for a second round of amplification with standard Illumina Nextera XT primers. For the first reaction, 5 µl of DNA template is mixed with 10µl of OneTaq 2x Master Mix, 1 µl of 25mM MgCl₂, 1 µl each of the forward and reverse primers (at 10mM concentration) and 2µl of molecular biology grade water. This 20µl mixture was amplified using the following conditions:

1. 94°C for 10 min
2. 94°C for 3 min
3. 55°C for 1 min
4. 68°C for 1 min
5. Repeat steps 2–4 for a total of 8 cycles
6. 68°C for 1 min
7. Hold at 4°C

The amplified product was purified using Ampure XP magnetic beads following the protocol in Ref. [6] (with 20 μ l of beads used per sample). Then, 5 μ l of the purified product was used as the template for a second reaction along with Illumina Nextera XT primers (1 μ l of 10mM stock for each primer), 1 μ l of 25mM MgCl₂, 10 μ l of OneTaq 2x Master Mix and 2 μ l of water with the following reaction conditions:

1. 94°C for 5 min
2. 94°C for 30 sec
3. 62°C for 30 sec
4. 68°C for 30 sec
5. Repeat steps 2–4 for a total of 25 cycles
6. 68°C for 5 min
7. Hold at 4°C

The PCR products were purified using Ampure XP beads as before. 10 μ l of each purified DNA was mixed to form a pool for Illumina sequencing, concentrated to 33 μ l DNA using Ampure beads as before with equal volume of beads as pooled sample. I then size selected via agarose gel extraction (QIAquick Gel Extraction Kit) to isolate the correct amplicon before submitting for sequencing.

Sequencing

My 10 DNA samples were sequenced on HiSeq with DNA samples of other experiments. An average of 2×10^5 total reads were obtained for my 9 replicated experiments (Replicate # 10 was excluded due to low coverage, see Table 3.5).

Table 3.5. Total barcode reads for 10 replicated experiments

Replicate#	1	2	3	4	5
Total reads from HiSeq	458,208	206,496	156,833	180,889	526,253
Total reads after barcode identification	331,331	137,156	108,375	133,037	373,659
Replicate#	6	7	8	9	10
Total reads from HiSeq	458,119	229,944	41,663	271,717	16,600
Total reads after barcode identification	313,684	158,452	16,613	189,674	21

Data analysis

The goal of this analysis is to understand the mean and variance in the barcode frequency data. To identify and count DNA barcode sequences, I processed the raw sequencing data file with a customized python program, BarcodeCounter, described in Ref. [102] and available at

<https://github.com/sandeepvenkataram/BarcodeCounter2>

I obtain the total reads after running the program, as shown in Table 3.5. Replicate-#10 was excluded due to low coverage (Table 3.5). Among the 26 clones used in the measurement experiment, The records of the barcode sequence of 2 clones from the previous yeast lineage tracking experiment is lost. Another 2 clones with the lowest frequencies were not represented in any of the replicates were also removed. In the follwoing data analysis, I use data of 22 clones presenting in 9 replicated experiments.

Mean of barcode frequency

For each barcode i in replicate n , I calculate its barcode frequency $f_{i,n}$ by dividing its barcode read count $r_{i,n}$ by the total number of reads in that replicate R_n .

$$f_{i,n} = \frac{r_{i,n}}{R_n} \quad (3.12)$$

The mean of observed barcode frequency, $\bar{f}_i$, for barcode i , is calculated from the average barcode frequency among 9 replicates.

$$\bar{f}_i = \frac{1}{9} \sum_{n=1}^9 f_{i,n} \quad (3.13)$$

The expectation of $\bar{f}_i$ is the true cell frequency of barcode i , p_i .

$$\mathbb{E}[f_{i,n}] = p_i \quad (3.14)$$

Variance of barcode read count

In barcode lineage tracking (BLT) experiment, the total read depth is a known quantity. I am interested in the variance of data resulting from the stochastic process of fixed total read depth. I therefore focus on each replicate since the read depth is fixed in the same measurement process. To calculate the variance, I first group the clones $\sigma_{n,p}^2$ of similar frequency p as set X_p . 5,5,4,5,2 clones present in groups of $p = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ and 10^{-5} . The variance of barcode read count is calculated from the clones in X_p .

$$\sigma_{n,p}^2 = \text{variance} \{r_{i,n}\} \forall i \text{ in } X_p. \quad (3.15)$$

To understand the variance to mean relationship, I derive the mean of read count for such group in replicate n .

$$\bar{r}_{n,p} = \text{mean} \{r_{i,n}\} \forall i \text{ in } X_p \quad (3.16)$$

3.5.2 Bayesian Filtering Algorithm

The algorithm 1 shows the overall procedure of the Bayesian filtering method. I implemented this method in a Python based software package, which is available at

https://github.com/HuanyuKuo/Bayesian_Filtering_Method

This package uses *stan* (c language) to build up statistical models (N-model, S-model) and utilizes *pystan* as an interface between Python and *stan*.

To compute the belief distribution, I conduct MCMC algorithm to generate samples from the belief distributions in both models. I choose No-U-Turn Sampler (NUTS) as my MCMC sampler. For each belief distribution, it sampled 4500 steps with 1000 burn-in steps. Sampling is repeated twice to check for convergence. The belief distribution is then an estimated density

from the MCMC sampling results.

To increase the computation efficiency, I implement multiprocessing for the MCMC sampling. The number of processors can be set manually depending on user's environment. In this study, 16-32 processors were used for running $10^4 - 10^5$ lineages.

Since the likelihood function is constructed from randomly selected lineages, using the sampling-based probability of each lineage, this function is noisy and is not analytically defined. To find an optimal position of a noisy function, I implement a python package- *Noisyopt*, which finds local optimization from the average over the stochasticity in a noisy function. My computation of the likelihood function also utilizes multi-processor to speed up the MCMC sampling.

This Bayesian filtering program outputs the inference of mean fitness trajectory and the estimated fitness effects (with uncertainties) for lineages identified as adapted.

3.5.3 Simulation Method (Bayesian)

I simulate the BLT experiment in a “harsh” scenario where selection is strong (such that mean population fitness reaches 12% by generation 140, genetic drift is strong (bottleneck size = 10 per barcoded lineage) and sequencing coverage is small (= 10x per cell at the bottleneck).

The simulation starts with $N_L = 10^5$ barcoded lineages. The initial population size of each lineage is randomly drawn from an exponential distribution with mean of N/N_L , where the total population size is $N = 2 \times 10^8 = 2000N_L$. 2,000 lineages carry a pre-existing beneficial mutation. The selection coefficient of each mutant is randomly drawn from a normal distribution with mean 5% and width (2 standard deviation) of 5%. For simplicity, I do not allow any new mutations to arise during the simulation. At each cycle, I dilute the total population by the diluting factor $D = 100$ to bottleneck size $N_b = 10N_L$. After dilution, lineages increase or decrease deterministically based on their s and the current mean-fitness $\bar{s}_t$ until the next dilution (6.6 generations of growth). The dilution-growth cycles are repeated for a total 18 times (~ 140 generations). To simulate barcode sequencing, individuals are randomly sampled from the

Algorithm 1. Bayesian Filtering Method

```
1: for cycle  $k = 0$  do
2:   for all lineage  $i$  do
3:     procedure INITIALIZATION OF BELIEF DISTRIBUTION ▷ subsection 3.2.3
4:        $P_{i0}(n) = \gamma\left(n; \text{mean} = r_0 \frac{N_0}{R_0}, \text{variance} = r_0(1 + 0.01 r_0) \frac{N_0^2}{R_0^2}\right)$ 
5:        $P_{i0}(s) = U(s; s_{min}, s_{max})$ 
6:     end procedure
7:     procedure PUTATIVE LINEAGE CLASS
8:       lineage belongs to NEU class
9:     end procedure
10:   end for
11: end for
12: for cycle  $k = 1, \dots, k_{end}$  do
13:   procedure ESTIMATING MEAN FITNESS EFFECT
14:      $\bar{s}_k = \arg \max L(\bar{s}_k)$  ▷ eqn 3.11
15:   end procedure
16:   for all lineage  $i$  do
17:     procedure PROJECTING TO PRIOR
18:       S-model  $P_k^{pr}(n_k, s | \mathbf{r}_{k-1})$  ▷ eqn 3.1, table 3.3
19:       N-model  $P_k^{pr}(n_k | \mathbf{r}_{k-1})$  ▷ eqn 3.4
20:     end procedure
21:     procedure UPDATING AS BELIEF(with data  $r_k$ )
22:       S-model  $P_k^b(n_k, s | \mathbf{r}_k)$  ▷ eqn 3.2
23:       N-model  $P_k^b(n_k | \mathbf{r}_k)$  ▷ eqn 3.5
24:     end procedure
25:     procedure PUTATIVE LINEAGE CLASS
26:       Bayes Factor  $B_k$  ▷ eqn 3.10
27:       classify lineage to {UNK, NEU, ADP} ▷ table 3.3
28:     end procedure
29:   end for
30: end for
31: Final calling of adapted lineages ▷ subsection 3.2.4
32: Estimating fitness effects
```

population every other cycle, and their barcodes are sequenced with a total read depth $R = 10^6$. I model the sampling (barcode measurement) using Negative binomial distribution (equation (3.9)) with $\varepsilon = 0.01$ (see Section 3.2.2).

Acknowledgments

Chapter 3, in full is currently being prepared for submission for publication of the material. Huan-Yu, Kuo; Sergey, Kryazhimskiy. The dissertation author was also the primary researcher and author of this material.

Chapter 4

A Method for Estimating the Arrival Rate for a Class of Adaptive Mutations

4.1 Introduction

Mutations are the source of all genetic variation including the variation that drives adaptative evolution. Thus, understanding the rate at which mutations arise is of fundamental important in evolutionary biology [3, 49, 65, 92]. While measuring mutation rates is not easy—primarily because mutations are rare events—several methods have been developed for estimating the rates of mutations, and robust estimates for the overall genome-wide rates are available for many organisms [3, 18, 52, 65, 86, 92].

The overall genome-wide mutation rate enters as a key quantity into many population-genetic models, but the assumption that there is a single mutation rate is false. The fact that different types of mutations happen at dramatically different rates has been well established, e.g., large chromosomal duplications and deletions are much more common than individual point mutations [40, 109]. In addition to the variation across types of mutations, recent studies have shown that the rates of mutations also vary across the genome [26, 40, 56, 66, 71, 72, 93, 97]. Importantly, mutation of different types and at different loci have different phenotypic consequences for the organism, e.g., aneuploidies are often associated with cancer [16, 35, 37]. Moreover, there is emerging evidence that the variation in the rates at which different mutations arrive is an important factor that determines how a population adapts [31, 80, 95, 103]. Despite

the importance of such variation, our understanding of how the rates of adaptive mutations vary across the genome is extremely poor, even for well-studied laboratory organisms, such as the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae*.

The challenge with measuring the rates of adaptive mutations is twofold. First, beneficial mutations are rare, comprising probably at most a few percent of all mutations [22, 58], which implies that finding adaptive mutations among random mutations would require impractically large sample sizes. It is, of course, easy to enrich a population for beneficial mutants by simply allowing it to evolve and thereby letting natural selection amplify the frequencies of beneficial mutants. However, the second challenge is that the frequencies of beneficial mutations are amplified by selection unevenly. The frequencies of mutations that provide strong fitness benefits increase quickly and robustly by natural selection, making them relatively easy to identify and isolate [13, 48, 53]. In contrast, weakly beneficial mutations are subject to stronger genetic drift (i.e., their frequencies fluctuate more widely) and they are often outcompeted by stronger adaptive mutants before reaching detectable frequencies [57], which makes them much more difficult to detect and isolate. The BLT approach allows us to detect beneficial mutations at extremely low frequencies, soon after they establish and before they are outcompeted [58]. Some of these mutations can subsequently be isolated using the approach developed by Venkataram et al [101]. In brief, we isolate the evolved population at the cycle when the most beneficial lineages are present, and sanger-sequenced clones to identify unique barcode clones, following by whole-genome sequence to identify the mutations.

This strategy for sampling adaptive mutations alleviates the strongest biases introduced by natural selection while still leveraging it to amplify the frequencies of adaptive mutations. Nevertheless, estimating the rate of beneficial mutations that occur at different genomic loci remains challenging. The main obstacle is that each new beneficial mutation randomly arises in a single individual in the population, which implies that, before the mutant can be detected, it must sufficiently expand in size while surviving genetic drift and overcoming the competition against the rest of the population whose fitness is continuously increasing. Therefore, an

inference of the mutation rate from BLT data requires a sufficiently accurate and tractable model of these complex stochastic dynamics. These dynamics are typically modeled with the Wright-Fisher model [23, 99, 107]. However, this model has limited analytical tractability, and various approximations are necessary to obtain quantitative results. One approach that has proven successful is to use the branching process theory for modeling frequency dynamics of a lineage while it is rare [17, 34]. Here, I extend this approach to model multiple lineages that carry independently arisen mutations from the same class that compete against the rest of the population whose mean fitness continuously increases (Section 4.2). I apply this theory to construct a likelihood function for the mutation rate of a class of beneficial mutations (e.g., mutations in the same gene), which is based on the number of times mutations from this class were sampled and observed. (Section 4.3). Finally, I apply this approach to data from a recent BLT experiment in yeast (Section 4.4.2).

4.2 Model of stochastic dynamics of a nascent mutant in an adapting population

4.2.1 Theory

Consider a single mutant with selective advantage s arising in a population of size N . During the small time interval from time t to $t + \delta t$, each mutant cell may die with probability δt , or give birth with probability $\delta t (1 + s - \bar{s}(t))$, where $\bar{s}(t)$ is the average selection coefficient in the population (“mean fitness”), which is assumed to be known. The dynamics of mutant cell number n can be derived from a general-birth-death process.

To solve this birth-death problem, I start by considering all possible cell numbers as

states $0, 1, 2, \dots, \infty$ along a Markov chain. The transition probability from state i to state $i + j$ is

$$T_{i+j,i}(\delta t) = \begin{cases} i \delta t & , \text{ if } j = -1 \\ i \delta t (1 + s - \bar{s}(t)) & , \text{ if } j = +1 \\ 1 - i \delta t (2 + s - \bar{s}(t)) & , \text{ if } j = 0 \\ 0 & , \text{ otherwise.} \end{cases} \quad (4.1)$$

I denote by $p_i(t)$ the probability that the mutant's population size at time t is i . The differential equations of each state probability can be expressed in terms of nearby state probabilities.

$$\frac{dp_i(t)}{dt} = \begin{cases} p_1(t) & , \text{ if } i = 0 \\ (i+1)p_{i+1}(t) + (i-1)(1+s-\bar{s}(t))p_{i-1}(t) - i(2+s-\bar{s}(t))p_i(t) & , \text{ otherwise} \end{cases} \quad (4.2)$$

I define the generating function as

$$\mathbb{G}(z, t) = \sum_{i=0}^{\infty} p_i(t) z^i, \quad (4.3)$$

which depends on two continuous variables, z and t , where $t \geq 0$, and the domain of z consists of the values where the sum converges. The generating the function obeys the following partial differential equation, which is derived from equation (4.2),

$$\frac{\partial \mathbb{G}(z, t)}{\partial t} = [(1 + s - \bar{s}(t)) (z^2 - z) + (1 - z)] \frac{\partial \mathbb{G}(z, t)}{\partial z} \quad (4.4)$$

Since the mutant starts from a single cell, it is reasonable to choose an initial condition $\mathbb{G}(z, 0) = z$.

The solution to equation (4.4) is

$$\mathbb{G}(z, t) = 1 + \frac{1}{\frac{e^{-\rho(t)}}{z-1} - \int_0^t d\tau (1 + \omega(\tau)) e^{-\rho(\tau)}} \quad (4.5)$$

with the effective growth rate

$$\omega(t) = s - \bar{s}(t) \quad (4.6)$$

and the effective growth factor

$$\rho(t) = \int_0^t d\tau \omega(\tau) \quad (4.7)$$

I next directly calculate the state probabilities $p_n(t)$ and the moments of cell number from the generating function.

$$p_n(t) = \begin{cases} \mathbb{G}(0, t) & , \text{ if } n = 0 \\ \frac{1}{n!} \left| \frac{\partial^n \mathbb{G}(z, t)}{\partial z^n} \right|_{z=0} & , \text{ if } n \geq 1 \end{cases} \quad (4.8)$$

The first moment of cell number (mean) is

$$\mathbb{E}(n) = \left| \frac{\partial \mathbb{G}(z, t)}{\partial z} \right|_{z=1} = \sum_{i=0}^{\infty} i p_i(t), \quad (4.9)$$

and the variance is

$$\sigma^2(n) = \left| \frac{\partial^2 \mathbb{G}(z, t)}{\partial z^2} \right|_{z=1} + \left| \frac{\partial \mathbb{G}(z, t)}{\partial z} \right|_{z=1} - \mathbb{E}(n)^2. \quad (4.10)$$

To simplify notations, I denote

$$a(t) = e^{-\rho(t)}, \quad (4.11)$$

and

$$b(t) = \int_0^t dt' e^{-\rho(t')} (1 + \omega(t')). \quad (4.12)$$

Taking equation (4.5) into (4.8), I derive the expressions of state probabilities in terms of a and

b.

$$p_n(t) = \begin{cases} 1 - \frac{1}{a+b} & , \text{ if } n = 0 \\ \frac{a}{b^2} \left[\frac{b}{a+b} \right]^{1+n} & , \text{ if } n \geq 1 \end{cases} \quad (4.13)$$

Similarly, take 4.5 into moment equation and derive

$$\mathbb{E}(n) = \frac{1}{a} = e^{\rho(t)}, \quad (4.14)$$

$$\sigma^2(n) = \frac{b-a-1}{a^2}. \quad (4.15)$$

The probability of survival is

$$P_{\text{surv}}(t) = 1 - p_0(t) = \frac{1}{a+b}. \quad (4.16)$$

The expected cell number of survival mutant is

$$\mathbb{E}(n | \text{Survive}) = \frac{a+b}{a} = 1 + \frac{b}{a}, \quad (4.17)$$

with corresponding state probabilities for $n \geq 1$

$$p_n(t | \text{Survive}) = \frac{a}{b} \left[\frac{b}{a+b} \right]^n = \frac{a}{a+b} \left[\frac{b}{a+b} \right]^{n-1}, \quad (4.18)$$

which is a geometric distribution.

Note that $1/a$ is the trajectory of the expected cell count (see equation (4.14)). If the mutant survives and grows, its population is usually much greater than one. Based on equation (4.17), we can assume that b is much greater than a for the most mutants. Thus $1/b$ is approximately the survival probability (see equation (4.16)).

Similarly to the procedure described in Refs. [17, 58], a transition from state n to

establishment time τ is defined as

$$n = \mathbb{E}(n | \text{Survive}) e^{-\rho(\tau)} = \frac{1}{P_{\text{surv}}(t)} e^{\rho(t) - \rho(\tau)}, \quad (4.19)$$

so that

$$\left| \frac{dn}{d\tau} \right| = n \frac{d\rho(\tau)}{d\tau} = n \omega(\tau). \quad (4.20)$$

The transition p.d.f. of τ from $p_n(t | \text{Survive})$ is said to be the probability distribution of establishment time for a survival mutant with s .

$$\begin{aligned} P(\tau | \text{Survive}) &= p_n(t | \text{Survive}) \left| \frac{dn}{d\tau} \right| \\ &= \omega(\tau) e^{-\rho(\tau)} \frac{a+b}{b} \text{Exp} \left[-e^{-\rho(\tau)} \frac{a+b}{a} \ln \left(1 + \frac{a}{b} \right) \right]. \end{aligned} \quad (4.21)$$

The mutant dynamics can be pictured as the follows. A random beneficial mutant, starting from single cell, might go extinct or survive with survival probability (equation (4.16)). For a mutant that survives, its population drifts stochastically before the establishment time τ , and grow deterministically in size (equation (4.19)) afterwards. The establishment time τ is a random variable distributed according to equation (4.21)

4.2.2 Approximation for the analytical solution

To derive the analytical expression of this model, it is necessary to work out the solution of $a(t)$ (equation 4.11) and $b(t)$ (equation 4.12). Since the mean-fitness trajectory $\bar{s}(t)$, which is measured explicitly, can be describes analytical function such as linear, exponential, or logistic function, $a(t)$ is easy to solve or compute numerically. Here I will focus on the solution of $b(t)$.

The function $b(t)$ is worked out as

$$\begin{aligned}
b(t) &= \int_0^t d\tau (1 + \omega(\tau)) e^{-\rho(\tau)} \\
&= \int_0^t d\tau e^{-\rho(\tau)} + \int_0^t d\tau \omega(\tau) e^{-\rho(\tau)} \\
&= \int_0^t d\tau e^{-\rho(\tau)} + \int_{\rho(0)}^{\rho(t)} d\rho e^{-\rho} \\
&= b_1(t) + (1 - a(t)),
\end{aligned} \tag{4.22}$$

where the left side is defined as

$$b_1(t) = \int_0^t d\tau e^{-\rho(\tau)}. \tag{4.23}$$

The solution of function $b_1(t)$ is approximated in two different regions, strong and weak fitness effects, as described below.

Strong fitness effect

For mutant with strong fitness effects ($\rho \gg 1$), its population expand quickly and does not decay (growth factor $\rho > 1$). In this region, the effect of $\omega(t)$ on small $e^{-\rho}$ can be ignored such that $\omega(\rho) \sim \omega(t=0)$ is a constant in the integrate of $e^{-\rho}$. The function $b_1(t)$ is approximated as

$$\tilde{b}_1(t) = \int_{\rho(0)}^{\rho(t)} \frac{d\rho}{\omega} e^{-\rho} = \frac{1}{\omega(0)} (1 - a(t)). \tag{4.24}$$

Substituting this approximation into equations (4.16)–(4.18) then yields

$$P_{\text{surv}}(t) = \frac{1}{1 + \frac{1}{\omega(0)} (1 - e^{-\rho(t)})} = \frac{\omega(0)}{1 + \omega(0) - e^{-\rho(t)}}, \tag{4.25}$$

$$\mathbb{E}(n | \text{Survive}) = e^{\rho(t)} + \frac{e^{\rho(t)} + 1}{\omega(0)}, \tag{4.26}$$

$$p_n(t | \text{Survive}) = \frac{\omega(0)}{1 + \omega(0)} e^{-\rho(t)} \left[\frac{1 + \omega(0)}{1 + \frac{\omega(0)}{1 - e^{-\rho(t)}}} \right]^n. \tag{4.27}$$

In long time limit, $a(t) = e^{-\rho}$ could be ignored as ρ increases, and $b(t)$ would reach a stationary value $1 + \frac{1}{\omega(0)}$. Therefore, in the limit of $e^{\rho(t)} \gg 1$, equations (4.16)–(4.18) can be further simplified.

$$P_{\text{surv}}(t) \sim \frac{\omega(0)}{1 + \omega(0)}, \quad (4.28)$$

$$\mathbb{E}(n | \text{Survive}) \sim \frac{1 + \omega(0)}{\omega(0)} e^{\rho(t)}, \quad (4.29)$$

$$p_n(t | \text{Survive}) \sim \frac{\omega(0)}{1 + \omega(0)} e^{-\rho(t)} \text{Exp} \left[-n \frac{\omega(0)}{1 + \omega(0)} e^{-\rho(t)} \right]. \quad (4.30)$$

$$P(\tau | \text{Survive}) \sim \omega(\tau) e^{-\rho(\tau)} \text{Exp} \left[-e^{-\rho(\tau)} \right]. \quad (4.31)$$

Note that my solutions in this region have similar as those obtained by Desai and Fisher (Ref. [17]) in the case of zero mean fitness $\bar{s} = 0$ (see Table 4.1 for comparison). Furthermore, in the limit of zero mean fitness ($\bar{s}(t) \rightarrow 0$) and $e^{\rho} \gg 1$, these functions and probabilities are equivalent.

Weak fitness effect

For mutant of a weak fitness effect ($\rho \ll 1$), its population dynamic includes three phases: (1) grow like “strong” mutant when the mean-fitness is still small ($s > \bar{s}(t)$), (2) slow down the expansion when the mean-fitness catches up to its fitness effect ($s \sim \bar{s}(t)$) and (3) decrease in size when this mutant is outcompeted by other mutants ($s < \bar{s}(t)$). Therefore, the behaviour of $b_1(t)$, corresponding to these phase, are approximated as

$$\tilde{b}_1(t) = \begin{cases} \frac{1}{\omega(t_0)} (1 - a(t)), & t < t_{c1}, \\ \frac{1}{\omega(t_0)} (1 - a(t_{c1})) + e^{-\rho^*} (t - t_{c1}), & t_{c1} \leq t < t_{c2}, \\ \frac{1}{\omega(t_0)} (1 - a(t_{c1})) + e^{-\rho^*} (t_{c2} - t_{c1}) + \frac{1}{-\omega(t_0)} (a(t_{c2}) - a(t)) & t > t_{c2}, \end{cases} \quad (4.32)$$

where t_c is chosen so that $\bar{s}(t_c) = s$ and $t_{c1} = t_c/2$, $t_{c2} = 3t_c/2$ and $\rho^* = (\rho(t_c) + \rho(t_{c1}))/2$.

Table 4.1. Solution of branching process with zero mean fitness (second column, all formula obtained in Ref. [17]) and with non-zero mean-fitness.

	Branching process ($\bar{s} = 0$)	Branching process ($\bar{s} \neq 0$)
Probability of extinction	$P[n = 0; t] = \frac{e^{st} - 1}{(1+s)e^{st} - 1}$	$P[n = 0; t] = \frac{e^{\rho(t)} - 1}{(1+\omega(0))e^{\rho(t)} - 1}$ $\omega = s - \bar{s}(t)$ $\rho(t) = \int_0^t \omega(t') dt'$
State probability	$p_n(t) = \frac{s^2 e^{st}}{[(1+s)e^{st} - 1][(1+s)e^{st} - 1 - s]} \times \left[\frac{(1+s)e^{st} - 1 - s}{(1+s)e^{st} - 1} \right]^n$	$p_n(t) = \frac{a}{b(a+b)} \left(\frac{b}{a+b} \right)^n$ $a = e^{-\rho(t)}$ $b \approx (1 + \frac{1}{\omega(0)})(1 - a)$
State probability for survival mutant	$p_n(t \text{Survive}) \approx \frac{s}{(1+s)e^{st} - 1 - s} \exp \left[-\frac{sn}{(1+s)e^{st} - 1} \right]$	$p_n(t \text{Survive}) \approx \frac{\omega(0)}{(1+\omega(0))e^{\rho(t)}} \exp \left[-\frac{\omega(0)n}{(1+\omega(0))e^{\rho(t)}} \right]$
Expected cell size	$\langle n \rangle = e^{st}$	$\langle n \rangle = e^{\rho(t)}$
Expected cell size for survival mutant	$\langle n \text{Survive} \rangle = e^{st} + \frac{e^{st} - 1}{s}$	$\langle n \text{Survive} \rangle = \frac{1+\omega(0)}{\omega(0)} e^{\rho(t)}$
Transition formula	$n \equiv \frac{1}{s} e^{s(t-\tau)}$	$n \equiv \frac{1+\omega(0)}{\omega(0)} e^{\rho(t)-\rho(\tau)}$
Probability of establishment time for survival mutant in long term limit	$P[\tau \text{Survive}] = \frac{s}{1+s} \exp \left[-s\tau_{est} - \frac{e^{s\tau_{est}}}{1+s} \right]$	$P[\tau \text{Survive}] = \omega(\tau) \exp \left[-\rho(\tau) - e^{-\rho(\tau)} \right]$

4.3 Likelihood function of mutation rate

The goal in this chapter is to estimate the rate of the same class mutations, who arise, survive and grow in the evolving population, then be isolated and sequenced in the BLT experiment[102]. To infer such mutation rate, it is necessary to model the probability of “clone” observations- resting from the the rate of mutation, the growth of mutant, and the sampling procedure. Here, I construct the likelihood function of mutation rate as the probability of clone sampling.

To keep the model as simple as possible, I assume that our population has the constant size N , mutations from the mutation class m arise at (an unknown) rate U_m per division. All mutations from the same class are assumed to confer the same (known) fitness effect s_m relative to the ancestor. After T generations of evolution, a total K isolates are sampled from the whole population, they are all genotyped and k_m of them are determined to have the adaptive mutation from the focal class.

To estimate the mutation rate U_m for mutation class m , I define the log-likelihood function

$$L(U_m) = \log P(k_m; s_m, U_m). \quad (4.33)$$

The sampling probability $P(k; s, U)$ for k clones depends on several factors: the number of lineages that independently arose in the population and that carry a mutation of the focal type, as well as on the frequencies of these lineages at the sampling time point T . It is difficult to obtain an analytical expression for this probability because it requires integrating over all these nuisance parameters. I therefore compute the sampling probability numerically.

In simulation, the population starts with N ancestor cells. Mutations with selection coefficient s (same class) would randomly arise in a rate of U (per genome per generation) across the evolving time. Once a mutation arises, it would randomly be extincted or survived with a survival probability (equation (4.25)) based on the model. If the mutation survives, it's

establishment time is a random variable draw from equation (4.21), thus the size at sampling time is determined (equation(4.19)). At the sampling time, I already have several mutations with various frequencies. Next, I randomly sample a total number of K clones, and record the number of clones (noted as j) that belong to this mutation. This simulation is repeated for $10^4 - 10^5$ runs to harvest all possible observations of j . Finally, I record the number of runs that I receive $j = k$, as the simulating observation j matches the true observation k . The sampling probability $P(k; s, U)$ is computed the fraction of matching runs ($j = k$) to the total runs ($10^4 - 10^5$).

4.4 Results and Discussions

4.4.1 The validity of strong- and weak-mutation approximations

First, I computationally demonstrate that both strong-mutation (equation (4.24)) and weak-mutation (equation (4.32)) approximations hold in some regimes. To this end, I compare the each analytical approximation to the numerically computed solution for a mutation that grows in a mean-fitness of logistic, exponential and linear trajectory. For all cases, the mean-fitness increases from 0 to 1 (figure 4.1A, figure 4.2A).

I find that, for a strongly beneficial mutant ($s = 0.8$), both strong and weak approximations are reasonably close to the true $b_1(t)$, but typically underestimate it (Figure 4.1). For a weakly beneficial mutant ($s = 0.4$), the strong approximation qualitatively fails to capture the shape of this function, whereas the weak approximation captures it well (Figure 4.2). Although the weak approximation for $b_1(t)$ is more widely applicable, it is computationally more intensive. Thus, if an analysis focuses on large-effect beneficial mutations, it is possible to apply the strong-effect approximation. I will make this assumption for the rest of this Chapter.

Next, I use numerical simulations to demonstrate the accuracy of the strong-mutation approximation for analytically computing the mutant's survival probability and the distribution of establishment times. In brief, I carry out branching process simulations of a nascent adaptive mutation in a population whose mean fitness increases from 0% to 5% over 300 generations.

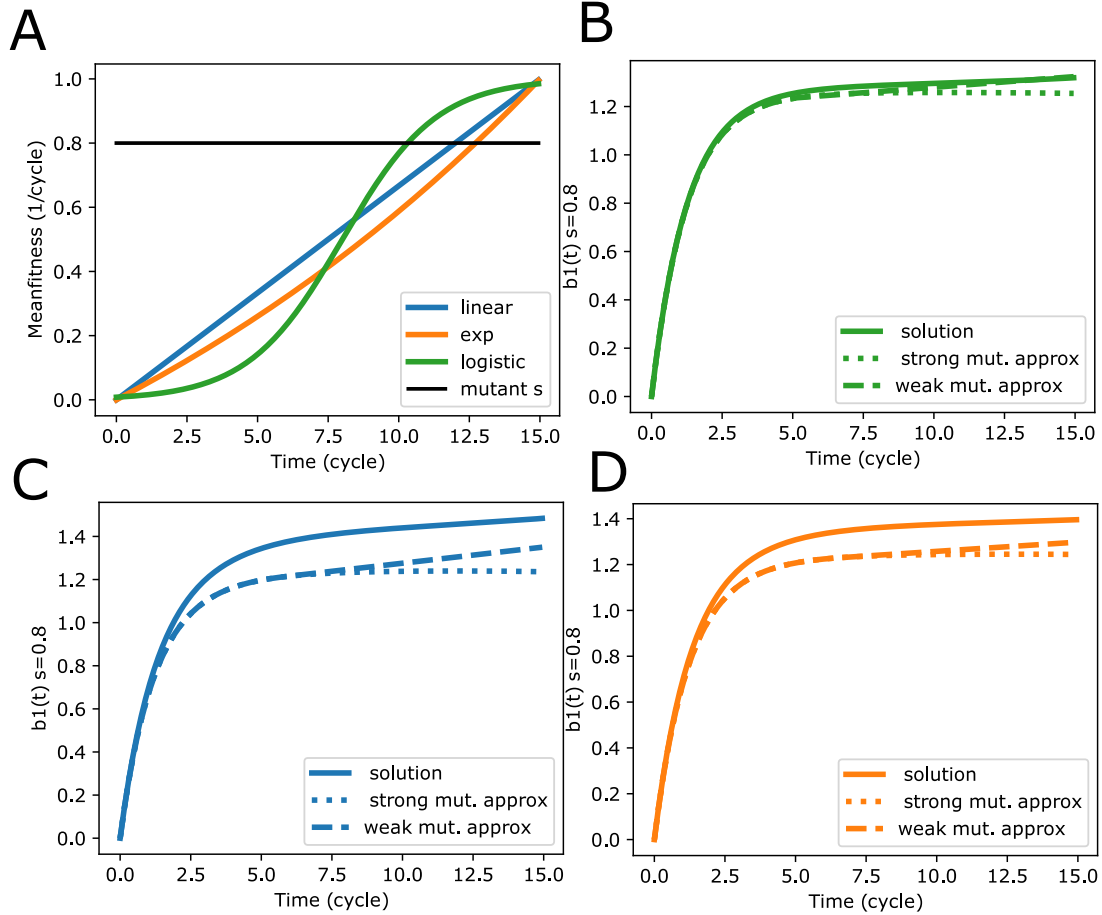


Figure 4.1. Strong and weak mutant approximations for function b_1 for a mutant with $s = 0.8$. I computationally test the validity of the strong mutant (given by the equation (4.24)) and weak mutant (given by the equation (4.32)) approximations for a mutant with $s = 0.8$ in three scenarios of mean fitness dynamics shown in panel (A): linear, exponential, logistic. Panel (B)–(D) show the numerically integrated function $b_1(t)$ (solid line) as well as the strong mutation (dotted line) and weak mutation (dashed line) approximations for the three scenarios. (B) Logistic mean fitness growth (green curve in panel (A)). (C) Linear mean fitness growth (green curve in panel (A)). (D) Exponential mean fitness growth (green curve in panel (A)).

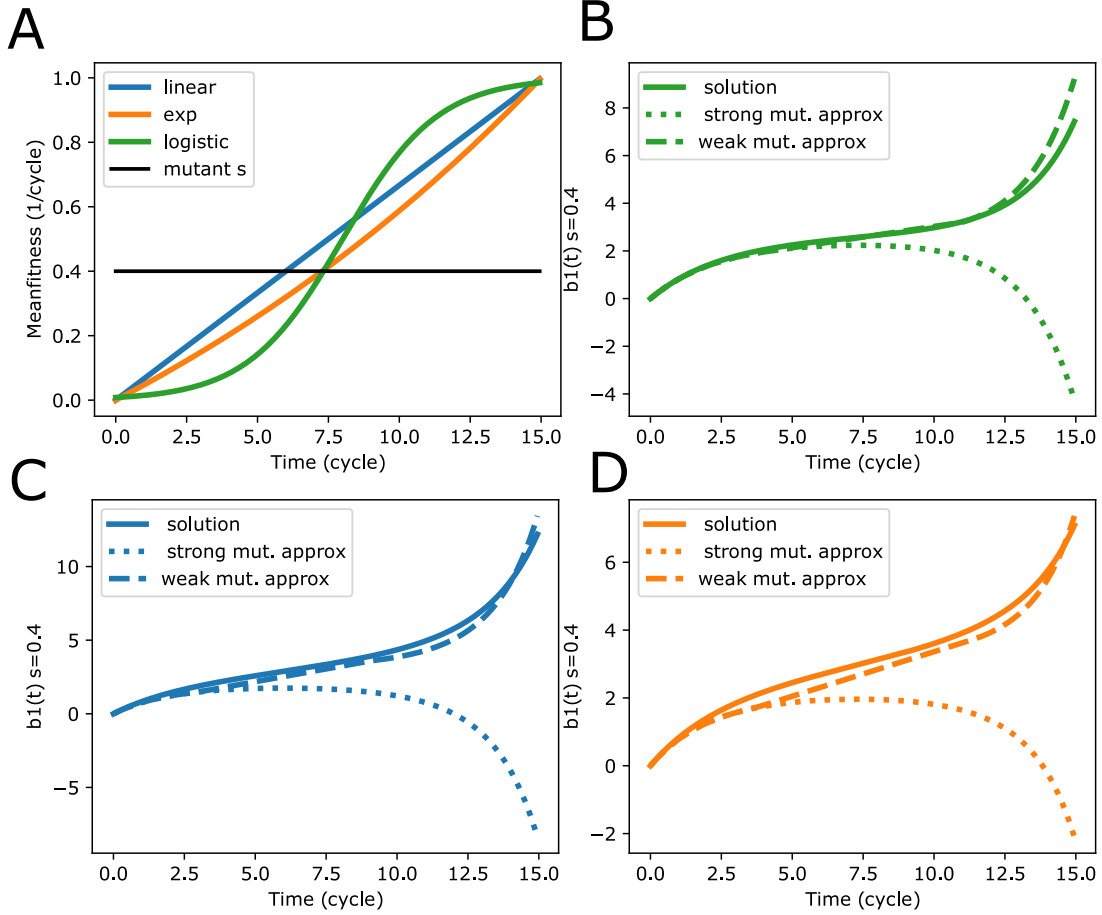


Figure 4.2. Strong and weak mutant approximations for function b_1 for a mutant with $s = 0.4$. Notations are the same as in Figure 4.1.

The mutant population starts with a single individual with a fitness benefit s relative to the rest of the population. If the mutant survives to the end, I compute the establishment time τ from the mutant's population size trajectory based on equation (4.19), as shown in Figure 4.3A. The simulation is repeated for $\sim 100,000$ runs until ~ 4000 runs have a survived mutant at the end. I then compute the empirical survival probability as a function of time and the empirical distribution of establishment times for those mutants that survived until the end of the simulation.

I find that the theoretical approximate solution based on equation (4.31) provides an excellent match to the empirical distribution of establishment times (Figure 4.3B). Likewise, the approximate survival probability obtained in equation (4.25) is in agreement with simulations (Figure 4.3C). As expected, the survival probability approaches a stable value in the long time

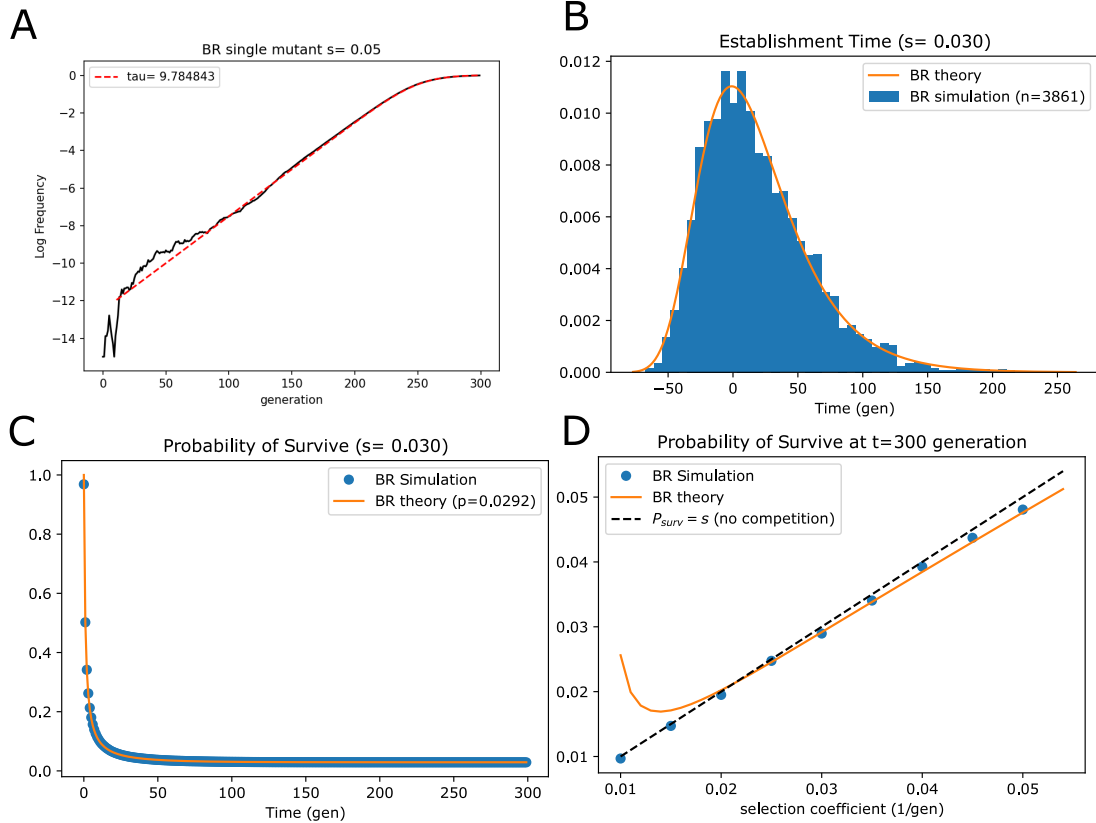


Figure 4.3. Theory with the strong-mutation approximation captures the distribution of establishment times and survival probability of a nascent adaptive mutant. The lineage size of a mutant with $s = 0$ is simulated using the branching process simulation with dynamically increasing mean fitness. (A) Trajectory of mutant population in one realization and the calculation of the establishment time. (B) A distribution of establishment times based on 10⁵ simulations (see text). (C) The probability of survival by time t . (D) Terminal survival probability as a function of to the fitness effect of the mutation.

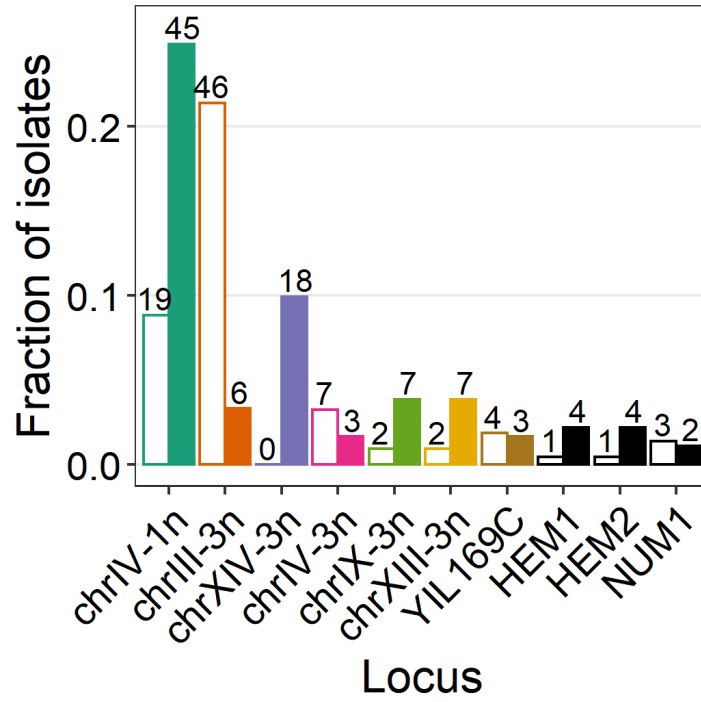


Figure 4.4. Distribution of adaptive mutations across the most common driver loci. Driver loci with 5 or more mutations are shown. Figur is reproduced from Ref. [102].

limit. It implies that, once the mutant escapes genetic drift while rare, its population grows deterministically and extinction is no longer possible.

Finally, I carry out these simulations for mutants with various fitness effects (s ranging from 1% to 5%). I find that my theory with the strong-mutation approximation (equation (4.28)) matches the empirical terminal survival probability well as long as $s \gtrsim 2\%$ (Figure 4.3D). I conclude that the analytical expressions for the survival probability until time t and for the distribution of establishment times obtained with the strong-mutation approximation can be applied for mutations with selection coefficients $s > 2\%$.

4.4.2 Rates of adaptive mutation in yeast

I apply the mutation-rate inference method to estimate the rates of adaptive mutations in yeast that were observed in a recently published BLT evolution experiment [102]. In this experiment, a barcoded population of yeast *Saccharomyces cerevisiae* was evolved in the presence

(C-condition) and absence (A-condition) of an ecological partner, the alga *Chlamydomonas reinhardtii*. Five replicate BLT experiments were carried out for each, A- and C-conditions. In each population, about 200 adapted clones were sampled at cycle 9 (~60 generation) from each of the conditions and their full genomes sequenced. Nine common mutation classes were discovered across both A and C mutants. Six classes of mutations, referred to as chrIV-1n, chrIII-3n, chrXIV-3n, chrIV-3n, chrIX-3n and chrXIII-3n, are aneuploidies. Three classes of mutations, referred to as YIL169C, HEM1, HEM2, are various single-nucleotide polymorphisms and small indels that occurred within or near the same gene coding region. Figure 4.4 shows the fraction of isolate for these 9 mutation classes in both conditions.

To estimate the mutation rate U_m of each mutation class m , I assume that all mutation from class m confer the same (known) fitness benefit s_m^A and s_m^C relative to the ancestor in the A- and C-condition, respectively. The fitness values and the number of isolation used are shown in Table 4.2. I define the log-likelihood function as

$$L(U_m) = \frac{1}{5} \sum_{i=1}^5 \log P(k_{im}^A; s_m^A, U_m) + \frac{1}{5} \sum_{i=1}^5 \log P(k_{im}^C; s_m^C, U_m), \quad (4.34)$$

which accounts for the fact that mutations are sampled from five replicate populations in the A-condition and from five replicates populations in the C-condition. The sampling probabilities $P(k_{im}^A; s_m^A, U_m)$ and $P(k_{im}^C; s_m^C, U_m)$ are computed numerically, as described in Section 4.3). In the simulation, all population have a constant size $N = 2 \times 10^6$, and all clones are sampled at time $T = 60$ generations. The mutation rate U_m is estimated by maximizing the likelihood function $L(U_m)$. The likelihood function for chrIII-3n mutations is shown as an example in Figure 4.5. The inferred rates of each mutation class are shown in Table 4.2 and Fig. 4.6. The rate of beneficial ranges from $10^{-6.65}$ to $10^{-4.14}$.

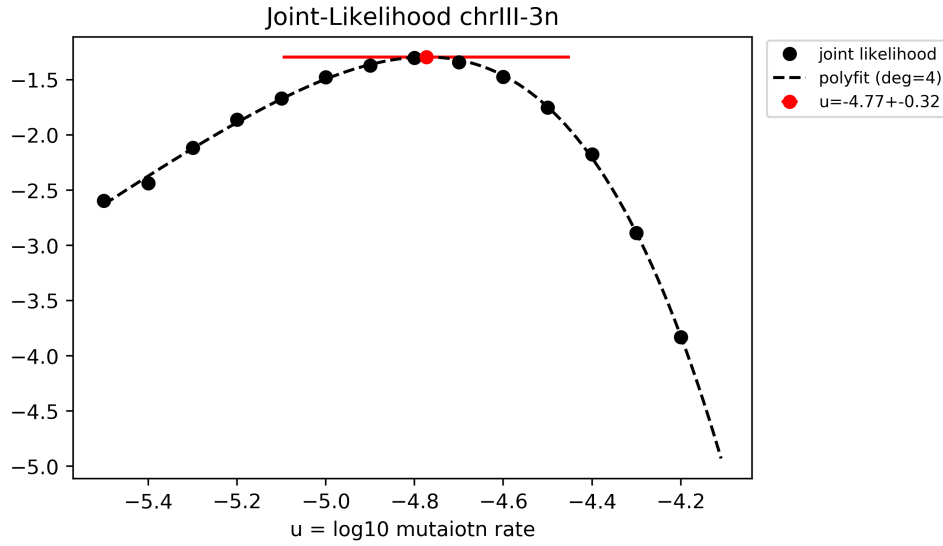


Figure 4.5. Estimation of the rate of beneficial mutations. The log-likelihood function for the estimation of the chrIII-3n beneficial mutation rate is shown as an example. Red point indicates the identified maximum value. Horizontal red bar shows the standard error calculated based on Fisher Information.

4.5 Discussion

In overall, I model the population dynamics of beneficial mutation in a competitive background. The dynamics is pictured as follows. For a mutant arises in a population, it might go extinct or survive in certain rate. For mutant who survives, it's population would drift stochastically before establishment time then grow deterministically afterward. I also work out the approximation of analytical solution in both strong and weak fitness region. The solutions are verified by simulation of branching process. Next, I apply this model to infer the mutation rate of beneficial mutations collected in yeast-algae BLT experiments. To infer the mutation rate, I construct likelihood function as the probability of observing clone of adaptive mutations. The mutation rate is inferred by maximizing the likelihood function. In result, I report the mutation rate from $10^{-6.65}$ to $10^{-4.14}$ for the top 9 common mutations in this experiment.

I find that the probability of observing adaptive mutation (Figure 4.6 A) is consistent with the number of mutant isolates (Figure 4.6 C) and depends on three factors: the fitness effect of

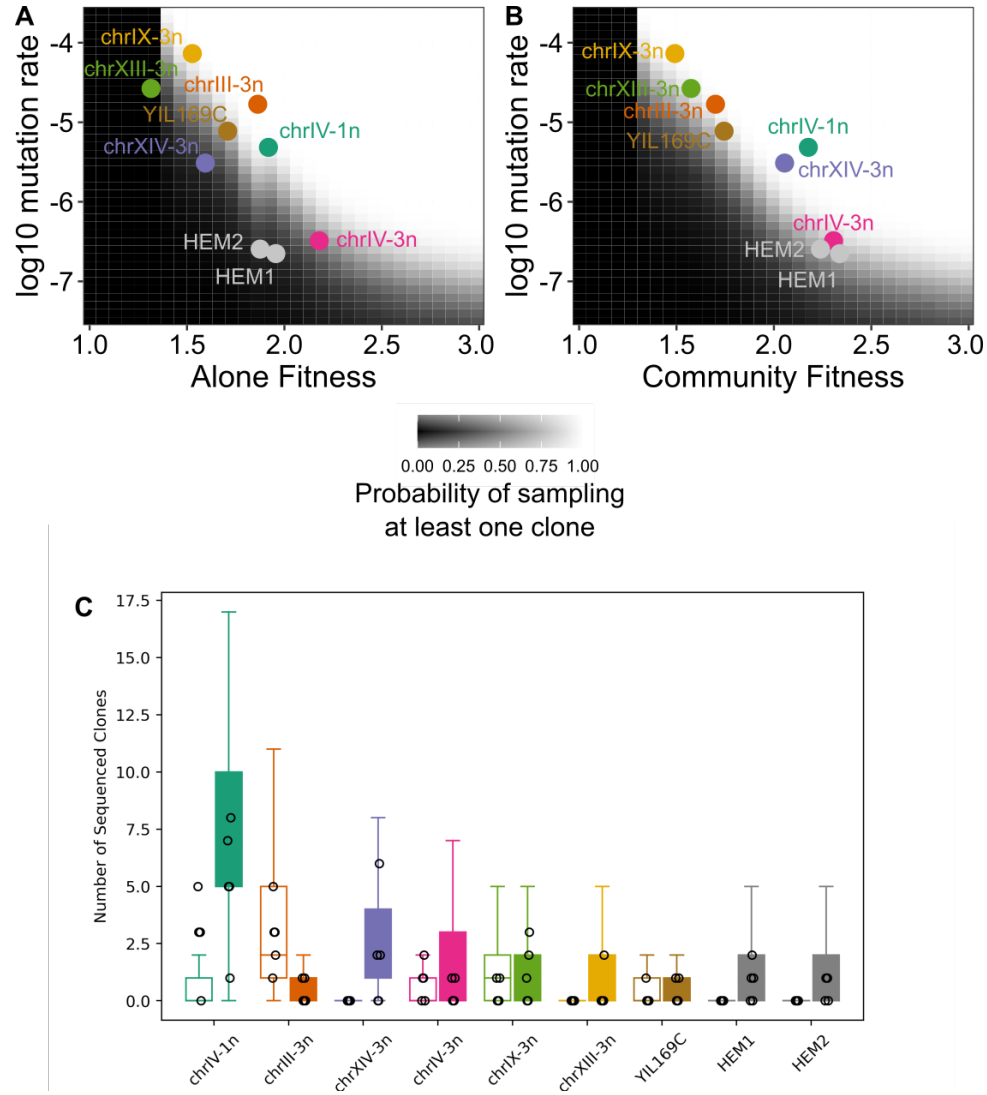


Figure 4.6. Probabilities of observing adaptive mutations at the most common driver loci in the whole-genome sequencing data. A. Shades of gray represent the probability of sampling at least one clone with a beneficial mutation that arises at a certain rate (y-axis) and provides a certain fitness benefit in the A-condition (x-axis). The most common driver loci are shown by points. The estimated beneficial mutation rate and the selection coefficient for each mutation class are given in Table S3 . B. Same as A but for the C-condition. The mutation rate for each locus is assumed to be the same in both conditions, but the selection coefficients vary. C. Black points show the number of sequenced clones with a given driver mutation found per replicate population in either A- (left) or C-condition (right, $n = 5$ replicate populations per condition). Box and whiskers show the distributions of these numbers found in our simulations (mid-line shows the median, boxes show the 25th and 75th quantiles, whiskers show the 5th and 95th quantiles)

mutations in the given class, the rate of mutations, and the strength of competition (mean-fitness trajectory). Indeed, it has been shown that the mean-fitness are different in two conditions (see Supplementary Figure 1 in Ref.[102]). The fact that the difference in the probabilities of observing A and C mutants also implies that the yeast populations in these two conditions are about to embark on distinct evolutionary trajectories. Therefore, the mutations that they carry are more likely to win the clonal competition towards fixation in the condition from which they were sampled. Thus, ecological interactions with the alga change the short-term evolutionary trajectory of yeast.

The workout in mutant growth model helps to simplify the simulation procedure. Since the population dynamics is stochastically in nature, the full simulation should consider the randomness of birth and death of a single cell for all generations. Simulating the entire population stochastically over hundreds generations is time consuming. However, with the workout of branching process model, one can picture the mutant dynamics by three questions: Whether this mutant survives? If yes, when does it escape the drift? If the mutant escapes the drift, what's the deterministic frequency of mutant? By answering the above question, the mutant growth can be simulated with only three quantities: survival rate, establishment time and the population frequency.

Lastly, in appendix B I demonstrate that the solutions derived from the branching process theory can also be applied to the Wright-Fisher model. The only difference between the Wright-Fisher and branching process is the survival rate and expected size. Mutant that grows in the Wright-Fisher model will have a double of survival probability but half of the expected population size to those in the Branching process model. Therefore, the population dynamics in the Wright-Fisher model is pictured as the follows. For a mutant arises in a population, it might go extinct or survive in certain rate (double of that in branching process). For mutant who survives, it's population would drift stochastically before establishment time (same as that branching process) then grow deterministically with frequency (half to that in branching process) afterward.

Acknowledgments

I am deeply thankful to Sandeep Venkataram for his inspiring discussions and generous provision of the experimental data utilized in Chapter 4.

Chapter 4, in its entirety, has been published in *Nature Ecology & Evolution* (2023): 1-12 (Ref. [102]) in the title of “Mutualism-enhancing mutations dominate early adaptation in a two-species microbial community.” Sandeep, Venkataram; Huan-Yu, Kuo; Erik F. Y. Hom; Sergey, Kryazhimskiy. The dissertation author was the researcher and author of this publication.

Table 4.2. Numbers of isolates carrying the most common adaptive mutations and the estimated mutation rates. Only clones with unique barcodes are shown. Column “Seq. isolates” shows the total number of sequenced isolates with unique barcodes sampled from each culture. Last row shows the estimated $u_m = \log_{10} U_m$.

Culture	Seq. isolates	Mutation class m /Locus									
		chrIV-1n	chrIII-3n	chrXIV-3n	chrIV-3n	chrIX-3n	chrXIII-3n	YIL169C	HEM1	HEM2	
A1	13	0	2	0	0	1	0	0	0	0	
A2	17	3	3	0	2	0	0	0	0	0	
A3	12	3	3	0	1	0	0	0	0	0	
A4	19	3	5	0	1	0	0	0	0	0	
A5	18	5	1	0	0	1	0	1	0	0	
C1	25	8	0	2	1	3	2	1	1	0	
C2	16	1	1	6	0	2	0	0	0	1	
C3	21	5	0	0	0	0	0	0	2	1	
C4	11	5	0	0	0	0	0	0	1	1	
C5	24	6	1	2	1	1	0	1	0	0	
	s_m^A	0.916	0.855	0.593	1.18	0.527	0.315	0.708	0.955	0.874	
	s_m^C	1.17	0.689	1.05	1.3	0.492	0.575	0.744	1.34	1.24	
	$\hat{u}_m$	-5.32	-4.77	-5.51	-6.49	-4.14	-4.58	-5.11	-6.65	-6.59	

Chapter 5

Estimating the Mutation Rate Spectrum from Lineage Tracking Data

5.1 Introduction

The rates at which mutations arise and how they affect fitness are fundamental evolutionary quantities that contribute to our understanding of adaptive evolution [22]. Despite its importance, the shape of the distribution of fitness effects (DFE), particularly its beneficial tail are poorly characterized.

Theoretical works in 1980's and 1990's proposed that the DFE of beneficial mutations might follow an exponential or normal distribution. Gillespie [30] first derived the exponential distribution of the DFE using extreme value theory. He argued that the “wildtype” has the highest fitness among all of its mutant neighbors in the home environment. As the environment changes, the fitness of both the wildtype and mutations is re-ranked. Since beneficial mutations are rare, Gillespie argued that if only one beneficial allele is accessible in the new environment, the fitness of the wildtype would be ranked second. By applying extreme value theory, the DFE is exponentially distributed if it belongs to the Gumbel family. Later, Orr [76, 78] generalized Gillespie's result to multiple beneficial alleles, and Joyce *et al.*[45] further extended the exponential prediction to include the Frechet and Weibull domains. Interestingly, Orr and Gillespie's work suggests that the exponentially distributed DFE holds regardless of other biological details, such as ancestor fitness, mutation rate, and environmental factors. It is worth emphasizing that

the exponential finding is based on one major assumption. To apply the extreme value theory, the fitness of the wild type must be in the tail of the distribution of its single-mutation neighbors, regardless of the environment. If this is true, the shape of the bulk of the distribution of fitness effects of mutations is irrelevant and only the properties of its tail matter. Whether this is a good assumption in practice remains unknown.

An alternative approach is to derive the DFE in an explicit model of a biological organism, such as the Fisher's geometric model (FGM) [24, 79]. In brief, FGM characterizes organisms as points in a multidimensional space of phenotypic traits. In this space, random mutations are represented by isotropic vectors that move an ancestral phenotype to the mutant phenotype. Fitness is assumed to decrease quadratically from a single phenotypic optimum. By applying FGM, the DFE can be well approximated by a shifted negative gamma distribution, which converges toward a normal distribution in high-dimensional phenotypic complexity. It is important to note that this prediction relies on strict assumptions such as isotropic mutations, independent phenotypic traits, the existence of an optimal phenotype, and a quadratic function for the fitness-to-phenotype map.

Experimental biologists have been interested in determining the empirical DFE and comparing it to the theoretical predictions ever since the techniques for isolating adaptive mutations became available. Some experiments that sample small numbers of nascent beneficial mutations have shown results that are supportive, similar, or opposite to the theoretical predictions. For instance, Kassen and Bataillon [47] derived 665 independent single-step mutations in the bacterium *P. fluorescens* using a conventional fluctuation-style assay. They measured the fitness effects of mutants ($5\% \leq s \leq 25\%$) and obtained an exponential tail of the DFE across different environments. Hietpas *et al.*[39] created point mutations in a region of the Hsp90 gene in *Saccharomyces cerevisiae* and identified 38 and 24 beneficial mutants ($1\% \leq s \leq 8\%$) in two conditions of elevated salinity. Their observations of 181 mutations (including deleterious, neutral, and adaptive) in different environments are supported by the one-dimensional FGM. Sanjuan *et al.*[88] created 91 single mutant clones of vesicular stomatitis virus using site-directed

mutagenesis. They performed competition experiments to measure the relative fitness of each mutant ($0\% \leq s \leq 6\%$) and concluded that the DFE of beneficial mutations follows a gamma distribution, which is similar to both the exponential prediction by Orr and Gillespie and the negative gamma prediction by FGM. On the other hand, Rokyta *et al.*[85] selecting 25 novel mutants of two sets of viruses, bacteriophage *ID11* and RNA phage $\phi 6$, rejected both hypotheses and found that their distribution of beneficial mutants ($3\% \leq s \leq 8\%$) is consistent with a uniform distribution.

Although the aforementioned studies have contributed to characterizing the DFE, the resolution of the DFE tail is limited by the number of beneficial mutations analyzed (ranging from tens to a few hundreds). These limitations arise from the rarity of beneficial mutations and the difficulty of detecting mutants with weak effects. Recently, the barcode lineage tracking (BLT) technique has made it possible to achieve a high resolution in the empirical measurement of the DFE, including the region of small effects. The BLT experiment utilizes unique DNA barcodes as heritable markers to track microbes that have evolved in serial batch culture. By deep-sequencing the barcodes, the barcode frequency can be tracked down to very low limits (ranging from 10^{-5} to 10^{-7}). When a novel adaptive mutation arises, it leads to an observable increase in its barcode frequency. Therefore, the fitness effects of beneficial mutations can be inferred from the rising trajectories.

BLT is a powerful technique that allows one to observe many lineages that carry individual adaptive alleles and is perfectly suited for estimating the DFE. However, doing so is not trivial. The naive way of estimating the DFE from BLT data is to simply bin adapted lineages by the estimated fitness effects of the mutations that they carry and construct a histogram of such lineages. I refer to this binned count data as the **observed DFE**. What is the relationship between the observed DFE and the true DFE? In theory, the DFE describes the probability that a newly arisen mutation has a fitness effect within a certain range. In contrast, the observed DFE estimates the probability that a mutation arose, escaped genetic drift while rare and then sufficiently expanded population size prior to the end of the experiment to be detected by the

experimenter. Thus, in addition to the origination probability the observed DFE accounts for the probability of the mutant lineage to survive genetic drift, expand to a certain size and then be detected as adapted in the BLT analysis. Since adaptive mutations with higher selection coefficients are more likely to survive genetic drift and expand faster, we expect that the observed DFE would be right-shifted compared to the true DFE. Fortunately, it is possible to correct for these biases.

By applying BLT, Levy *et al.*[58] identified approximately 25,000 lineages of *Saccharomyces cerevisiae* and acquired novel beneficial mutations in short-term evolution spanning 168 generations. They used a branching process model to estimate the survival probabilities and the population sizes of mutant lineages. They used these quantities to estimate the true DFE from the observed DFE. Contrary to both hypotheses mentioned above, they observed a multimodal DFE, with most mutants confined to a narrow range of fitness effects ($2\% \leq s \leq 5\%$), and two peaks in the fitness ranges of 7–8% and 10–11%. Blundell *et al.*[7] used a similar approach and also detected a multimodal DFE ($3\% \leq s \leq 12\%$ in carbon limitation and $2\% \leq s \leq 7\%$ in nitrogen limitation) from beneficial mutants in *Saccharomyces cerevisiae*.

These previous attempts to estimate the DFE have two important shortcomings. First, both studies relied on the branching process under the assumption that the population's mean fitness stays constant. However, these authors in fact observed a mean-fitness increase of 12% and 3% in carbon and nitrate-limited environments, respectively. Ignoring the mean fitness increase leads to overestimating the chances of survival for weakly beneficial mutants which in turn leads to an underestimation of the rate at which weakly beneficial mutants arise. The second shortcoming of the previous analysis is that it does not account for the detection process. Since weak mutations are more likely to occur in small numbers and go undetected, a failure to model the detection process is also expected to result in an underestimation of the rate of weakly beneficial mutations.

The goal of this chapter is to derive a better estimate for the DFE from the BLT data. More precisely, I describe a method for estimating the **mutation rate spectrum** of beneficial

mutations which is defined as the rate at which mutations with certain fitness effects originate (per unit time per individual). The theoretical DFE is the normalized mutation rate spectrum, i.e., spectrum divided by the total rate at which mutations arise. I propose a method for inferring the mutation rate spectrum from BLT experiments. The first part of the method is the Wright-Fisher model, which describes the expansion of the mutant population in a clonal interference regime. The second part is the empirical calculation of the detection probability for mutants. Together, this method aims to bridge the gap between experimentally measured DFE and the theoretical mutational rate spectrum.

5.2 Estimating the mutation rate spectrum

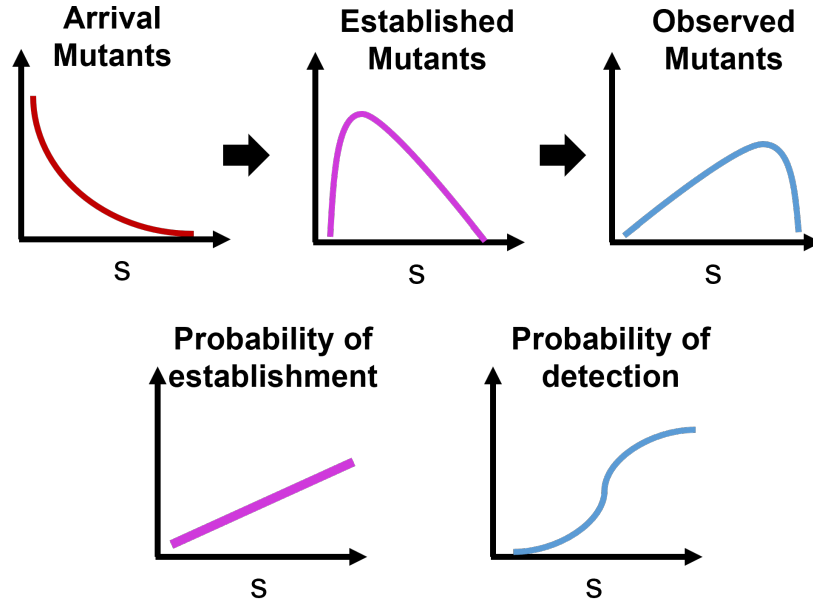


Figure 5.1. Schematic plot for the observed DFE resulting from the arrival mutants, probability of mutant establishment $P_{est}(s)$ and probability of detection $P_{det}(s)$.

Suppose that beneficial mutations occur in the population at rate U_b (per genome per generation) and that a new beneficial mutation provides a fitness advantage s with probability $D_t(s)$ (here, subindex “t” stands for “true” DFE). In other words, a mutation with selection coefficient s arises at rate $\mu(s) = U_b D_t(s)$ per generation. I refer to $\mu(s)$ as the mutation rate

spectrum mDFE.

To infer the $\mu(s)$, I consider the processes by which a nascent beneficial mutant linked to a barcode is detected as an adapted lineage in the BLT experiment. To be detected, the mutant must survive genetic drift while rare (biological bias) and be identified as adapted in the BLT analysis (technical bias). Due to these biases, the observed DFE is not identical to the original mutation rate spectrum (Figure 5.1). To model the biological bias, I model how mutants survive and establish by the Wright-Fisher model. The model provides analytical description for the establishment probability of a mutant as a function of its selection coefficient s . To model the technical bias, I use simulations to compute the empirical probability that an established mutant is detected as adapted. With the descriptions of establishment and detection probabilities, the mutation rate spectrum $\mu(s)$ can be derived from the observed DFE, as follows.

Consider a population of size N that evolved for t generations. The expected number of beneficial mutations with fitness effect s that arise during this time is

$$n_b(s, t) = \mu(s) N t. \quad (5.1)$$

Each mutation will either go extinct or will establish. Denote the probability that a mutant that arose at time t_0 establishes by time $t \geq t_0$ as $P_{est}(s, t, t_0)$. The stochastic process of establishment is modeled as the Wright-Fisher (see Section 5.3.1). Then, the total number of mutations that establish at by the sampling time t is

$$n_{b,est}(s, t) = \mu(s) N \int_0^t dt_0 P_{est}(s, t, t_0), \quad (5.2)$$

and the total number of established mutants up to the last sampling time T is

$$n_{b,est}(s) = \mu(s) N \frac{2}{T} \int_t^T dt \int_0^t dt_0 P_{est}(s, t, t_0) = \mu(s) N T P_{est}(s), \quad (5.3)$$

where

$$P_{est}(s) = \frac{2}{T^2} \int_t^T dt_0 \int_{t_0}^T dt P_{est}(s, t, t_0) \quad (5.4)$$

Since the Wright-Fisher model provides a framework of stochastic process, it is easy to simulate the cell frequency trajectories $f(s, t, t_0)$ of mutants and their corresponding barcode read count data. With the simulated data, I run it through the Bayesian Filtering Method (chapter 3) for estimating fitness and identifying adaptive lineages. After repeated runs, the value of detection probability is empirically computed analogous to equation (5.4)

$$P_{det}(s) = \frac{2}{T^2} \int_t^T dt_0 \int_{t_0}^T dt P_{det}(s, t, t_0), \quad (5.5)$$

where $P_{det}(s, t, t_0)$ is the probability that an established mutant with selection coefficient s that arose at time t_0 will be detected by time t .

Together, the number of mutant that occurs, survives drift and grow to detectable size is

$$n_{b,est,det}(s) = \mu(s) NT P_{est}(s) P_{det}(s). \quad (5.6)$$

It is also the expected number of unique barcode which is identified as adapted in the Wright-Fisher & Bayesian framework. Finally, the mutation rate spectrum can be inferred as

$$\tilde{\mu}(s)ds = \frac{\text{number of observed barcode with in } [s, s + ds]}{NT P_{est}(s) P_{det}(s)}. \quad (5.7)$$

5.3 Results

5.3.1 Analytical approach to the Wright-Fisher model for mutant population expanding in a background of increasing mean-fitness

The population dynamics are generally modeled with the Wright-Fisher model. However, this model has limited analytical tractability, and various approximations are necessary to obtain quantitative results. One approach that has proven successful is to use the branching process

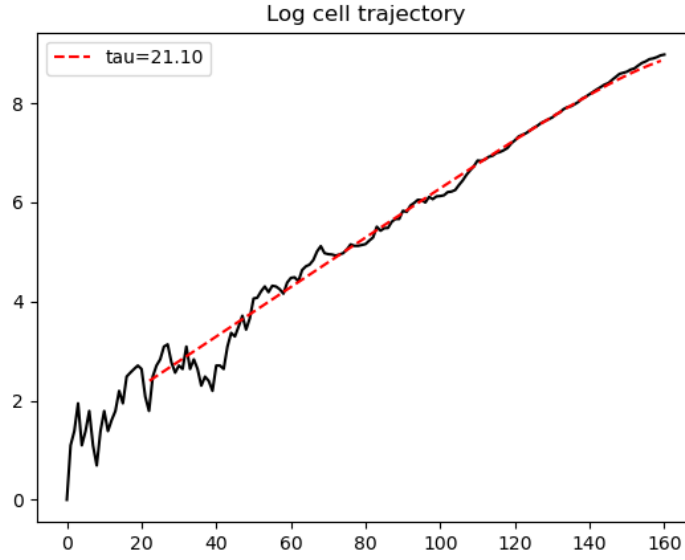


Figure 5.2. A population trajectory of mutant ($s = 5\%$, arising at $t_0 = 0$ generation) over simulating time (black line, y-axis: log population size, x-axis: time in generation), fitted by theoretical trajectory (red dash line) with best-fit of establishment time $\tau = 21.1$ generation.

theory for modeling the frequency dynamics of a lineage while it is rare. In thesis chapter 4, I extend this analytical approach to model multiple lineages that carry independently arisen mutations from the same class that competes against the rest of the population whose mean fitness continuously increases. Here, I further extend my analytical solutions of branching process theory to the Wright-Fisher model.

First, I demonstrate that the Wright-Fisher model has double the survival probability (half the genetic drift) than the branching process for mutant of the same selective advantage s . I run WF simulation with a single targeted mutant of s and $N - 1$ ancestor cell. The population are undergo an increasing mean fitness due to random mutants. At the end time $T = 160$ generation, if the targeted mutant survives, the population trajectory (Figure 5.2 as example) is recorded.

In the long run, a mutant population has only two fates: either establish and expand or go extinct. As expected, the survival probability of a nascent mutant (Figure 5.3) decreases over time and reaches a stable value. I refer to such stable value as long-term survival probability.

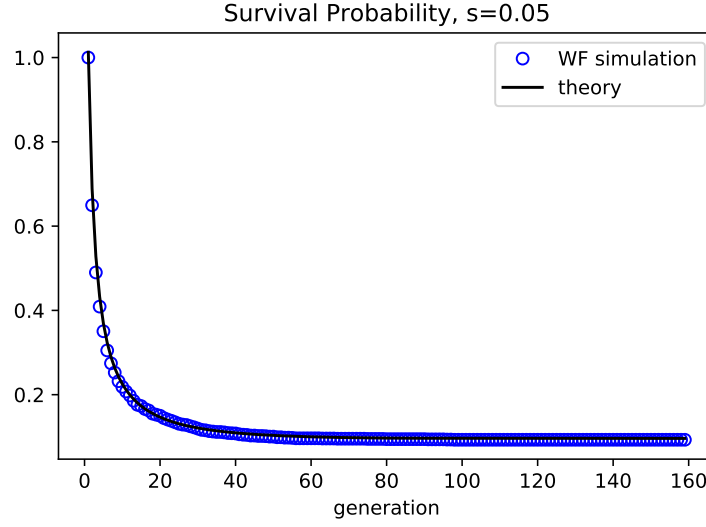


Figure 5.3. Survival probability for a mutant of $s = 5\%$ arising at $t_0 = 0$ in the Wright-Fisher model (blue circle) over time, compared to the theoretical prediction (black line).

	Branching process	Wright-Fisher model
Zero mean-fitness ($\bar{s} = 0$)	$s/(1 + s)$	$2s/(1 + s)$
Increasing mean-fitness $\omega = s - \bar{s}(t_0)$	$\omega/(1 + \omega)$	$2\omega/(1 + \omega)$

Table 5.1. The long-term survival probability for a mutant of s , arising at time $t = t_0$ and expanding in a zero mean-fitness ($\bar{s} = 0$) or an increasing mean-fitness background, compared between branching process and the Wright-Fisher model.

The long-term survival probability is also the fixation probability for the mutant expanding in zero mean fitness. The long-term survival probability in the branching process theory is $s/(1 + s)$ [17]. In the Wright-Fisher model, the fixation probability is expected to be twice as large ($\sim 2s$) [20, 36, 50, 107]. To learn such a relationship in an increasing mean-fitness background, I repeat the above procedure for targeted mutants with different s and t_0 . As a result, for the mutant arises at time t_0 , its long-term survival probability is $2\omega/(1 + \omega)$ for the mutant, where ω is the selective advantage s relative to the mean-fitness value $\bar{s}(t_0)$ (Table 5.1 & Figure 5.4).

The establishment time τ (when population reach $n_c = 1/\text{long term survival probability}$) is a random variable. To learn the population stochasticity of WF model, I take the mutant

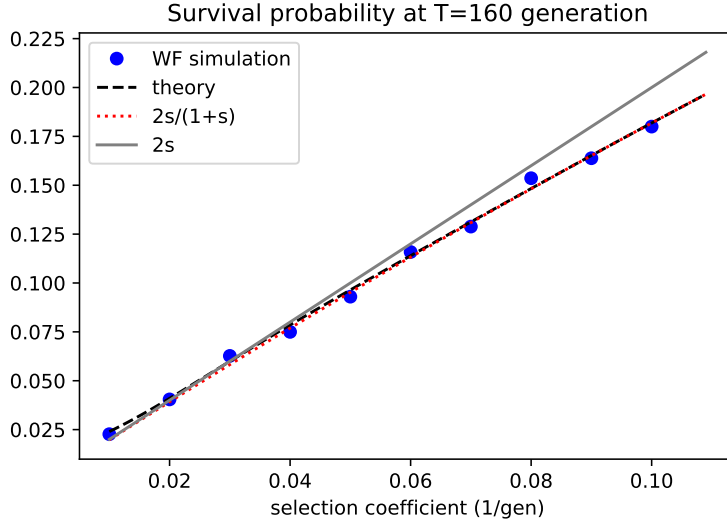


Figure 5.4. Survival probability at time $T = 160$ generation for a mutant arising at $T = 0$ in the Wright-Fisher simulation (blue circle), theoretical long-term stable survival probability with competitive background (red dot line), and theoretical survival probability at 160 generations (black dashed line).

trajectories (Figure 5.2) to infer τ , and construct the distribution of establishment $P(\tau)$ by repeated runs. Figure 5.5 shows that the $P(\tau)$ in WF simulation is the same as the analytically description in the branching process. It indicates that a mutant (of given s) would take same time in the two models to pass the genetic drift, despite of their differences in drift strength.

5.3.2 The establishment probability in the Wright-Fisher model

The establishment probability $P_{est}(s, t, t_0)$ of mutant (with s , arising at t_0) at time t can be derived in two ways. First, establishment occurs when the mutant population exceeds a threshold size n_c , leading to the probability equation

$$P_{est}(s, t, t_0) = \int_{n_c}^{\infty} dn P(n|s, t, t_0). \quad (5.8)$$

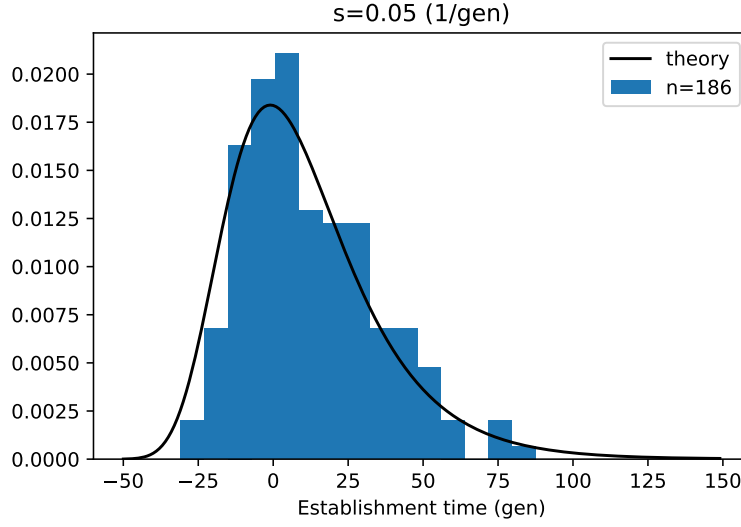


Figure 5.5. Probability distribution of established time for a mutant of $s = 5\%$ arising at $t_0 = 0$ in the Wright-Fisher model (blue histogram), compared to the theoretical distribution (black line) derived from the branching process.

Second, establishment is determined by the mutant being established at time t , meaning that $t_0 + \tau < t$. Consequently, the probability equation becomes

$$P_{est}(s, t, t_0) = \int_{-\infty}^{t-t_0} d\tau P(\tau|s, t, t_0). \quad (5.9)$$

Theoretically, these two approaches yield equivalent results, and the choice of method depends on convenience. The former method is suitable for Wright-Fisher (WF) simulations, while the latter is advantageous for analytical predictions. With the establishment probability $P_{est}(s, t, t_0)$, the overall establishment probability $P_{est}(s)$ (Figure 5.1) can be calculated using Equation 5.4.

To validate the analytical description of $P_{est}(s)$, I conducted a Wright-Fisher simulation to obtain the simulated $P_{est}(s)$. The simulation commenced with 2×10^6 ancestral cells, where beneficial mutations randomly arose based on a general exponential distribution (Figure 5.8) during the evolutionary period from 0 to 160 generations. For each arising mutation, if its population size ever exceeded the critical size, the mutant was marked as established in the WF simulation. Subsequently, the analytically described $P_{est}(s)$ from both methods were compared

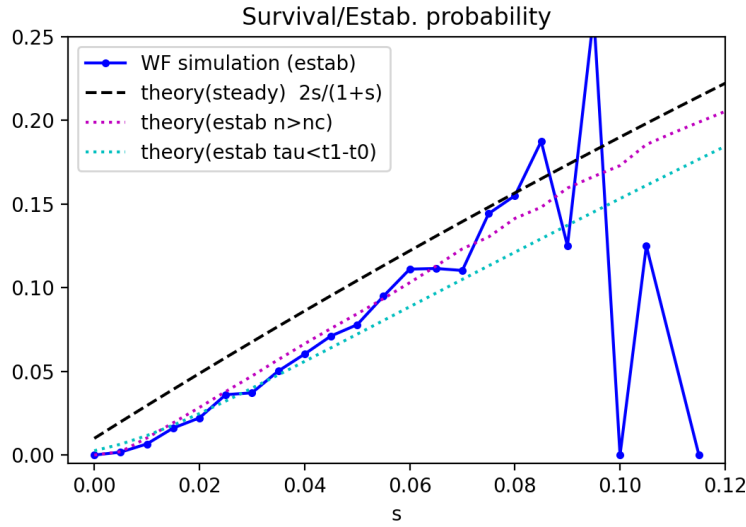


Figure 5.6. Establishment probability for mutants that randomly arise during the simulating time in the Wright-Fisher model (blue circle), compared to the theory (cyan dash line).

to the establishment probability in the WF simulation.

The results showed good agreement between both ways of calculating establishment probabilities (from population size: pink dot-line in Figure 5.6, from establishment time: cyan dot-line in Figure 5.6) and the simulation $P_{est}(s)$ (blue circles in Figure 5.6). It is worth noting that the establishment probability lies below the expectation on long-term survival probability (black dashed line, $2s/(1+s)$) with zero mean-fitness. This discrepancy arises due to the increase in mean fitness, resulting in a lower likelihood of survival for mutants that arise late in evolution. Furthermore, in this case, the establishment probability diminishes to zero when s is below 1.2%. Mutants with a fitness effect lower than this threshold are not anticipated to be established or observed.

Figure 5.7 illustrates the resulting established distribution of fitness effects (DFE). The predicted established DFE (cyan dashed line in Figure 5.1) is derived by multiplying the establishment probability (Figure 5.6) with the arising mutants (Figure 5.8) used in this case.

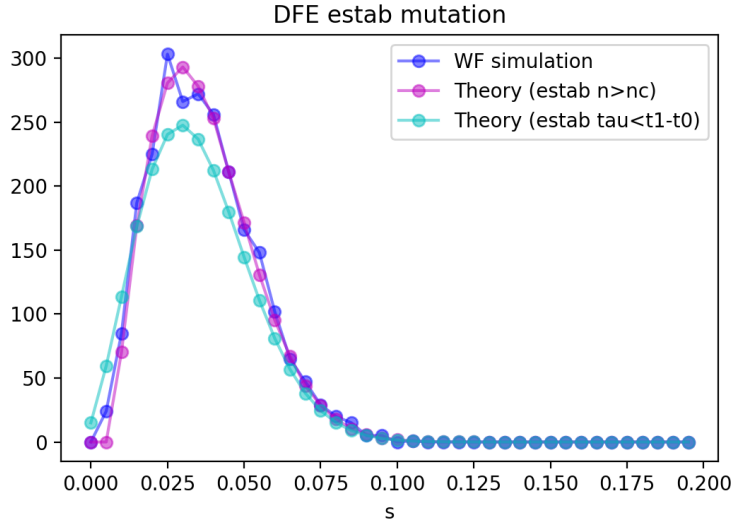


Figure 5.7. Distribution of fitness effect (DFE) for the established mutants that randomly arise during the simulating time in the Wright-Fisher model (blue circle), compared to the theory (cyan circle).

5.3.3 Empirical calculation on the detection probability of established mutants

Detecting adaptive lineages in the BLT experiment involves multiple procedures, such as barcode sampling, fitness effect inference, and software analysis to identify adaptive lineages. To determine the detection probability as a function of s , two layers of challenges must be considered. First, there is inherent randomness in mutation occurrence and size, as well as in barcode sampling, all of which contribute to the stochastic nature of lineage detection. Second, the specific details of the software used for analysis influence the inference procedures, making it nontrivial to quantify the computational process and its success rate. Consequently, deriving analytical descriptions for detection probabilities is challenging. However, performing all stochastic processes and detection procedures on a computer is straightforward. Thus, calculating an empirical probability of detection through repeated procedures becomes a more viable approach.

To estimate the probability of detection, I propose an empirical calculation method, as

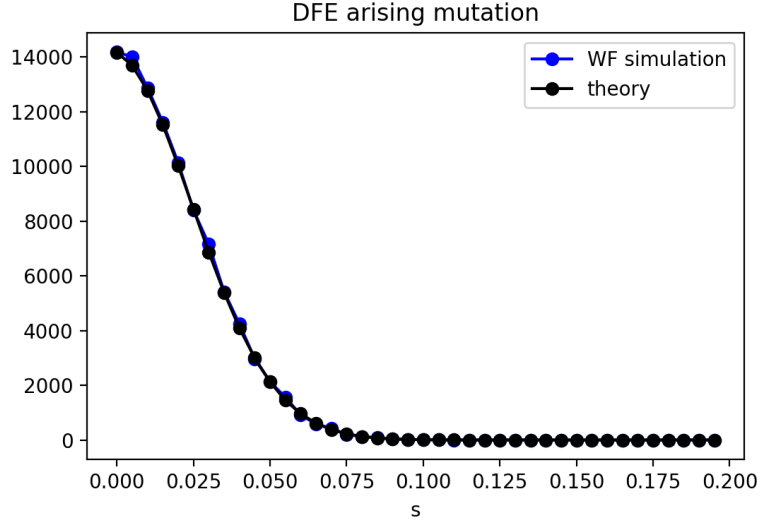


Figure 5.8. Arising beneficial mutations $NT\mu(s)$ from input function (black circle line) and simulation sampling results (blue circle line) with population size $N = 2 \times 10^6$, total time $T = 160$, and mutational rate spectrum as general exponential distribution $\mu(s) = U_b e^{-(s/\sigma)^\beta} / \sigma / \Gamma(1 + \beta^{-1})$ with $U_b = 3 \times 10^{-4}$, $\beta = 2$, $\sigma = 3.8\%$.

illustrated in Figure 5.9B. Since the establishment probability has already been derived (Section 5.3.2), I focus on the detection probability for mutants that have already been established. Starting with a single mutant cell with fitness effect s , I simulate the lineage trajectory using the Wright-Fisher model, which involves randomly sampling the arrival and the establishment times. Subsequently, I sample the barcode read count for a given established lineage to model experimental sample collection. Utilizing the obtained barcode read data, I apply my Bayesian filtering program for analysis. If the program identifies the lineage as adaptive, it is considered a successful detection. I repeat this process 1000 times for mutants with the same fitness effect. The success rate in these 1000 runs serves as an empirical estimate of the detection probability.

The empirical probability of detection is illustrated in Figure 5.10. For strongly beneficial mutants ($s > 5\%$), the detection probability stabilizes at about 84%. The maximum detection probability is less than 100% due to the requirement of having evidence present in at least two consecutive sampling times to confidently identify an adaptive lineage. Thus, mutants that establish late in the experiment will never be called as adapted. For mutants with medium fitness

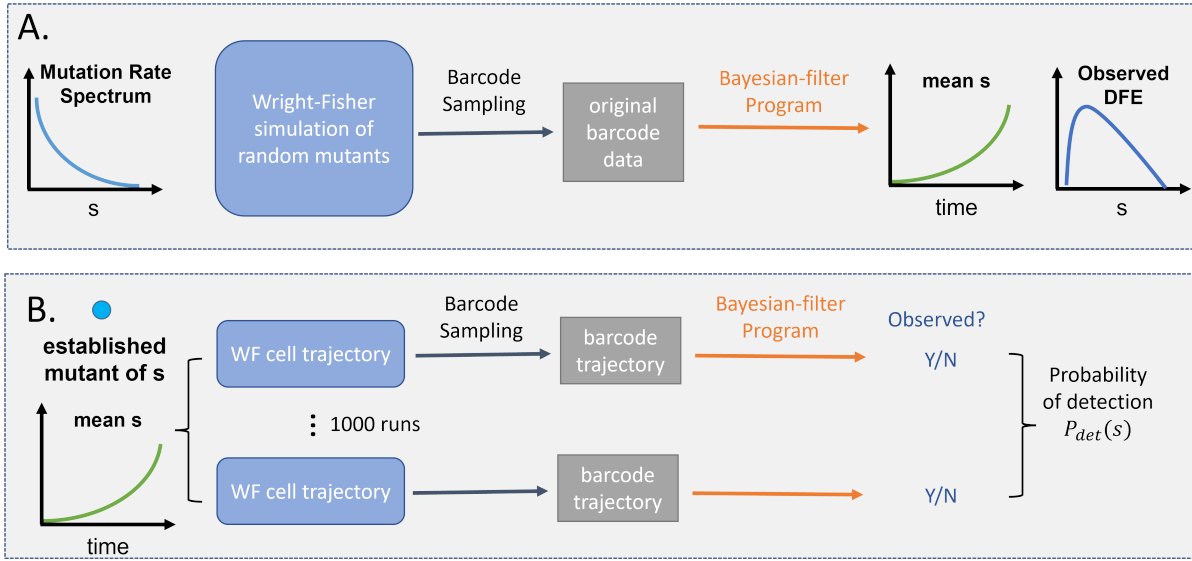


Figure 5.9. Schematic Plots (A) Obtaining the inferred mean-fitness trajectory and observed DFE by applying the Bayesian filtering program on simulated barcode data, generated by Wright-Fisher model with random mutants drawn from a known mutational rate spectrum. (B) Estimating the probability $P_{det}(s)$ of detecting established mutation of s , by empirically computations of successful observation events among 1000 runs of random cell trajectories of established mutant expanding under known mean-fitness.

effects ($1\% \leq s \leq 5\%$), the detection probability is sensitive to the fitness effects, ranging from 8% to 80% for s in the range of 1% to 5%. Interestingly, in this study, the detection probability reaches zero (as determined by a logistic function fit, represented by the black dashed line in Figure 5.10) at $s = 0.3\%$, indicating that mutants with effects less than 0.3% are undetectable.

It is important to note that the empirical calculation of the detection probability requires prior knowledge of the mean-fitness trajectory. This is logical since mutant population dynamics are influenced by the mean-fitness. When the mean-fitness increases rapidly, weak mutants, although established, will occur at low frequencies in the population and are less likely to be detected. Thus, the mean-fitness trajectory significantly affects the detection probability. Consequently, to estimate the detection probability, it is necessary to specify the mean-fitness. In this study, the mean-fitness trajectory was obtained by applying the Bayesian program to the simulated barcode data. In practical applications, the original data represents experimental

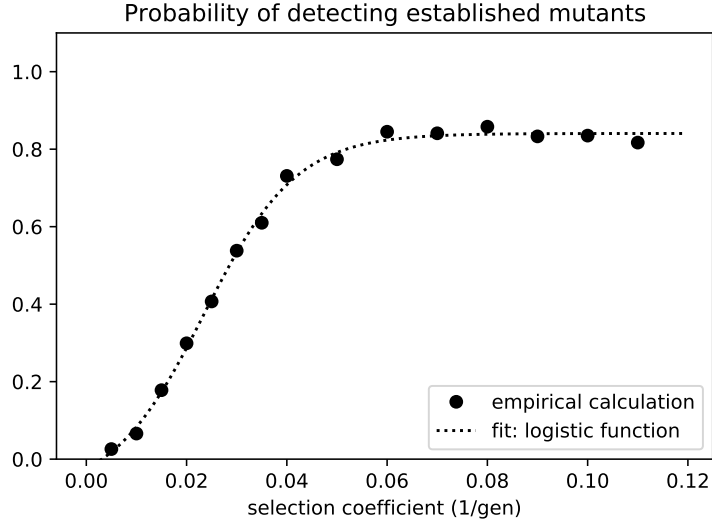


Figure 5.10. Probability of detecting established mutants with fitness effect s by empirical computation (black circle), fit by logistic function (black dashed line) $P_{det}(s) = (0.840 - 0.09e^{-109(s-0.0235)}) / (1 + e^{-109(s-0.0235)})$

measurements of sequencing counts.

In the WF simulation performed in this study, the experiment began with 2×10^6 ancestor cells, each carrying a unique DNA barcode (a total of 2×10^4 barcodes). Beneficial mutations arose randomly based on a general exponential distribution over the evolution time from 0 to 160 generations (Figure 5.8). Following the standard Wright-Fisher model, offspring from both ancestors and mutants were randomly selected for the next generation, while the total population size remained fixed ($N = 2 \times 10^6$). For each mutation, if its population size exceeded the critical threshold, the mutant was marked as established. Every 16 generations, barcode read counts were randomly sampled using the negative-binomial measurement process described in Section 3.2.2. Subsequently, the Bayesian filtering program was applied to infer the mean-fitness trajectory (Figure 5.11) and the observed distribution of fitness effects (Figure 5.12). The entire schematic process is depicted in Figure 5.9A.

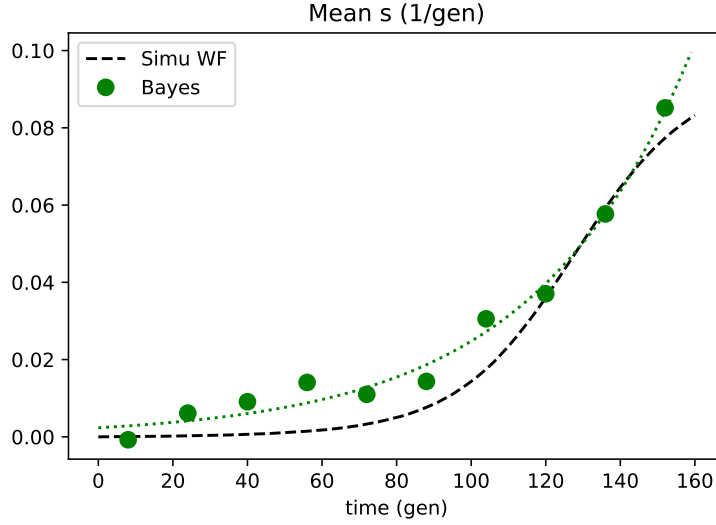


Figure 5.11. Mean fitness trajectory for the Wright-Fisher simulation (black dash), Bayesian filtering program inference (green circle), and logistic function fit (green dotted-line) to the Bayesian inference.

5.3.4 Inference of the mutational rate spectrum

The estimation of the mutational rate spectrum, denoted as $\mu(s)$, relies on various factors, including the population size ($N = 2 \times 10^6$) and the total duration of the experiment ($T = 160$ generations). Additionally, it depends on the observed distribution of fitness effects (DFE) as illustrated in figure 5.12, the probability of mutant establishment as depicted in figure 5.6, and the probability of mutant detection as shown in figure 5.10. Utilizing all the previously obtained results, I derived an estimated $\mu(s)$ by back-calculating the observed DFE based on equation 5.7. However, the resulting estimation curve (red circle line, Figure 5.13) deviates from the theoretical $\mu(s)$ (black line, Figure 5.13) used in this study. Specifically, for strong mutants ($s > 8\%$), the mutation rate is overestimated by a factor of roughly 10, whereas for weak and medium mutants ($s < 5\%$), my inference of $\mu(s)$ is an underestimate.

To investigate the underlying cause of the observed discrepancy, I conducted a comparison between various elements. This involved analyzing the observed distribution of fitness effects (DFE) obtained from the Bayesian framework, as well as the DFE established through back-

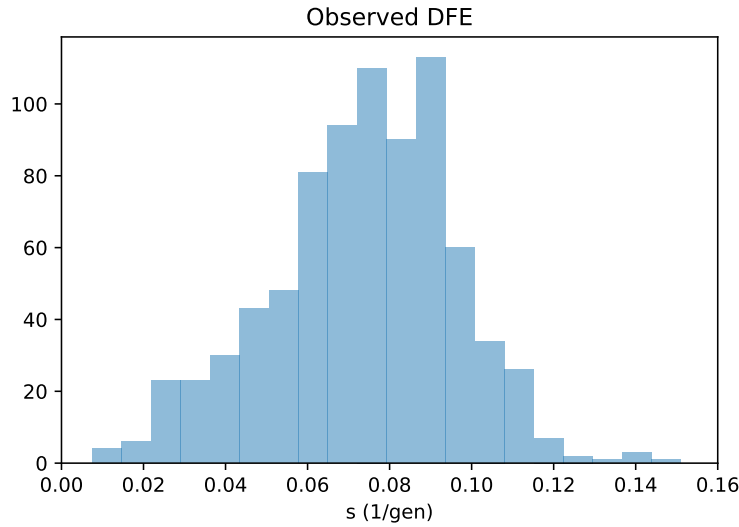


Figure 5.12. Observed distribution of fitness effect (DFE) by applying the Bayesian filtering program to the Wright-Fisher simulation data.

calculation. These were then compared to the established DFE obtained from simulation, along with its forward-calculated observed DFE. The findings, as depicted in figures 5.15 and 5.16, indicated that the detection probability did not contribute to the observed disparity between the observation (Figure 5.15) and simulation (Figure 5.16) spectra. Notably, the observed DFE exhibited peaks around $s = 8\%$, whereas the simulated DFEs peaked around $s = 4\%$. To directly compare the differences, the fitness effects inferred by Bayesian analysis for the identified lineages ($n=378$) are plotted against the corresponding simulated fitness values (Figure 5.17). Surprisingly, the inferred value of s was, on average, found to be greater than the simulated value by 2.8%. It is noteworthy that despite yielding acceptable results for the inference on the mean-fitness trajectory (Figure 5.11), the Bayesian filtering program employed in this study was unable to provide accurate estimations for s .

The main concern is whether the inaccurate inference of s is the primary cause of the observed failure. To assess this, it is important to evaluate the quality of the mutational rate estimation when s is accurate. For this purpose, all the inferred s values were manually adjusted downward by 2.8% (orange histogram, Figure 5.18). Then, the back-calculation of the shifted

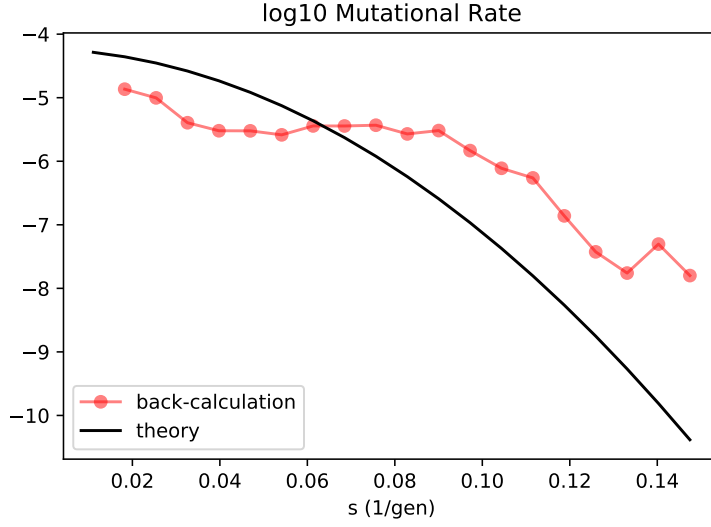


Figure 5.13. Mutation rate spectrum: theory (black line, $\mu(s) = U_b e^{-(s/\sigma)^\beta} / \sigma \Gamma(1 + \beta^{-1})$ with $U_b = 3 \times 10^{-4}$, $\beta = 2$, $\sigma = 3.8\%$) and the inference (red circle line) by back-calculating the observed DFE.

observed DFE was performed, yielding a new mutational spectrum. As shown in figure 5.14, the new spectrum (red circle) aligns with the theoretical values (black line) within the same order of magnitude. It should be noted, however, that this analysis does not provide a precise estimation of the mutation rate spectrum. Specifically, the estimated value for weak mutations tends to be underestimated, which may be attributed to the loss of weak mutants resulting from the left-shifting. Despite this imperfection, the inference of the mutation rate can be considered reliable if more accurate values for s are employed. Furthermore, the results indicate that both the work on detection probability and establishment probabilities, as well as the back-calculation procedure, can be deemed dependable.

5.4 Discussion

In this study, I present a comprehensive framework for inferring the underlying mutational rate spectrum for nascent beneficial mutations in adaptive evolution. The key idea is to back-calculate the observed distribution of fitness effect (DFE) by incorporating the detec-

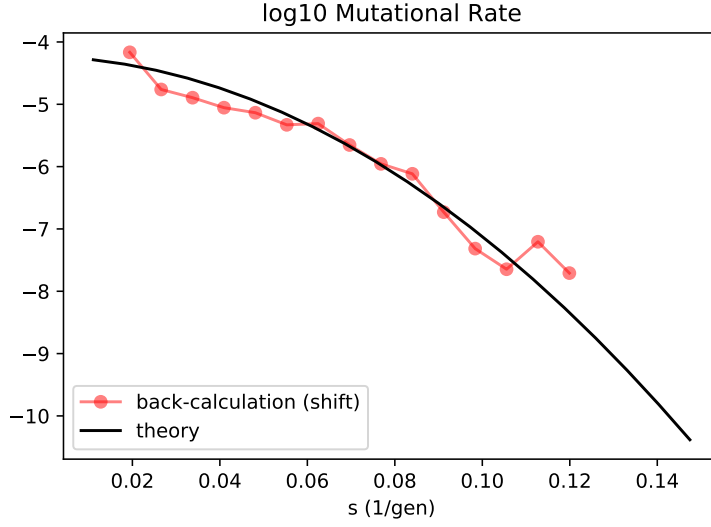


Figure 5.14. Mutation rate spectrum: theory (black line, $\mu(s) = U_b e^{-(s/\sigma)^\beta} / \sigma \Gamma(1 + \beta^{-1})$ with $U_b = 3 \times 10^{-4}$, $\beta = 2$, $\sigma = 3.8\%$) and the inference (red circle line) by back-calculating the shifted, observed DFE.

tion probability and establishment probability of adaptive mutants. The probability of mutant establishment is solvable using the Wright-Fisher model, and I provide an analytical solution for the establishment probability, which I validate through simulations of the Wright-Fisher model. Furthermore, I empirically compute the probability of detecting an established mutant. By integrating these findings, an inferred mutational rate spectrum is derived for a simulation study. Unfortunately, the inferred mutational spectrum is inaccurate, and the analysis suggests that the inaccurate inference of fitness is the primary cause.

A remaining challenge is the failure to accurately estimate fitness using the Bayesian filtering program. This failure may be attributed to the mismatch between the model used in this study (Wright-Fisher) and the model used in the Bayesian program (serial batch culture). The serial batch culture model assumes deterministic population growth during the growth phase, followed by dilution to the bottleneck size in each transfer. Genetic drift in this model is influenced by the dilution factor and the effective population size. I chose the serial batch model for my Bayesian program as it closely resembles the experimental process in laboratory

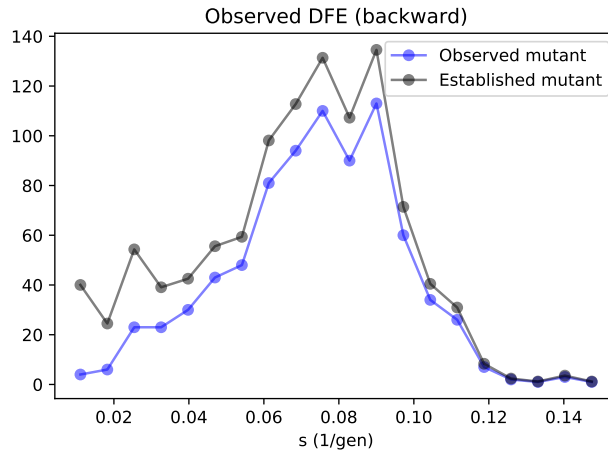


Figure 5.15. Observed DFE (blue circle) and its backward-calculating established DFE (black circle)

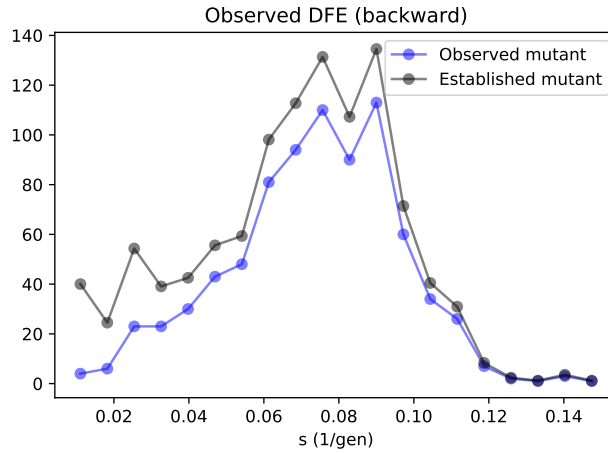


Figure 5.16. Established DFE in simulation (black square) and its forward-calculating observed DFE (blue square)

evolution. However, a drawback of the serial batch model is the lack of analytical solutions to describe the establishment probability and subsequent population dynamics. In contrast, the Wright-Fisher model offers solvability, but it may not accurately capture drift in microbial evolution experiments.

There are two potential approaches to address the model discrepancy. The first option is to utilize the Wright-Fisher model as the underlying framework in the Bayesian program, which is relatively straightforward to implement. However, a concern arises regarding whether the WF-

Bayesian method can accurately characterize the fitness of mutants in laboratory evolution. To evaluate its applicability, it may be worthwhile to apply the WF-Bayesian method to simulation data of the serial-batch culture. If the method yields satisfactory results for serial-batch culture simulations, the first approach would be advantageous.

The second approach involves developing an analytical description for the mutation population in the serial-batch model. If an approximate solution for the establishment probability can be derived, it would provide a significant advantage. Currently, I have obtained preliminary results for the long-term survival probability in the serial-batch culture, with values of $p = 3.4s$ for the serial-batch model with a dilution factor $D > 8$ (Wright-Fisher model: $p = 2s$, branching process: $p = s$). However, further work is required to investigate lineage dynamics and obtain the establishment probability, in order to determine the compatibility of stochasticity to the WF and branching process.

Lastly, this study leaves several important questions unanswered. Firstly, there is a lack of confidence or uncertainty estimation in the inference of the mutational rate spectrum. It is crucial to assess the reliability and uncertainty of the results obtained. Furthermore, the current analysis is limited to a specific mutational rate spectrum, namely the general exponential distribution, while other shapes of distributions have not been explored. In this study, the lower bound of fitness for mutant establishment was found to be $s = 1.2\%$, while the lower bound of fitness for detection was determined to be $s = 0.3\%$. However, it remains unclear what factors contribute to these observed limits of s . Exploring these factors would contribute to a more robust, accurate, and interpretable inference of the mutational rate spectrum.

Acknowledgments

I would also like to extend my heartfelt gratitude to my esteemed committee members, namely Sergey Kryazhimskiy, Terence Hwa, Elena Koslover, Vineet Bafna, and Wouter-Jan Rappel, for their invaluable support and guidance in propelling the progress of the work presented

in this chapter.

Chapter 5 is a work in progress that contains unpublished material coauthored with Sergey, Kryazhimskiy. The dissertation author was the primary author of this chapter.

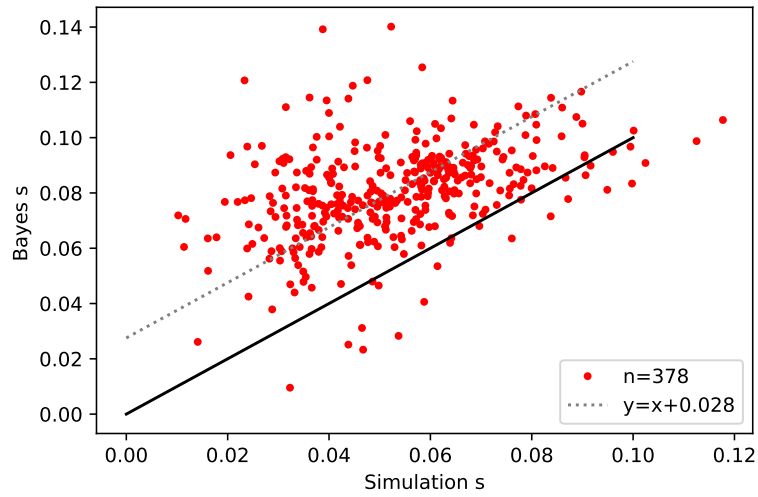


Figure 5.17. Inaccurate estimated fitness effect by Bayesian filtering program compared to the simulation value

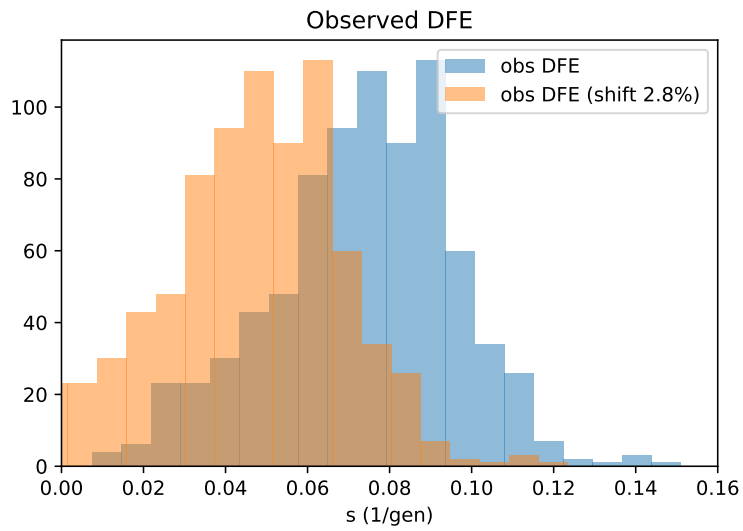


Figure 5.18. The observed DFE (blue histogram) and the observed DFE left-shifted by 2.8% (orange histogram).

Appendix A

Barcode Frequency Dynamics

In the Levy-Blundell method, the inference of mean fitness $\bar{s}_k$ at sampling time k is based on the dynamics of neutral barcodes ($s = 0$). This model assumes that the expected (arithmetic mean) of barcode read μ , calculated from lineages of the same barcode read count r_{k-1} at the sampling time $k - 1$, is given as

$$\frac{\mu}{R_k} = \frac{r_{k-1}}{R_{k-1}} e^{-\bar{s}_k \Delta t_k}, \quad (\text{A.1})$$

where Δt_k is time between two sampling events, R_k & R_{k-1} are the total read count at time point k & $k - 1$, and s is the selection coefficient of lineages.

To examine this assumption, I derive barcode dynamics from cell dynamics. We first denote $\mu = \mathbb{E}(r_k | r_{k-1})$ as the expected barcode read count for lineages selected on r_{k-1} at the previous cycle $k - 1$. The expected value can be broken down to the integration over read count r_k and cell number n_k , as shown in equation A.2.

$$\begin{aligned} \mathbb{E}(r_k | r_{k-1}) &= \int dr_k r_k P(r_k | r_{k-1}) \\ &= \int dn_k \int dr_k r_k P(r_k | n_k) P(n_k | r_{k-1}) \\ &= \int dn_k \mathbb{E}(r_k | n_k) P(n_k | r_{k-1}) \end{aligned} \quad (\text{A.2})$$

Since the expected observation of a lineage is linearly proportional to its size, it has $\mathbb{E}(r_k | n_k) = R_k n_k / N_k$, where R_k and N_k denotes total read depth and total population size. I thus take $\mathbb{E}(r_k | n_k)$

into equation A.2, simplify the equation, and write it as an integration of cell size n_{k-1} at the previous cycle $k-1$, as shown in equation A.3.

$$\begin{aligned}\mathbb{E}(r_k|r_{k-1}) &= \frac{R_k}{N_k} \int dn_k n_k P(n_k|r_{k-1}) \\ &= \frac{R_k}{N_k} \int dn_{k-1} \int dn_k n_k P(n_k|n_{k-1}) P(n_{k-1}|r_{k-1}) \\ &= \frac{R_k}{N_k} \int dn_{k-1} \int \mathbb{E}(n_k|n_{k-1}) P(n_{k-1}|r_{k-1})\end{aligned}\tag{A.3}$$

Based on the cell frequency dynamics (equation 1.21), the expected cell size is given as

$$\mathbb{E}(n_k|n_{k-1}) = N_k \frac{n_{k-1}}{N_{k-1}} e^{(s-\bar{s}_k)\Delta t_k}.\tag{A.4}$$

Taking equation A.4 into equation A.3, I have

$$\begin{aligned}\mathbb{E}(r_k|r_{k-1}) &= \frac{R_k}{N_k} \frac{N_k}{N_{k-1}} e^{(s-\bar{s}_k)\Delta t_k} \int dn_{k-1} n_{k-1} P(n_{k-1}|r_{k-1}) \\ &= \frac{R_k}{N_{k-1}} e^{(s-\bar{s}_k)\Delta t_k} \mathbb{E}(n_{k-1}|r_{k-1})\end{aligned}\tag{A.5}$$

Note that the expected size $\mathbb{E}(n_{k-1}|r_{k-1})$ depends on the prior distribution of population $P(n_{k-1})$ and the measurement process $P(r_{k-1}|n_{k-1})$ for barcode read.

$$\mathbb{E}(n_{k-1}|r_{k-1}) = \int dn_{k-1} n_{k-1} \frac{P(r_{k-1}|n_{k-1}) P(n_{k-1})}{P(r_{k-1})}.\tag{A.6}$$

To learn the property of $\mathbb{E}(n_{k-1}|r_{k-1})$, I consider a simplest case where the prior size $P(n_{k-1})$ is exponential distributed, and the measurement process $P(r_{k-1}|n_{k-1})$ is Poisson distributed with mean of $R_{k-1}n_{k-1}/N_{k-1}$. For the simplest case, I calculate and derive equation 2.7 as

$$\mathbb{E}(n_{k-1}|r_{k-1}) = \frac{N_{k-1}}{R_{k-1}} (1 + r_{k-1}),\tag{A.7}$$

thus the expected read is given as

$$\mathbb{E}(r_k|r_{k-1}) = \frac{R_k}{R_{k-1}} (1 + r_{k-1}) e^{(s-\bar{s}_k)\Delta t_k}. \quad (\text{A.8})$$

If the sampling process is noisy (assuming $P(r_{k-1}|n_{k-1})$ as a negative binomial (NB) distribution with a mean of $R_{k-1}n_{k-1}/N_{k-1}$ and variance of $2\kappa \times \text{mean}$), I derive the expected read as

$$\mathbb{E}(r_k|r_{k-1}) \approx \frac{R_k}{R_{k-1}} (2\kappa + r_{k-1}) e^{(s-\bar{s}_k)\Delta t_k}. \quad (\text{A.9})$$

The function has an offset (1 for Poisson and 2κ for NB distribution) determined by the measurement noise for two simple cases. I show that the expected value (arithmetic mean) of the read count is not linear. Thus neutral decline assumption A.1 is incorrect in general.

Appendix B

Prediction of mutant population dynamics in Wright-Fisher model

Here, I demonstrate that the solutions derived from the branching process theory can also be applied to the Wright-Fisher simulations. The Wright-Fisher simulation is as follows. I start with a single mutant cell (with fitness effect s) in a fixed population of size N . All parents cell deliver offspring stochastically in a competitive environment where mean fitness increases from 0% to 5% during 300 generations. At the end of each generation, offspring are randomly selected to the next generations. The population trajectory of mutation from generation 0 to 300 is record. The simulation is repeated for $\sim 100,000$ runs until ~ 4000 runs have survival mutant at the end. Similarly to the approaches in the branching process, the establishment time τ is draw from the population trajectory (Figure B.1A) and the survival probability is computed.

In the Wright-Fisher model, mutation's survival probability is twice of that in the branching process model ($P_{\text{surv}}^{\text{WF}} = 2P_{\text{surv}}^{\text{BR}}$), but half the expected population size ($\mathbb{E}^{\text{WF}}(n | \text{Survive}) = \frac{1}{2} \mathbb{E}^{\text{BR}}(n | \text{Survive})$). However, the dynamics of the branching process can be corrected accordingly and be used to successfully predict the mutant's behavior in the Wright-Fisher model (Figure B.1). Therefore, mutant dynamics in Wright-Fisher can be pictured as followings. A random beneficial mutant, starting from single cell, might go extinct or survive with survival probability $P_{\text{surv}}^{\text{WF}}$. For a mutant that survives, its population drifts stochastically before the establishment time τ , and grow deterministically in size $\mathbb{E}^{\text{WF}}(n | \text{Survive})$ afterwards. The estab-

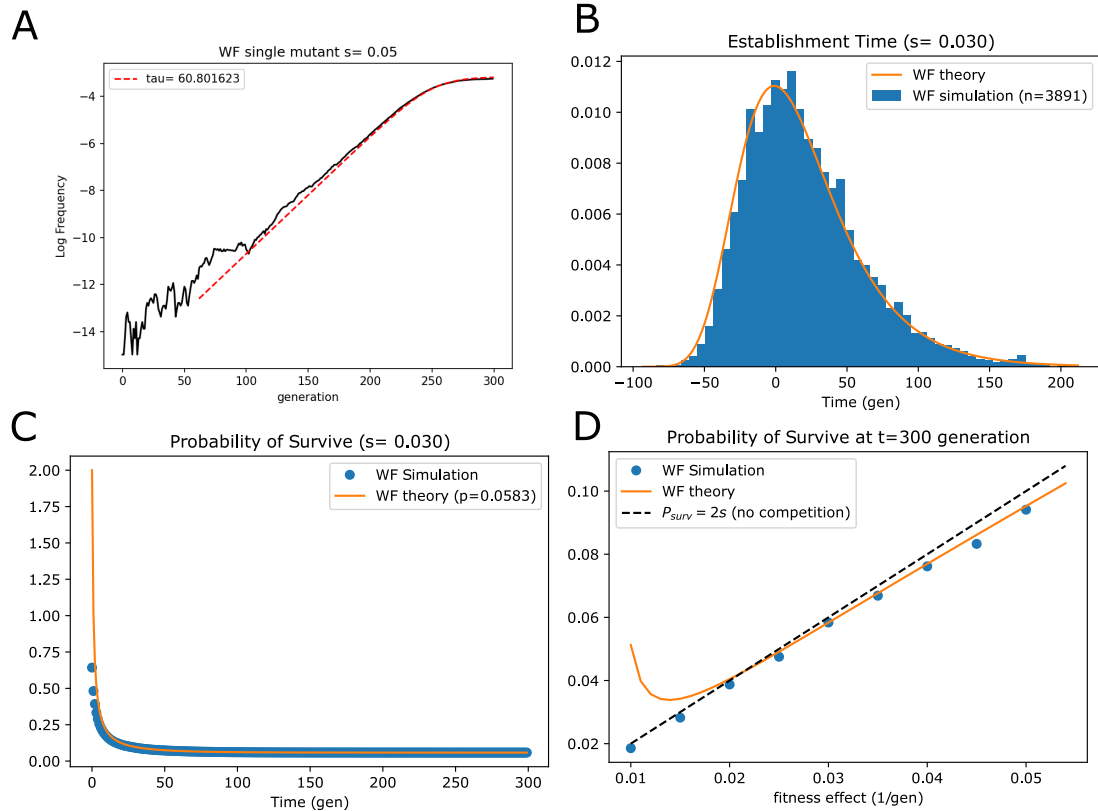


Figure B.1. Wright-Fisher simulation (A) trajectory of mutant population (B) distribution of establishment time (C) survival probability over time (D) survival probability at the end time to the fitness effect.

lishment time is a random variable, which is sampled from the same probability distribution as in branching process model.

Bibliography

- [1] Dimitra Aggeli, Yuping Li, and Gavin Sherlock. Changes in the distribution of fitness effects and adaptive mutational spectra following a single first step towards adaptation. *Nature Communications*, 12(1):5193, December 2021.
- [2] Joao A. Ascensao, Kelly M. Wetmore, Benjamin H. Good, Adam P. Arkin, and Oskar Hallatschek. Quantifying the Adaptive Potential of a Nascent Bacterial Community, February 2022.
- [3] Charles F Baer, Michael M Miyamoto, and Dee R Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, 8(8):619–631, 2007.
- [4] Jeffrey E Barrick and Richard E Lenski. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827–839, 2013.
- [5] João Barroso-Batista, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B Xavier, and Isabel Gordo. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS genetics*, 10(3):e1004182, 2014.
- [6] Michael Baym, Sergey Kryazhimskiy, Tami D Lieberman, Hattie Chung, Michael M Desai, and Roy Kishony. Inexpensive multiplexed library preparation for megabase-sized genomes. *PloS one*, 10(5):e0128036, 2015.
- [7] Jamie R. Blundell, Katja Schwartz, Danielle Francois, Daniel S. Fisher, Gavin Sherlock, and Sasha F. Levy. The dynamics of adaptive genetic diversity during the early stages of clonal evolution. *Nature Ecology & Evolution*, 3(2):293–301, February 2019.
- [8] Kirsten Bomblies and Catherine L Peichel. Genetics of adaptation. *Proceedings of the National Academy of Sciences*, 119(30):e2122152119, 2022.
- [9] Guoshuai Cai, Shoudan Liang, Xiaofeng Zheng, and Feifei Xiao. Local sequence and sequencing depth dependent accuracy of rna-seq reads. *BMC bioinformatics*, 18:1–12, 2017.

- [10] Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 117(9):4682–4692, 2020.
- [11] Deborah Charlesworth, Brian Charlesworth, and MT Morgan. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–1632, 1995.
- [12] Luis-Miguel Chevin. On measuring selection in experimental evolution. *Biology Letters*, 7(2):210–213, April 2011.
- [13] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034):1190–1192, 2011.
- [14] Nate J. Cira, Michael T. Pearce, and Stephen R. Quake. Neutral and selective dynamics in a synthetic microbial community. *Proceedings of the National Academy of Sciences*, 115(42):E9842–E9848, October 2018.
- [15] James F. Crow and Motoo Kimura. An Introduction to Population Genetics Theory. In *Population (French Edition)*, volume 26, page 977, September 1971.
- [16] Subhajyoti De and Franziska Michor. Dna replication timing and long-range dna interactions predict mutational landscapes of cancer genomes. *Nature biotechnology*, 29(12):1103–1108, 2011.
- [17] Michael M Desai and Daniel S Fisher. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798, 2007.
- [18] John W Drake. A constant rate of spontaneous mutation in dna-based microbes. *Proceedings of the National Academy of Sciences*, 88(16):7160–7164, 1991.
- [19] Santiago F. Elena and Richard E. Lenski. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, 4(6):457–469, June 2003.
- [20] Warren John Ewens. *Mathematical population genetics: theoretical introduction*, volume 27. Springer, 2004.
- [21] Adam Eyre-Walker. The genomic rate of adaptive evolution. *Trends in ecology & evolution*, 21(10):569–575, 2006.
- [22] Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, 2007.

- [23] Ronald A Fisher. Xxi.—on the dominance ratio. *Proceedings of the royal society of Edinburgh*, 42:321–341, 1923.
- [24] Ronald A Fisher. The genetical theory of natural selection. clarendon, 1930.
- [25] Simon E Fisher and Constance Scharff. FOXP2 as a molecular window into speech and language. *Trends in Genetics*, 25(4):166–177, 2009.
- [26] Patricia L Foster, Andrew J Hanson, Heewook Lee, Ellen M Popodi, and Haixu Tang. On the mutational topology of the bacterial genome. *G3: Genes, Genomes, Genetics*, 3(3):399–407, 2013.
- [27] Evgeni M Frenkel, Benjamin H Good, and Michael M Desai. The fates of mutant lineages and the distribution of fitness effects of beneficial mutations in laboratory budding yeast populations. *Genetics*, 196(4):1217–1226, 2014.
- [28] Douglas J. Futuyma. *Evolutionary Biology*. Sinauer Associates, Sunderland, Mass., 3rd ed edition, 1998.
- [29] Philip J Gerrish and Richard E Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144, 1998.
- [30] John H Gillespie. Molecular evolution over the mutational landscape. *Evolution*, pages 1116–1129, 1984.
- [31] Kevin Gomez, Jason Bertram, and Joanna Masel. Mutation bias can shape adaptation in large asexual populations experiencing clonal interference. *Proceedings of the Royal Society B*, 287(1937):20201503, 2020.
- [32] Benjamin H Good and Michael M Desai. The impact of macroscopic epistasis on long-term evolutionary dynamics. *Genetics*, 199(1):177–190, 2015.
- [33] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, 2017.
- [34] Benjamin H Good, Igor M Rouzine, Daniel J Balick, Oskar Hallatschek, and Michael M Desai. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proceedings of the National Academy of Sciences*, 109(13):4950–4955, 2012.
- [35] David J Gordon, Benjamin Resio, and David Pellman. Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics*, 13(3):189–203, 2012.
- [36] John Burdon Sanderson Haldane. A mathematical theory of natural and artificial selec-

tion, part v: selection and mutation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23, pages 838–844. Cambridge University Press, 1927.

- [37] Arbel Harpak, Anand Bhaskar, and Jonathan K Pritchard. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS genetics*, 12(12):e1006489, 2016.
- [38] Matthew Hegreness, Noam Shores, Daniel Hartl, and Roy Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617, 2006.
- [39] Ryan T Hietpas, Claudia Bank, Jeffrey D Jensen, and Daniel NA Bolon. Shifting fitness landscapes in response to altered environments. *Evolution*, 67(12):3512–3522, 2013.
- [40] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756–766, 2011.
- [41] Christopher JR Illingworth and Ville Mustonen. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLoS pathogens*, 8(12):e1003091, 2012.
- [42] Weronika Jasinska, Michael Manhart, Jesse Lerner, Louis Gauthier, Adrian W. R. Serojijos, and Shimon Bershtein. Chromosomal barcoding of *E. coli* populations reveals lineage diversity dynamics at high resolution. *Nature Ecology & Evolution*, 4(3):437–452, March 2020.
- [43] Milo S. Johnson, Alena Martsul, Sergey Kryazhimskiy, and Michael M. Desai. Higher-fitness yeast genotypes are less robust to deleterious mutations. *Science*, 366(6464):490–493, October 2019.
- [44] Milo S. Johnson, Sandeep Venkataram, and Sergey Kryazhimskiy. Best practices in designing, sequencing, and identifying random DNA barcodes. *J Mol Evol*, 2023.
- [45] Paul Joyce, Darin R Rokya, Craig J Beisel, and H Allen Orr. A general extreme value theory model for the adaptation of dna sequences under strong selection and weak mutation. *Genetics*, 180(3):1627–1643, 2008.
- [46] Katy C Kao and Gavin Sherlock. Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nature genetics*, 40(12):1499–1504, 2008.
- [47] Rees Kassen and Thomas Bataillon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature genetics*, 38(4):484–488, 2006.

- [48] Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196, 2011.
- [49] Motoo Kimura. On the evolutionary adjustment of spontaneous mutation rates*. 9(1):23–34. Publisher: Cambridge University Press.
- [50] Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.
- [51] Alexander Klug and Joachim Krug. Conflicting effects of recombination on the evolvability and robustness in neutrally evolving populations. *PLOS Computational Biology*, 18(11):e1010710, 2022.
- [52] Fyodor A Kondrashov and Alexey S Kondrashov. Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1169–1176, 2010.
- [53] Sergey Kryazhimskiy, Daniel P Rice, Elizabeth R Jerison, and Michael M Desai. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522, 2014.
- [54] Sergey Kryazhimskiy, Gašper Tkačik, and Joshua B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences*, 106(44):18638–18643, 2009.
- [55] Gregory I Lang, David Botstein, and Michael M Desai. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(3):647–661, 2011.
- [56] Gregory I Lang and Andrew W Murray. Mutation rates across budding yeast chromosome vi are correlated with replication timing. *Genome biology and evolution*, 3:799–811, 2011.
- [57] Gregory I Lang, Daniel P Rice, Mark J Hickman, Erica Sodergren, George M Weinstock, David Botstein, and Michael M Desai. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.
- [58] Sasha F. Levy, Jamie R. Blundell, Sandeep Venkataram, Dmitri A. Petrov, Daniel S. Fisher, and Gavin Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519(7542):181–186, March 2015.
- [59] Fangfei Li, Marc L Salit, and Sasha F Levy. Unbiased fitness estimation of pooled barcode or amplicon sequencing studies. *Cell systems*, 7(5):521–525, 2018.
- [60] Yuping Li, Dmitri A. Petrov, and Gavin Sherlock. Single nucleotide mapping of trait space

- reveals Pareto fronts that constrain adaptation. *Nature Ecology & Evolution*, 3(11):1539–1551, November 2019.
- [61] Yuping Li, Dmitri A Petrov, and Gavin Sherlock. Single nucleotide mapping of trait space reveals pareto fronts that constrain adaptation. *Nature ecology & evolution*, 3(11):1539–1551, 2019.
 - [62] Yuping Li, Sandeep Venkataram, Atish Agarwala, Barbara Dunn, Dmitri A. Petrov, Gavin Sherlock, and Daniel S. Fisher. Hidden Complexity of Yeast Adaptation under Simple Evolutionary Conditions. *Current biology: CB*, 28(4):515–525.e6, February 2018.
 - [63] Yuping Li, Sandeep Venkataram, Atish Agarwala, Barbara Dunn, Dmitri A Petrov, Gavin Sherlock, and Daniel S Fisher. Hidden complexity of yeast adaptation under simple evolutionary conditions. *Current Biology*, 28(4):515–525, 2018.
 - [64] Laurence Loewe. Quantifying the genomic decay paradox due to muller’s ratchet in human mitochondrial dna. *Genetics Research*, 87(2):133–159, 2006.
 - [65] Michael Lynch, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster. Genetic drift, selection and the evolution of the mutation rate. 17(11):704–714. Number: 11 Publisher: Nature Publishing Group.
 - [66] Iñigo Martincorena, Aswin SN Seshasayee, and Nicholas M Luscombe. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, 485(7396):95–98, 2012.
 - [67] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531, 2014.
 - [68] Lauren MF Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer as an evolutionary and ecological process. *Nature reviews cancer*, 6(12):924–935, 2006.
 - [69] Rosario Miralles, Philip J Gerrish, Andrés Moya, and Santiago F Elena. Clonal interference and the evolution of RNA viruses. *Science*, 285(5434):1745–1747, 1999.
 - [70] Ellie Mok, Eugene Wee, Yuling Wang, and Matt Trau. Comprehensive evaluation of molecular enhancers of the isothermal exponential amplification reaction. *Scientific reports*, 6(1):37837, 2016.
 - [71] J Grey Monroe, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, Julia Hildebrandt, Manuela Neumann, Daniel Kliebenstein, et al. Mutation bias reflects natural selection in arabidopsis thaliana. *Nature*, 602(7895):101–105, 2022.

- [72] Rob W Ness, Andrew D Morgan, Radhakrishnan B Vasanthakrishnan, Nick Colegrave, and Peter D Keightley. Extensive de novo mutation rate variation between individuals and across the genome of *chlamydomonas reinhardtii*. *Genome Research*, 25(11):1739–1749, 2015.
- [73] Alex N Nguyen Ba, Ivana Cvijović, José I Rojas Echenique, Katherine R Lawrence, Artur Rego-Costa, Xianan Liu, Sasha F Levy, and Michael M Desai. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*, 575(7783):494–499, 2019.
- [74] Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1):263–286, 1992.
- [75] Virginie Orgogozo. Replaying the tape of life in the twenty-first century. *Interface focus*, 5(6):20150057, 2015.
- [76] H Allen Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526, 2003.
- [77] H. Allen Orr. Fitness and its role in evolutionary genetics. *Nature reviews. Genetics*, 10(8):531–539, August 2009.
- [78] H Allen Orr. The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1195–1201, 2010.
- [79] H Allen Orr and Jerry A Coyne. The genetics of adaptation: a reassessment. *The American Naturalist*, 140(5):725–742, 1992.
- [80] Joshua L Payne, Fabrizio Menardo, Andrej Trauner, Sonia Borrell, Sebastian M Gygli, Chloe Loiseau, Sebastien Gagneux, and Alex R Hall. Transition bias influences the evolution of antibiotic resistance in mycobacterium tuberculosis. *PLoS biology*, 17(5):e3000265, 2019.
- [81] Joel R Peck, Guillaume Barreau, and Simon C Heath. Imperfect genes, fisherian mutation and the evolution of sex. *Genetics*, 145(4):1171–1199, 1997.
- [82] Pleuni S Pennings, Sergey Kryazhimskiy, and John Wakeley. Loss and recovery of genetic diversity in adapting populations of hiv. *PLoS genetics*, 10(1):e1004000, 2014.
- [83] Gwenaél Piganeau and Adam Eyre-Walker. Estimating the distribution of fitness effects from dna sequence data: implications for the molecular clock. *Proceedings of the National Academy of Sciences*, 100(18):10335–10340, 2003.
- [84] Raul Rabadan, Gyan Bhanot, Sonia Marsilio, Nicholas Chiorazzi, Laura Pasqualucci, and

- Hossein Khiabani. On statistical modeling of sequencing noise in high depth data to assess tumor evolution. *Journal of statistical physics*, 172:143–155, 2018.
- [85] Darin R Rokyta, Craig J Beisel, Paul Joyce, Martin T Ferris, Christina L Burch, and Holly A Wichman. Beneficial fitness effects are not exponential for two viruses. *Journal of molecular evolution*, 67:368–376, 2008.
- [86] W. A. Rosche and P. L. Foster. Determining mutation rates in bacterial populations. 20(1):4–17.
- [87] JM Ruijter, C Ramakers, WMH Hoogaars, Y Karlen, O Bakker, MJB Van den Hoff, and AFM2665230 Moorman. Amplification efficiency: linking baseline and bias in the analysis of quantitative pcr data. *Nucleic acids research*, 37(6):e45–e45, 2009.
- [88] Rafael Sanjuán, Andrés Moya, and Santiago F Elena. The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proceedings of the National Academy of Sciences*, 101(22):8396–8401, 2004.
- [89] Martijn F Schenk, Mark P Zwart, Sungmin Hwang, Philip Ruelens, Edouard Severing, Joachim Krug, and JAGM De Visser. Population size mediates the contribution of high-rate and large-benefit mutations to parallel evolution. *Nature Ecology & Evolution*, 6(4):439–447, 2022.
- [90] Stephan Schiffels, Gergely J Szöllösi, Ville Mustonen, and Michael Lässig. Emergent neutrality in adaptive asexual evolution. *Genetics*, 189(4):1361–1375, 2011.
- [91] Ulrich Schlecht, Zhimin Liu, Jamie R Blundell, Robert P St. Onge, and Sasha F Levy. A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nature communications*, 8(1):15586, 2017.
- [92] Paul D Sniegowski, Philip J Gerrish, Toby Johnson, and Aaron Shaver. The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066, 2000.
- [93] John A Stamatoyannopoulos, Ivan Adzhubei, Robert E Thurman, Gregory V Kryukov, Sergei M Mirkin, and Shamil R Sunyaev. Human mutation rate associated with dna replication timing. *Nature genetics*, 41(4):393–395, 2009.
- [94] David L Stern. The genetic causes of convergent evolution. *Nature Reviews Genetics*, 14(11):751–764, 2013.
- [95] Jay F Storz, Chandrasekhar Natarajan, Anthony V Signore, Christopher C Witt, David M McCandlish, and Arlin Stoltzfus. The role of mutation bias in adaptive molecular evolution: insights from convergent changes in protein function. *Philosophical Transactions of the*

Royal Society B, 374(1777):20180238, 2019.

- [96] Natalja Strelkova and Michael Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682, 2012.
- [97] Way Sung, Matthew S Ackerman, Jean-François Gout, Samuel F Miller, Emily Williams, Patricia L Foster, and Michael Lynch. Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. *Molecular biology and evolution*, 32(7):1672–1683, 2015.
- [98] Abigail R. Tirrell, Katelyn M. Vendrely, Lisa A. Checkley, Sage Z. Davis, Marina McDew-White, Ian H. Cheeseman, Ashley M. Vaughan, François H. Nosten, Timothy J. C. Anderson, and Michael T. Ferdig. Pairwise growth competitions identify relative fitness relationships among artemisinin resistant *Plasmodium falciparum* field isolates. *Malaria Journal*, 18(1):295, August 2019.
- [99] Tat Dat Tran, Julian Hofrichter, and Jürgen Jost. An introduction to the mathematical structure of the wright–fisher model of population genetics. *Theory in Biosciences*, 132(2):73–82, 2013.
- [100] Kimberly S. Vasquez, Lisa Willis, Nate J. Cira, Katharine M. Ng, Miguel F. Pedro, Andrés Aranda-Díaz, Manohary Rajendram, Feiqiao Brian Yu, Steven K. Higginbottom, Norma Neff, Gavin Sherlock, Karina B. Xavier, Stephen R. Quake, Justin L. Sonnenburg, Benjamin H. Good, and Kerwyn Casey Huang. Quantifying rapid bacterial evolution and transmission within the mouse intestine. *Cell Host & Microbe*, 29(9):1454–1468.e4, September 2021.
- [101] Sandeep Venkataram, Barbara Dunn, Yuping Li, Atish Agarwala, Jessica Chang, Emily R. Ebel, Kerry Geiler-Samerotte, Lucas Hérisant, Jamie R. Blundell, Sasha F. Levy, Daniel S. Fisher, Gavin Sherlock, and Dmitri A. Petrov. Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell*, 166(6):1585–1596.e22, September 2016.
- [102] Sandeep Venkataram, Huan-Yu Kuo, Erik FY Hom, and Sergey Kryazhimskiy. Mutualism-enhancing mutations dominate early adaptation in a two-species microbial community. *Nature Ecology & Evolution*, pages 1–12, 2023.
- [103] Sandeep Venkataram, Ross Monasky, Shohreh H Sikaroodi, Sergey Kryazhimskiy, and Betül Kaçar. Evolutionary stalling and a limit on the power of natural selection to improve a cellular module. *Proceedings of the National Academy of Sciences*, 117(31):18582–18590, 2020.
- [104] Liyun Wang, Daming Fan, Wei Chen, and Eugene M. Terentjev. Bacterial growth, detachment and cell size control on polyethylene terephthalate surfaces. *Scientific Reports*,

5(1):15159, October 2015.

- [105] Benjamin A Wilson, Nandita R Garud, Alison F Feder, Zoe J Assaf, and Pleuni S Pennings. The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. *Molecular ecology*, 25(1):42–66, 2016.
- [106] Michael J Wiser, Noah Ribeck, and Richard E Lenski. Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367, 2013.
- [107] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.
- [108] Bin Wu, Chaitanya S Gokhale, Matthijs van Veelen, Long Wang, and Arne Traulsen. Interpretations arising from Wrightian and Malthusian fitness under strong frequency dependent selection. *Ecology and Evolution*, 3(5):1276–1280, May 2013.
- [109] Yuan O Zhu, Mark L Siegal, David W Hall, and Dmitri A Petrov. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, 111(22):E2310–E2318, 2014.