UNIVERSITY OF CALIFORNIA
RIVERSIDE

Inferring Fitness From Genetic Time Series Data

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Physics

by

Zhenchen Hong

September 2023

Dissertation Committee:

Dr. John Paul Barton, Chairperson
Dr. Thomas Kuhlman
Dr. Roya Zandi

The Dissertation of Zhenchen Hong is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my deepest gratitude and appreciation to the many individuals and institutions who have contributed to the realization of this doctoral journey. First and foremost, I am immensely thankful to my advisor, Dr. John Paul Barton, whose guidance, unwavering support, and intellectual insights have been the cornerstone of this research endeavor. Your dedication to pushing the boundaries of knowledge and your commitment to fostering my growth as a researcher have been truly inspiring.

I extend my heartfelt thanks to the members of my dissertation committee, Dr. Thomas Kuhlman and Dr. Roya Zandi, for their invaluable feedback, constructive criticism, and thought-provoking discussions that have undoubtedly enriched the quality of this work. The diversity of perspectives and expertise each of you brought to the table has been instrumental in shaping the depth and breadth of my research.

I am grateful for the camaraderie and intellectual exchange with my fellow colleagues and peers at University of California, Riverside and Hong Kong University of Science and Technology. The stimulating debates, exchange of ideas, and late-night discussions have played an integral role in shaping my academic journey. Special appreciation goes to my lab mates, Dr. Kai Shimagaki, Dr. Yawei Qin, Dr. Yunxiao Li, Dr. Marco Garcia, Brian Lee, Liz Finney, Yirui Gao, Kevin Yang, Dr. Edwin Rodríguez, and amazing collaborators, Dr. Muhammad Saqib Sohail, Dr. Raymond HY Louie and Dr. Matthew R McKay, for their continuous support and friendship.

I extend my thanks to University of California, Riverside and the National Institute of General Medical Sciences of the National Institutes of Health for their financial support, enabling me to carry out this research effectively during the last six years. Their investment in my work has been pivotal in bringing this thesis to fruition.

My gratitude also goes to the numerous experts, scholars, and professionals who graciously shared their insights during interviews and surveys, providing me with the empirical foundation upon which this research is built.

Last but not least, I want to express my deepest gratitude to my parents, Shiyu Hong and Liqing Lei for their unwavering encouragement, and understanding throughout this journey. Thank you to Han Yang, for all her love and support. Your belief in me has been my driving force, and I am truly blessed to have you all by my side.

In sum, this thesis is a culmination of the collective efforts of many individuals and institutions, and it is with the utmost humility that I acknowledge and appreciate their contributions. While I take responsibility for any shortcomings that might persist, their support and guidance have undeniably propelled this work to new heights. Thank you all for being an integral part of this academic odyssey.

To my love for all the support.

ABSTRACT OF THE DISSERTATION

Inferring Fitness From Genetic Time Series Data

by

Zhenchen Hong

Doctor of Philosophy, Graduate Program in Physics
University of California, Riverside, September 2023
Dr. John Paul Barton, Chairperson

Evolution is the process by which populations of organisms undergo genetic changes over successive generations to fit to the environment. The genome of an organism contains its complete set of genetic instructions, including the information necessary for its development, functioning, and response to the environment. Thus, understanding the genetic variants on genomes responsible for adaptation during evolution is crucial, especially for comprehending the dynamics of fast-evolving pathogens or cancers. For example, the quick evolution of high risk pathogens, such as HIV-1 and influenza, is more likely to undergo the accumulation of advantageous mutations and enable them to evade the human immune system's defenses. In evolutionary biology, fitness refers to the measure of an organism's reproductive success adapting to the environment and its ability to contribute its genetic material to future generations. However, due to thousands to billions of base pairs on genomes, and the specific arrangements, estimating the fitness of genetic variants is a challenging task. Moreover, epistatic interactions, the effects of a genetic variant that depend on the presence of the other variants in one genetic sequence, elevate the level of challenge.

Although researchers are using advanced quantitative methods to decode these interactions, challenges still exist because of the increasing dimensions of the fitness landscape and difficulties in interpreting quantitative measurements.

To quantify the mutational effects of genetic variants, this work presents a method, Marginal Path Likelihood (MPL), inferring fitness parameters from observed evolutionary histories of genetic sequences. By extending the inference framework with epistatic interactions, this approach quantitatively measures the probability of an evolutionary path using a path integral derived from statistical physics, and estimates the fitness parameters, including the relative fitness (selection coefficient) and fitness that differs from the sum of the fitness effects of each individual mutant (epistasis), that best explain an observed evolutionary trajectory with Bayesian theorem. With the help of evolutionary simulation and mutagenesis experiments, this approach proves to be more consistent and explanatory than the current state-of-art methods, even within finite-sampling scenarios. In mutagenesis experiments, a large scale of genetic variants are generated and helps us to explore the functional consequences of numerous genetic variants simultaneously. Then, In this work, a pipeline package, popDMS, is also reported to process this kind of genetic time series data automatically.

Contents

List of Figures **xi**

List of Tables **xix**

1 Introduction **1**

1.1 Motivation 1

1.2 Quantifying mutational effects 2

1.3 Relationship with statistical physics 5

2 Inferring epistasis from genetic time-series data **8**

2.1 Inference framework overview 10

2.2 Design of the inference framework 12

2.2.1 Evolutionary model 12

2.2.2 Path integral: 15

2.2.3 Genotype-level path integral 16

2.2.4 Allele-level path integral 18

2.2.5 Marginal path likelihood (MPL) estimator with epistasis 34

2.2.6 Combining multiple independent observations 37

2.2.7 Asymmetrical mutation probabilities 38

2.2.8 Equivalence with the genotype estimate 44

2.2.9 Robustness 45

2.2.10 Simulation setup 47

2.2.11 Data Availability 47

2.3 Results 48

2.3.1 Accurate estimation of epistasis and selection coefficients 48

2.3.2 Analysis of a more complex five-locus epistatic fitness landscape . . 53

2.3.3 Robustness to sampling parameters 57

2.3.4 Comparison with a model that does not account for epistasis 62

2.3.5 Computational complexity 63

2.4 Discussion 65

3 Inferring mutational effects with deep mutational scanning data 68

3.1 Traditional mutagenesis experiments 68

3.1.1 Alanine mutagenesis experiments 68

3.1.2 Shortcomings of Alanine mutagenesis experiments 71

3.2 Deep mutational scanning (DMS) experiments 72

3.2.1 Introduction of DMS experiments 72

3.2.2 Advantages of DMS experiments and comparison with traditional mutagenesis techniques 74

3.3 Related data sets resources 77

3.3.1 Multiplex assays of variant effect database (MaveDB) 77

3.3.2 Bloom's lab DMS data 78

3.4 State-of-the-art analytical methods inferring mutational effects with DMS data 79

3.4.1 *dms_tools* and *dms_tools*2, ratio method 81

3.4.2 *Enrich* and *Enrich*2, ratio-regression method 83

3.4.3 Drawbacks of popular analytical tools inferring mutational effects . . 85

3.5 popDMS infers mutation effects with DMS experiments data 87

3.5.1 Inferring mutation effects by popDMS 89

3.5.2 Sequencing error correction 96

3.5.3 Regularization optimization 98

3.5.4 Simulation shows the ratio methods are sensitive to sampling noise. 99

3.5.5 Inferences by popDMS are more consistent than preferences across different replicates in real DMS experiments. 103

3.5.6 popDMS with multiple replicates 106

3.5.7 Main differences between popDMS and existed inference methods in single variant mutational effects inference. 107

3.5.8 Inference of epistatic interactions with popDMS and consensus gauge 110

3.5.9 Comparison of epistatic interaction between popDMS and functional score methods. 111

3.6 Discussion 113

4 Conclusions 116

Bibliography 121

List of Figures

2.1 MPL can accurately infer selection coefficients and pairwise epistasis terms. (A) shows distribution of inferred selection coefficients and pairwise epistasis terms for various forms of epistasis when both selection coefficients are positive, while (B) shows the same for the case when both selection coefficients are negative. The results were obtained for a two-locus system with selection coefficients s_1, s_2 corresponding to the mutant alleles at locus 1 and 2 respectively, pairwise epistasis term s_{12}, per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$, and the initial population consisted of only the WT genotype (00). The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. The dashed lines represent the true selection coefficients (s_1 and s_2) and epistasis term (s_{12}). In these simulations, $s_1 = s_2$, hence the histograms of the estimates of the two have a significant overlap shown in grey color. 49

2.2 Higher data variability leads to more accurate inference. (A) shows a sample run (*left* panel) of a two-locus system (negative epistasis scenario) where all genotypes are well represented in the data, as indicated by the magnitude of the diagonal entries of the integrated covariance matrix (*center* panel). This leads to accurate estimation of the epistasis term and the selection coefficients (*right* panel). The vertical bars in the *right* panel indicate the 95% confidence intervals while the horizontal bars indicate the true selection coefficients and epistasis terms. (B) shows a sample run (*left* panel) of a two-locus system (positive epistasis scenario) where the double mutant genotype has limited variation, as indicated by the magnitude of the bottom right entry of the integrated covariance matrix (*center* panel). This leads to low accuracy in the estimate of the epistasis term. The selection coefficient estimates are still accurate because the single mutant genotypes, although present at low frequencies, are well represented in the data as indicated by the first two entries of the diagonal of the integrated covariance matrix (*center* panel). The results were obtained for a two-locus system with selection coefficients s_1, s_2 corresponding to the mutant alleles at locus 1 and 2 respectively, pairwise epistasis term s_{12}, per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$ and the initial population consisted of only the WT genotype. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. . . . 50

2.3 MPL can accurately estimate individual fitness parameters (selection coefficients and epistasis terms) and/or their sums depending on the variation present in the population. The results are for a two-locus system with positive sign epistasis (selection coefficients s_1, s_2 and pairwise epistasis term s_{12}). The boxplots of inferred selection coefficients and epistasis terms are shown on white background in each panel, while those of their sums are shown on grey background. The red bars indicate the true values of the respective terms. The boxplots show the standard data summary (minimum, first quartile, median, third quartile, maximum). In order to control data variability, both the per locus mutation probability and the per locus recombination probability were set to zero. The initial population contained the genotypes indicated above each panel, and the frequency of each non-WT genotype in the initial population was set to 10% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 150$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. 51

2.4 MPL can accurately estimate individual selection coefficients and pairwise epistasis terms and/or their sums depending on the variation present in the population. Results are for a two-locus system with (A) negative sign epistasis, (B) positive sign epistasis, (C) negative epistasis, (D) positive epistasis. The boxplots of inferred selection coefficients and epistasis terms are shown on white background in each panel, while those of their sums are shown on grey background. The red lines indicate the true values of the respective terms. The boxplots show the standard data summary (minimum, first quartile, median, third quartile, maximum). In order to control data variability, both the per locus mutation probability and the per locus recombination probability were set to zero. The initial population contained the genotypes indicated above each panel, and the frequency of each non-WT genotype in the initial population was set to 10% of the population size. The sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, with $T = 150$ generation used for inference. 54

2.5 The fraction of selection coefficients and epistasis terms that are accessible depends on data variability. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by varying the number of unique genotypes in the initial population to either 5, 10 or 20). (C) shows the fraction of accessible selection coefficients as a function of data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. 56

2.6 The average fraction of accessible selection coefficients and pairwise epistasis terms increases with increasing data variability (controlled by changing the number of unique genotypes in initial population to either five, ten or twenty). The selection coefficients and pairwise epistasis terms of a five-locus system were classified into three categories based on the reduced row-Echelon form (see 2.2) of the integrated covariance matrix in (2.65). Results are for a five-locus system (i.e., five selection coefficients and ten pairwise epistasis terms) simulated in Figure 2.8. 57

2.7 The fraction of selection coefficients and epistasis terms accessible in low data variability scenarios can be increased by combining multiple independent replicates. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by using either 1, 3 or 5 replicates for inference). (C) shows the fraction of accessible selection coefficients increases with the increase in data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. The initial population contained five unique genotypes. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. 58

2.8 The fraction of selection coefficients and epistasis terms that are accessible depends on data variability. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by varying the number of unique genotypes in the initial population to either 5, 10 or 20). (C) shows the fraction of accessible selection coefficients as a function of data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. 59

2.9 MPL performs well on dense as well as sparse fitness landscapes. (A) shows true (model) fitness landscapes with varying density of pairwise epistasis terms, while (B) shows the corresponding mean AUROC of detecting accessible beneficial and deleterious selection coefficients. Results are for a five-locus system where the initial population contained twenty unique genotypes, with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. The selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). Error bars indicate the standard error of the mean. 60

2.10 MPL is robust to variation in sampling parameters. The *left* and *right* panels show the mean AUROC performance of detecting accessible beneficial and deleterious selection coefficients, respectively. Results are for a five-locus system with the fitness landscape shown in Figure 2.8A. The initial population contained twenty unique genotypes with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$ and $T = 100$ generations were used for inference. All results were averaged over 1000 Monte Carlo runs. 61

2.11 MPL is robust to variation in sampling parameters. The *left* and *right* panels show the mean AUROC performance of detecting beneficial and deleterious accessible epistasis terms respectively. Results are for a five-locus system with the fitness landscape shown in Figure 2.8A. The initial population contained twenty unique genotypes with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. 62

2.12 Ability of MPL to accurately identify selection coefficients is robust to the density of non-zero epistasis terms in the fitness landscape. Comparison of MPL with MPL (no epistasis) method of [155] shows that both methods have similar performance when the underlying fitness landscape has few epistatic interactions. As the density of non-zero pair-wise epistasis terms increases in the underlying fitness landscape, there is a continuous degradation in the performance of MPL (no epistasis) while performance of MPL remains robust. The *top* and *bottom* panels show the mean AUROC performance of detecting beneficial and deleterious selection coefficients from the rest respectively. Error bars indicate the standard error of the mean. Results are for a five-locus system where the initial population contained 20 unique genotypes, with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. The sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, with $T = 100$ generation used for inference. All simulation results were computed over 1000 Monte Carlo runs. 64

2.13 MPL outperforms the MPL (without epistasis) method in classification of beneficial and deleterious accessible selection coefficients even on data with low variability. The figure shows the mean AUROC performance of the two methods. Error bars indicate the standard error of the mean. Results are for a five-locus system with a fully connected fitness landscape. The initial population contained five unique genotypes, with both the per locus mutation probability μ and per locus recombination probability r set to zero. 65

3.1 Main steps in Deep mutational scanning experiments. 73

3.2 popDMS pipeline workflow. 88

3.3 **Regularization optimization sample of TpoR data set.** popDMS will run the inference with different regularization magnitudes, depending on the least frequency of each data set. With each regularization, the pipeline will calculate the average Pearson correlation of mutational effects inferred by popDMS across replicates. Then the pipeline will choose the least regularization having the correlation drawdown less than 10% of overall correlation range. The red circle is the optimized regularization magnitude for further analysis. 99

3.4 **MPL is a more consistent method than the other mutational effects inference methods in the population genetic simulation with finite sampling.** **a**, Allele frequency trajectories of one beneficial mutant in the simulation. The red curve is the true evolution trajectory of this mutant with large population size, while the grey curves are the trajectories of finite sampling of this mutant. The stochastic trajectories of finite sampling have high variance when apply the inference methods of mutational effects. **b**, Performance comparison of different mutational effects inference methods with finite sampling. As the generations used for inference increased, the Pearson correlation coefficients increase for all the methods, and MPL always has the higher Pearson correlation coefficients than the other inference methods. . 102

3.5 **popDMS overview.** **a**, Mutational effects heatmap example of protein T7-longE4BU. Blue, white, and red boxes represent genetic variants that were deleterious, neutral, or beneficial, respectively, during the selection process; grey represents not observed variants; and dotted entries represent the wild-type residue.**b**, Inference consistency comparison of total 25 different experiments in 15 research data sets of human and virus proteins. popDMS infers more consistent mutational effects than the ratio/regression methods within all DMS experiments. **c**, Scatter plots of HIV envelop BF520 mutational effects inferred by popDMS across three experimental replicates. The correlation coefficients are high and don't come from some outliers misleading the correlations. **d**, Scatter plots of HIV envelop BF520 mutational effects inferred by dms_tool2 package across three experimental replicates. 105

3.6 **Visualization of single variant mutational effects and comparison of single variants mutational effects inference. a,** Logo plot of mutational effects in HIV Env BG505 example. **b,** Comparison of HIV Env BG505 mutational effects inferred by popDMS and preference method. Example site is 596, left logo plot is by popDMS, and right logo plot is by preference. **c,** Example frequency trajectories of top selection coefficients inferred by popDMS at site 596. $\triangle frequency$ is the frequency change between the post-selection stage and pre-selection stage, f_0 is the initial frequency of each variant. 108

3.7 **Heatmap of epistatic strength across methods for WW domain data set.** Upper right is the absolute epistatic strength across sites inferred by ratio method, while the lower left is inferred bu popDMS. The similar pattern of strong epistatic interaction is observed in this heatmap by two methods listed. The colorbar indicates how strong the asbolute epistatic strength is. 111

3.8 **Comparison of epistatic interaction inferred by popDMS vs. ratio method.** **a,** Scatter plot of element-wise epistasis inference by popDMS and ratio method. **b,** Histogram of pairs distance. Overall histogram is the distance distribution of all possible site pairs. The distances of popDMS and ratio method are the site distance of the site pairs with top 50 epistasis inference by different methods. The distance unit is angstrom (Å). t-statistics and p-value suggest whether the average distance of top epistasis inferred is significantly different from the average of overall pair sites distance. 112

List of Tables

2.1 Probabilities of recombination events in (2.51). The factor of $\frac{1}{2}$ arises because there is 50% chance of choosing genotypes c or d in the recombination process. The denominator is $L-1$ as there are only $L-1$ possible locations along the sequence length where a breakpoint (bpt.) can occur. 31

3.1 Bloom's lab data sets information 90

3.2 MaveDB data sets information 91

Chapter 1

Introduction

1.1 Motivation

Evolving populations are the core of biological reproduction processes, shaping the diversity of life on our planet. These populations are in a constant state of flux, undergoing genetic changes and adaptations driven by selective pressures and environmental factors. Exploring the intricacies of evolving populations provides crucial insights into the mechanisms of evolution, the emergence of new traits, and the persistence of species in changing environments.

An example of the importance of evolving populations can be seen in the influenza virus[127, 175]. The ability of the influenza virus to evolve is significant because it poses challenges for public health. The rapid evolution of the flu virus makes it difficult to develop long-lasting vaccines, as the circulating strains may differ from those included in the vaccine. Consequently, regular updates to the influenza vaccine are required to match the most prevalent strains. The ongoing evolution of influenza highlights the need for continual surveillance, research, and preparedness to effectively manage and control the disease.

A genome encompasses all the genetic material within an organism, comprising DNA (or RNA in certain viruses). It consists of genes as well as other elements responsible for regulating gene activity. Genetic variants on DNA or RNA sequences form the foundation of evolving populations, serving as the essential raw material that undergoes natural selection. Natural selection is proposed by Charles Darwin, referring to the process by which certain traits or characteristics become more or less common in a population over time, depending on their impact on survival and reproductive success. Over time of the evolutionary of genetic sequences, favorable traits become more prevalent, as they enhance an organism's ability to survive and reproduce, while less advantageous traits decrease in corresponding genetic variants frequency.

1.2 Quantifying mutational effects

In evolutionary biology, fitness effects refer to the measure of an organism's reproductive success in adapting to the environment and the ability to contribute its genetic information to the next generations. The fitness of a genetic sequence represents the relative ability of an organism with such a genetic sequence to survive, reproduce, and pass on its genes to future generations compared to other individuals with different genetic sequences. Fitness landscapes are intimately linked with the natural selection[11, 147]. A fitness landscape visualizes how the fitness of organisms or genetic sequences within a population relates to their genetic makeup. It represents the multidimensional space where each point corresponds to a specific combination of traits or genetic variants. Favorable traits that enhance a sequence's fitness lead to higher peaks or advantageous regions on the landscape. Conversely, less advantageous traits correspond to lower peaks or valleys. The process of natural selection can drive individuals with particular genetic sequences towards these higher fitness peaks through the differential survival and reproduction of individuals with more beneficial traits. Consequently, the dynamics of natural selection and the exploration of the fitness landscape influence the evolutionary paths that populations follow, allowing them to adapt and optimize their fitness according to their environments.

Naturally, the fitness landscapes can be complicated. Real fitness landscapes often exhibit ruggedness, with multiple peaks and valleys representing different combinations of traits or genetic variants because of thousands to billions of genetic base pairs on the sequences. By simplifying the fitness landscapes, this thesis focuses on how to infer the fitness effects, including the relative fitness of individual mutant to wild-type (selection coefficient) and fitness that differs from the sum of the fitness effects of each individual mutant (epistasis), from dynamic time series genetic sequence data. To achieve this, my research goal is to develop and implement efficient computational methods to interpret evolutionary dynamics quantitatively. The selection coefficients follow the additive rule, under which the fitness of any haplotype is a sum of selection coefficients at each sequence locus. The selection coefficient provides insights into the strength and direction of selection pressures, and by studying selection coefficients, researchers can predict the evolutionary trajectory of populations, estimate the rate at which genetic variation is changing, and gain a better understanding of how new traits or adaptations arise and spread[93, 108]. A high positive selection coefficient indicates a beneficial variant that enhances an organism's fitness, while a negative selection coefficient reflects a detrimental variant that reduces fitness.

In addition to more complicated scenarios, this work also includes the epistatic interaction in the inference framework. Understanding the nature of epistasis is vital for having insights into the complexity of genetic interactions and their impact on phenotypic outcomes[129]. By inferring selection coefficients and epistatic interactions, it provides insights into the evolutionary trajectories of genetic sequences, helping researchers to gain a deeper understanding of the dynamics of genetic variants along the evolution and how natural selection shapes the genetic composition of populations over time.

Recent advances in experiments have greatly improved the availability of temporal genetic data. This increased availability has the potential to enhance the accuracy and precision of detecting selection. However, it is important to note that inferring selection from temporal genetic time series data still poses technical challenges. Multiple factors interact and contribute to the observed patterns of genetic variation, including but not limited to numerous genetic variants combination, selection, mutation, genetic drift, and genetic linkage. Disentangling the effects of these factors and accurately quantifying selection coefficients and epistasis require advanced computational approaches and sophisticated statistical models. Certain existing computational methods for inferring fitness from population dynamics either ignore the effects of linkage or face significant computational challenges[49, 56, 84]. Especially in fitness models that incorporate pairwise epistasis terms, the number of model parameters to be inferred grows quadratically with the length of the genetic sequence. This means that as the sequence length increases, the number of parameters that need to be estimated also increases rapidly. Thus, developing a new approach is necessary to infer the fitness effects more efficiently, reliably, and in a more interpretable way.

1.3 Relationship with statistical physics

Evolutionary biology and statistical physics employ similar theoretical methodologies despite studying distinct phenomena. Statistical thermodynamics and the evolution of allele frequencies under mutation, selection, and random drift exhibit a strong analogy[10, 148, 172]. In statistical thermodynamics, the behavior of a system of particles is described using statistical mechanics, which employs probability distributions to study the collective properties of a large number of particles. The Boltzmann distribution, for example, relates the energy states of particles to their probabilities of occupation. Similarly, in population genetics, the evolution of allele frequencies in a population can be described using mathematical models that involve probabilities. Evolutionary models such as the Wright-Fisher model incorporate mutation, selection, and random drift to study changes in allele frequencies over time. In both cases, the probabilistic nature of the systems allows for the analysis of how macroscopic properties emerge from the behavior of individual constituents. This emergence of macroscopic behavior is a common theme in statistical mechanics and population genetics.

Some statistical physics methods could provide a robust framework for analyzing large-scale data sets and inferring the underlying principles governing genetic variation. I will leverage innovative approaches from statistical physics to implement reliable, scalable, and interpretable method, Marginal Path Likelihood (MPL)[155], for inferring the fitness impacts of mutations using temporal genetic data and extending the inference framework with epistatic interaction. This method will incorporate key factors including genetic linkage and epistasis. The approach involves quantifying the probability of an evolutionary path, which encompasses the mutant allele frequencies at each time point, using a path integral method derived from statistical physics. By analytically inverting this expression, we can identify the parameters that are most likely responsible for generating a given evolutionary path.

More challenges come from sequencing data quality, as different sequencing approaches offer insights into evolving populations. Low-throughput techniques, such as alanine scanning, focus on analyzing specific mutations by individually introducing alanine substitutions[105, 121]. While alanine scanning can be a valuable tool for studying specific aspects of protein sequences, alanine scanning is limited in scale, not providing detailed information about the entire protein structure or the arrangement of other amino acids. In contrast, high-throughput methods like Deep Mutational Scanning (DMS) allow for the simultaneous measurement of the functional effects of thousands of mutations across an entire protein or genome[60, 176]. DMS provides a comprehensive view of mutational effects and their interactions, enabling researchers to examine the dynamics of evolving populations on a larger scale. It involves creating a comprehensive library of protein variants, where each genetic site contains all possible single amino acid substitutions. These variants are then subjected to functional or phenotypic assays to measure their effects on protein stability, activity, binding affinity, or other desired properties. Although there exist some popular analytical methods or packages for DMS data analysis[16, 55, 58, 75, 141], these methods lack theoretical support and are sensitive to finite sampling with small frequencies. To improve the reliability and interpretability to infer the mutational effects, this work built an inference pipeline, popDMS, to help with analyzing DMS data.

Overall, quantifying the effects of mutations is essential for evolutionary biology in understanding the fitness consequences of mutations, explaining the process of adaptation and the emergence of new traits within a population. Although there are differences in the specific systems and processes studied, the underlying principles and mathematical tools employed in statistical thermodynamics and the evolution of allele frequencies exhibit a strong analogy, highlighting the power of probability and statistical methods in understanding complex systems. By unraveling the complex selection coefficients and epistasis using statistical physics methods, researchers can gain deeper insights into the adaptive processes that shape the diversity of life and the mechanisms behind evolutionary innovations. Continued advancements in statistical physics methods and high-throughput sequencing technologies promise to further enhance our understanding of evolving populations, opening new avenues for research and discovery in the field of evolutionary biology.

Chapter 2

Inferring epistasis from genetic time-series data

Epistasis refers to fitness effect of mutant alleles that differ from the sum of the fitness effects of each individual mutant [26, 36, 106, 129]. Epistasis therefore causes the fitness effect of a mutation to depend on the genetic background on which it arises.Theoretical and experimental studies have shown that epistasis can play a role in speciation [65, 173] and adaptation [31, 73], and that it is intertwined with the evolutionary advantages of recombination [37, 95]. Epistasis is not uncommon in nature, and signatures of strong epistasis have been observed in laboratory evolution and site-directed mutagenesis experiments [13, 68, 91, 143].

Epistasis makes fitness landscapes more complex, shaping evolution [35, 129]. For example, epistasis may make certain mutational pathways more difficult to traverse while others become more readily accessible, depending on the sequence background [182, 179, 129, 143, 35, 126]. A better understanding of epistasis could therefore help to characterize the evolutionary dynamics of novel viral strains capable of evading immune responses [83], pathogens that develop drug resistance [81, 192] and tumor growth in cancers [187, 174], as well as the adaptation of populations under lab settings [41].

Advances in sequencing technologies over the past decades have made it possible to obtain detailed, time-resolved population-level sequence data, enabling the study of evolving populations in fine detail. Examples of such data include those obtained from evolving populations in vitro [8], ones sampled from naturally-infected hosts [122, 190, 82, 186], and time-resolved global influenza evolutionary records [7]. These evolving populations contain multiple polymorphic loci, making the epistasis between mutant alleles a potential factor in the evolutionary dynamics of the population.

A complicating factor in inferring epistasis from such time-series data is the presence of linkage effects. Genetic linkage can arise by chance as a consequence of shared inheritance or for functional reasons due to epistatic interactions between linked loci. Linkage can be especially strong when recombination is low, selection is strong, and novel mutations frequently appear and compete in a population [38, 125, 152]. The ability to distinguish the effects of epistasis from linkage due to chance is therefore important for the reliable inference of fitness from genetic time-series data.

The large majority of existing methods for inferring the fitness effects of mutations from genetic data ignore epistasis in their modeling. Hence they do not estimate epistasis, nor do they account for epistatic effects when estimating the fitness advantage of an allele. Most existing methods are based on single-locus models which assume independent evolution of loci [18, 114, 117, 49, 100, 160, 56, 170, 52, 145, 67, 88, 166, 193], thus they are unable to directly account for genetic linkage or epistasis. A few methods [84, 167, 155] have been developed that consider the joint evolution of multiple loci, but these assume additive fitness models. Hence, while they account for genetic linkage, they do not consider epistasis. A notable exception are the methods that use an extension of the multi-locus approach of [84] to account for epistasic interactions [83, 82, 87]. These are based on a deterministic evolutionary model and, while an important advancement, require the use of computationally intensive numerical optimization methods.

2.1 Inference framework overview

Here we present a novel method that provides a closed-form, analytical solution for estimates of selection coefficients and pairwise epistatic interactions from genetic time-series data. Due to its analytical form, our approach is straightforward to implement and computationally efficient for moderate numbers of loci. Our method is based on an extension of the marginal path likelihood (MPL) framework [155] to account for epistasis. Here we use a path integral method derived from statistical physics [137] to efficiently represent the likelihood of an observed trajectory of single and double mutant allele frequencies. We then apply Bayesian theory to estimate the fitness parameters that best explain an observed evolutionary trajectory.

We model a population evolving under the Wright-Fisher (WF) model with mutation, selection, and recombination. First, we define $\mathbf{x}(t)$ as the vector of single and double mutant allele frequencies observed at generation t. For a system with L loci labeled by

$i = 1, 2, ..., L$, the first L entries of $\mathbf{x}(t)$ represent mutant allele frequencies $x_i(t)$, and entries from $L + 1$ to $R = L(L + 1)/2$ represent the frequencies of individuals in the populations with mutant alleles at loci i and j, denoted $x_{ij}(t)$. Under WF dynamics the probability of observing a trajectory or 'path' $(\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_K))$ conditioned on $\mathbf{x}(t_0)$ is given by

$$P\left((\mathbf{x}(t_k))_{k=1}^{K} | \mathbf{x}(t_0)\right) = \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right). \tag{2.1}$$

We approximate the probability in (2.1) with a path integral. The first step of this approach is to approximate the WF process by a diffusion process [92, 47, 164, 74, 165]. Under this approximation, the transition probabilities that appear on the right-hand side of (2.1) can be approximated by the transition probability density, ϕ, of a diffusion process [45], multiplied by a constant scaling term. In principle, $P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right)$ can be approximated using numerical integration techniques to solve the diffusion equations [18, 114, 52]. Such approaches, however, are computationally intensive and lead to expressions that are difficult to treat analytically, even at the single locus level. Instead, the path integral approach we take allows efficient computation of (2.1). It discretizes the transition probability density for small time steps, with the resulting approximate density taking a Gaussian form. Taking a Gaussian prior for the selection coefficients and epistasis parameters, and applying the maximum a posteriori criterion, we obtain an analytical expression for the estimates of selection coefficients and epistasis terms given the observed allele frequency trajectories (see 2.2 for details):

$$\hat{\mathbf{s}} = \left[\mathrm{C}_{\text{int}} + \gamma I\right]^{-1} \times \left[\Delta\mathbf{x} - \mu\ \mathrm{v}_{\text{int}} - r\ \eta_{\text{int}}\right]\ . \tag{2.2}$$

Here, $\hat{\mathbf{s}}$ is a vector of estimated selection coefficients and pairwise epistasis terms, C_{int} denotes the covariance matrix of single and double mutant allele frequencies integrated over time, γ is a regularization parameter, and I is the identity matrix. $\Delta\mathbf{x}$ gives the difference between the single and double mutant allele frequencies at the last and first time points. Finally, μ and r are the per-locus per-generation mutation and recombination rates, respectively, while $\mathrm{v}_{\mathrm{int}}$ and η_{int} are functions of single and double mutant allele frequencies integrated over time. We give explicit details for each of these terms in 2.2 and also show that the same analytical expression for the estimates of fitness parameters is obtained from both the allele and genotype-level analyses of the WF dynamics.

Below, we use simulations to demonstrate that our approach accurately infers fitness parameters using data from populations evolving under selection, mutation, recombination, epistasis, and nontrivial genetic linkage. We also show under which conditions reliable inference of selection and epistasis is possible. In cases where low data variability precludes the accurate inference of some fitness parameters, MPL is still able to infer their collective fitness contributions.

2.2 Design of the inference framework

2.2.1 Evolutionary model

We consider a population of N individuals evolving under a WF model with selection, mutation and recombination. Each individual is represented by a sequence of length L. The loci are assumed to be bi-allelic where each locus is either 0 (wild-type (WT)) or 1 (mutant), thus resulting in $M = 2^L$ genotypes.

We consider a fitness model that accounts for epistasis arising due to pairwise interactions between alleles at different loci. The Wrightian fitness f_a of the ath genotype can then be written as

$$f_a = 1 + \sum_{i=1}^{L} s_i g_i^a + \sum_{i=1}^{L} \sum_{j=i+1}^{L} s_{ij} g_i^a g_j^a, \tag{2.3}$$

where s_i and s_{ij} denote the time-invariant selection coefficients and pairwise epistasis terms respectively, and g_i^a represents the allele (either 0 or 1) at the ith locus of the ath genotype. We can compactly denote the selection coefficients and epistasis terms in a single vector as

$$\mathbf{s} = \big(s_1, \cdots, s_L, s_{12}, \cdots, s_{(L-1)L}\big) \tag{2.4}$$

where the first L elements are the selection coefficients while the last $L(L-1)/2$ elements are pairwise epistasis terms. Similar to the notation adopted in the main text, we differentiate between non-italic and italic scalar notation to facilitate sequential indexing throughout the supplementary text. Thus we write

$$\mathbf{s} = \Big(\mathrm{s}_1, \cdots, \mathrm{s}_L, \mathrm{s}_{L+1}, \cdots, \mathrm{s}_R\Big) \tag{2.5}$$

where $R = L(L+1)/2$ and we have $\mathrm{s}_e = s_i$ for $e \in \{1, \ldots, L\}$, and $\mathrm{s}_e = s_{ij}$ for $e \in \{L+1, \ldots, R\}$, with obvious association between indices e and (i, j).

The population is completely specified by the $M \times 1$ genotype frequency vector $\mathbf{z}(t) = (z_1(t), \cdots, z_M(t))$, where $z_a(t) = n_a(t)/N$ and $n_a(t)$ denotes the number of individuals in the population that belong to genotype a at generation t. Let r be the probability of recombination per locus per generation. The frequency of genotype a at generation t after recombination is given by

$$y_a(t) = (1-r)^{L-1} z_a(t) + \big(1 - (1-r)^{L-1}\big)\, \psi_a(\mathbf{z}(t)) \tag{2.6}$$

where $(1-r)^{L-1}z_a(t)$ represents the fraction of genotype a not undergoing recombination, $\left(1-(1-r)^{L-1}\right)\psi_a(\mathbf{z}(t))$ the fraction of genotype a arising as a result of recombination, and the factor $L-1$ arises as there are $L-1$ possible recombination breakpoints. The quantity $\psi_a(\mathbf{z}(t))$ is the probability that a recombination event results in an individual of genotype a and is a function of the composition of the population at generation t. We represent this quantity as

$$\psi_a(\mathbf{z}(t)) = \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}z_c(t)z_d(t) \tag{2.7}$$

where $R_{a,cd}$ is the probability that genotypes c and d recombine to form genotype a and is a function of the number of breakpoints and the particular genotypes a, c and d. We describe this in detail later in the document, when we calculate the recombination term in (2.43) and (2.49).

Under WF dynamics, the probability of observing genotype frequencies $\mathbf{z}(t+1)$ at generation $t+1$, given genotype frequencies of $\mathbf{z}(t)$ at generation t is

$$P\left(\mathbf{z}(t+1)\middle|\mathbf{z}(t)\right) = N!\prod_{a=1}^{M}\frac{\left(p_a(\mathbf{z}(t))\right)^{Nz_a(t+1)}}{(Nz_a(t+1))!} \tag{2.8}$$

with

$$p_a(\mathbf{z}(t)) = \frac{y_a(t)f_a + \sum_{b\neq a}\left(\mu_{ba}y_b(t)f_b - \mu_{ab}y_a(t)f_a\right)}{\sum_{b=1}^{M} y_b(t)f_b}. \tag{2.9}$$

Here μ_{ba} is the probability of genotype b mutating to genotype a, and $y_a(t)$ is the frequency of genotype a after recombination

$$y_a(t) = (1-r)^{L-1}z_a(t) + \left(1-(1-r)^{L-1}\right)\psi_a(\mathbf{z}(t)), \tag{2.10}$$

where r is the recombination probability per locus per generation and $\psi_a(\mathbf{z}(t))$ is the probability that a recombination of two individuals results in an individual of genotype a.

We assume the genotype frequencies are observed at non-consecutive generations t_k, with $k \in \{0, 1, \dots, K\}$. Then, the probability that the genotype frequency vector follows a particular evolutionary path $(\mathbf{z}(t_1), \mathbf{z}(t_2), \cdots, \mathbf{z}(t_K))$, conditioned on the initial state $\mathbf{z}(t_0)$, is

$$P\left((\mathbf{z}(t_k))_{k=1}^{K} \, | \mathbf{z}(t_0) \right) = \prod_{k=0}^{K-1} P\left(\mathbf{z}(t_{k+1}) | \mathbf{z}(t_k)\right). \tag{2.11}$$

This expression can be used to infer evolutionary parameters. However, the inference problem is difficult due to the intractability of the fractional form of (3.9). Following the approach used in [155], we simplify the inference problem using a path integral. This allows us to obtain closed-form estimates of selection coefficients and epistasis terms. Even though the WF dynamics is defined at the genotype level (2.11), here we develop its simplified allele-level version for transparency. We show in the supplement that both the genotype and allele-level analyses lead to the same expression for the estimate of fitness parameters. For ease of exposition, we assume here that the probability of mutating from a WT to mutant allele is the same as that from mutant allele to WT, which we denote by μ. However, this assumption can be easily relaxed in 2.2 for details where we derive the estimator with asymmetrical mutation probabilities.

2.2.2 Path integral:

We model the evolution of both the single and the double mutant allele frequencies. In the allele-level path integral, these are required to obtain estimates of the selection coefficients and the pairwise epistasis terms.

The probability of observing a path of allele frequencies $(\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_K))$ conditioned on $\mathbf{x}(t_0)$ is given by

$$P\Big(\left(\mathbf{x}(t_k)\right)_{k=1}^{K} |\mathbf{x}(t_0)\Big) = \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)\right). \tag{2.12}$$

In principle, we can use the above expression to infer evolutionary parameters. However, the expression above is unyielding for the purpose of parameter inference due to the large dimensionality of the genotype space, which grows exponentially with sequence length, as well as intractability of the fractional form of the right hand side of (2.11). To simplify the problem, we use a path integral to approximate the probability in (2.11).

The first step of the approach consists of approximating the WF process by a diffusion process, as commonly done in population genetics [92, 47, 164, 74, 165]. Specifically, assume the population is large and that

$$\mathrm{s}_e = \frac{\bar{\mathrm{s}}_e}{N} + O\left(\frac{1}{N^2}\right), \quad \mu = \frac{\bar{\mu}}{N} + O\Big(\frac{1}{N^2}\Big) \quad r = \frac{\bar{r}}{N} + O\Big(\frac{1}{N^2}\Big), \tag{2.13}$$

and consequently

$$h_a = \frac{\bar{h}_a}{N} + O\Big(\frac{1}{N^2}\Big), \tag{2.14}$$

where $\bar{s}_e$, $\bar{\mu}$, $\bar{r}$, and $\bar{h}_a$ are constants that are independent of N. Under this scaling, we have

$$y_a(t) = z_a(t) - r(L-1)\big(z_a(t) - \psi_a(\mathbf{z}(t))\big) + O\left(\frac{1}{N^2}\right) \tag{2.15}$$

where L is also assumed to be constant with regards to N.

2.2.3 Genotype-level path integral

In [155], the authors derived the genotype-level path integral to approximate the probability of observing a trajectory of genotype frequencies $(\mathbf{z}(t_1), \mathbf{z}(t_2), \cdots, \mathbf{z}(t_K))$. Since this analysis applies equally to the current work, we give a brief summary here. We approximated the transition probability of the WF evolutionary process, using standard diffusion theory [47], by the transition probability density of a diffusion process, i.e.,

$$\check{\mathbf{Z}}(\tau) = \big(\check{Z}_1(\tau), \ldots, \check{Z}_M(\tau)\big) := \check{\mathbf{Z}}(\lfloor N\tau \rfloor), \quad \tau \geq 0 \tag{2.16}$$

taken in the limit $N \to \infty$. Here $\lfloor \cdot \rfloor$ denotes the floor function and τ is a continuous time variable with units of N generations, with one generation in discrete time (i.e., from t to $t+1$) thus taking

$$\delta\tau = \frac{1}{N} \tag{2.17}$$

continuous time units. The genotype-level diffusion process was found to be characterized by the drift vector $\bar{d}(\check{\mathbf{z}}(\tau))$ with ath entry

$$\begin{aligned} \bar{d}_a(\check{\mathbf{z}}(\tau)) = {} & \check{z}_a(\tau)\left(\bar{h}_a - \sum_{b=1}^{M} \bar{h}_b \check{z}_b(\tau)\right) + \bar{\mu}\left(\sum_{b=1, d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1, d_{ab}=1}^{M} \check{z}_a(\tau)\right) \\ & - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\left(\check{\mathbf{z}}(\tau)\right)\big), \end{aligned} \tag{2.18}$$

and diffusion matrix $\bar{C}(\check{\mathbf{z}}(\tau))$ with (a,b)th entry

$$\bar{C}_{ab}(\check{\mathbf{z}}(\tau)) = \frac{1}{2}\begin{cases} \check{z}_a(\tau)(1 - \check{z}_a(\tau)) & a = b \\ -\check{z}_a(\tau)\check{z}_b(\tau) & a \neq b. \end{cases} \tag{2.19}$$

The time evolution of the transition probability density of the diffusion process $\check{\mathbf{Z}}(\tau)$ is described by the Kolmogorov forward equation (also known as the Fokker-Planck equation).

Discretizing the transition probability density of the diffusion over small $\delta t \Delta t$ (equivalently large $\frac{N}{\Delta t}$), we approximated the probability of observing a trajectory of genotype frequencies $(\mathbf{z}(t_1), \mathbf{z}(t_2), \cdots, \mathbf{z}(t_K))$ conditioned on $\mathbf{z}(t_0)$ as

$$\begin{aligned} P\left((\mathbf{z}(t_k))_{k=1}^{K} | \mathbf{z}(t_0) \right) &= \prod_{k=0}^{K-1} P(\mathbf{z}(t_{k+1}) | \mathbf{z}(t_k)) \\ &\approx \prod_{k=0}^{K-1} \left[\frac{1}{\sqrt{\det C(\mathbf{z}(t_k))}} \left(\frac{N}{2\pi \Delta t_k} \right)^{M/2} \prod_{a=1}^{M} \mathrm{d}z_a(t_{k+1}) \right] \times \\ &\qquad \exp\left(-\frac{N}{2} S\left((\mathbf{z}(t_k))_{k=0}^{K} \right) \right) \end{aligned} \tag{2.20}$$

with

$$\begin{aligned} S\left((\mathbf{z}(t_k))_{k=0}^{K} \right) = \sum_{k=0}^{K-1} \frac{1}{\Delta t_k} \sum_{a=1}^{M} \sum_{b=1}^{M} &[z_a(t_{k+1}) - z_a(t_k) - d_a(\mathbf{z}(t_k)) \Delta t_k] \times \\ &\left(C^{-1}(\mathbf{z}(t_k)) \right)_{ab} [z_b(t_{k+1}) - z_b(t_k) - d_b(\mathbf{z}(t_k)) \Delta t_k], \end{aligned}$$

where $\Delta t_k = t_{k+1} - t_k$ and we have defined $d_a(\mathbf{z}(t_k)) := \frac{\bar{d}_a(\mathbf{z}(t_k))}{N}$, $\left(C(\mathbf{z}(t_k))\right)_{ab} := 2\left(\bar{C}(\mathbf{z}(t_k))\right)_{ab}$. This is the path integral representation of the genotype dynamics

2.2.4 Allele-level path integral

In [155], the authors modeled the evolution of the single mutant frequencies by applying linear combinations to the genotype frequencies described by (2.20), while assuming the double mutant frequencies were known. The single and double mutant frequencies relate to the genotype frequencies via

$$x_i(t) = \sum_{a=1}^{M} g_i^a z_a(t), \qquad x_{ij}(t) = \sum_{a=1}^{M} g_i^a g_j^a z_a(t), \tag{2.21}$$

where $x_i(t)$, and $x_{ij}(t)$, are the single and the double mutant frequencies at locus i and locus-pair (i, j) respectively at generation t. Here, in contrast, we model the evolution of both the single and double mutant frequencies which additionally requires the knowledge of the triple and quadruple mutant frequencies. These are related to the genotype frequencies via

$$x_{ijk}(t) = \sum_{a=1}^{M} g_i^a g_j^a g_k^a z_a(t), \quad x_{ijkl}(t) = \sum_{a=1}^{M} g_i^a g_j^a g_k^a g_l^a z_a(t), \tag{2.22}$$

where $x_{ijk}(t)$ and $x_{ijkl}(t)$ are the triple and the quadruple mutant frequencies at locus-triplet (i, j, k) and locus-quartet (i, j, k, l) respectively at generation t.

We concatenate the single and double mutant allele frequencies in a R length vector, where $R = L(L+1)/2$, as

$$\mathbf{x}(t) = \left(x_1(t), \cdots, x_L(t), x_{12}(t), x_{13}(t), \cdots, x_{(L-1)L}(t)\right). \tag{2.23}$$

Similar to the notation in the main text, we write

$$\mathbf{x}(t) = \Big(\mathrm{x}_1(t), \cdots, \mathrm{x}_L(t), \mathrm{x}_{L+1}(t), \cdots, \mathrm{x}_R(t)\Big) \tag{2.24}$$

to facilitate sequential indexing for notation convenience. Note that we differentiate between non-italic and italic scalar notation, as described in (2.4) and (2.5). Clearly, from (2.23) and (2.24), we have $\mathrm{x}_e(t) = x_i(t)$ for $e \leq L$, and $\mathrm{x}_e(t) = x_{ij}(t)$ for $L < e \leq R$.

To simplify the presentation, we also define U as an $M \times R$ mapping matrix where the ath row of U, i.e., $\mathbf{u}_a = \left(\mathrm{u}_1^a, \cdots, \mathrm{u}_L^a, \mathrm{u}_{L+1}^a, \cdots, \mathrm{u}_R^a\right)$, is given by

$$\mathbf{u}_a = \Big(g_1^a,\ \cdots,\ g_L^a,\ g_1^a\ g_2^a,\ \cdots,\ g_1^a\ g_L^a,\ g_2^a\ g_3^a,\ \cdots,\ g_2^a\ g_L^a,\ \cdots, g_{L-1}^a\ g_L^a\Big). \tag{2.25}$$

Note that g_i^a refers to the allele at the ith locus while $g_i^a g_j^a$ refers to the pair of alleles at locus-pair (i, j) in genotype a.

Next, we define a random vector comprising of the single and double mutant allele frequencies, i.e., $\mathbf{X}(t) = \big(X_1(t), \ldots, X_L(t), X_{12}(t), \ldots, X_{(L-1)L}(t)\big)$, which from (2.21) is related to the random genotype frequency vector by

$$\mathrm{X}_e(t) = \sum_{a=1}^{M} \mathrm{u}_e^a Z_a(t) \tag{2.26}$$

where u_e^a denotes the eth entry of $\mathbf{u}_a$.

Thus, $\mathbf{x}(t)$ is a realization of the random vector $\mathbf{X}(t)$. Similarly, the allele-level continuous process can be shown to be related to the genotype-level continuous time process (2.16) using the transformation above, and is given as

$$\check{\mathbf{X}}(\tau) = \big(\check{X}_1(\tau), \ldots, \check{X}_L(\tau), \check{X}_{12}(\tau)\check{X}_{13}(\tau), \ldots, \check{X}_{(L-1)L}(\tau)\big) := \mathbf{X}(\lfloor N\tau \rfloor), \quad \tau \geq 0 \tag{2.27}$$

taken as $N \to \infty$.

The time evolution of the transition probability density, ϕ, of the allele-level diffusion, is governed by the Kolmogorov forward equation

$$\frac{\partial \phi}{\partial \tau} = \left[-\sum_{e=1}^{R} \frac{\partial}{\partial \check{\mathrm{x}}_e} \bar{\mathrm{d}}_e(\check{\mathbf{x}}(\tau)) + \sum_{e=1}^{R}\sum_{f=1}^{R} \frac{\partial}{\partial \check{\mathrm{x}}_e}\frac{\partial}{\partial \check{\mathrm{x}}_f} \bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau)) \right] \phi, \tag{2.28}$$

where $\bar{C}(\check{\mathbf{x}}(\tau))$ and $\bar{\mathbf{d}}(\check{\mathbf{x}}(\tau))$ are the diffusion matrix and the drift vector associated with the allele-level diffusion process that describes the conditional change in the single and double mutant frequencies.

The diffusion matrix of the allele-level diffusion process is of size $R \times R$ and can be partitioned into four sub-matrices, i.e., the upper left $L \times L$ matrix, the upper right $L \times \frac{L(L-1)}{2}$ matrix, lower left $\frac{L(L-1)}{2} \times L$ matrix and the lower right $\frac{L(L-1)}{2} \times \frac{L(L-1)}{2}$ matrix. The definition and interpretation of these matrices is given below. Recalling (2.25), we note that the eth element of $\mathbf{u}_a$ refers to the allele at locus i for $1 \leq e \leq L$, and to the alleles at locus-pair (i, j) for $L < e \leq R$, i.e.,

$$u_e^a = \begin{cases} g_i^a & 1 \leq e \leq L \\ \\ g_i^a g_j^a & L < e \leq R. \end{cases} \tag{2.29}$$

The elements of the upper left sub-matrix of the diffusion matrix $\bar{C}(\check{\mathbf{x}}(\tau))$, i.e., $1 \leq e \leq L$ and $1 \leq f \leq L$, are given as

$$\begin{aligned} \bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau)) &:= \sum_{a=1}^{M}\sum_{b=1}^{M} \mathrm{u}_e^a \mathrm{u}_f^b \bar{\mathrm{C}}_{ab}(\check{\mathbf{z}}(\tau)) \\ &= \frac{1}{2}\sum_{a=1}^{M} g_i^a g_j^a \frac{\check{z}_a(\tau)(1-\check{z}_a(\tau))}{N} - \frac{1}{2}\sum_{a=1}^{M}\sum_{b=1, b\neq a}^{M} g_i^a g_j^b \frac{\check{z}_a(\tau)\check{z}_b(\tau)}{N} + O\left(\frac{1}{N^2}\right) \\ &= \frac{1}{2}\frac{\check{x}_{ij}(\tau) - \check{x}_i(\tau)\check{x}_j(\tau)}{N} + O\left(\frac{1}{N^2}\right), \end{aligned} \tag{2.30}$$

which measure the scaled joint variability between the number of mutants at loci i and j. We note here that the upper left sub-matrix here is the same as the diffusion matrix in [155] where only the evolution of the single mutant allele frequency was modeled.

Following similar steps, it can be shown that the entries of the upper right sub-matrix of the diffusion matrix $\bar{C}(\check{\mathbf{x}}(\tau))$, i.e., for $1 \leq e \leq L$ and $L < f \leq R$, are given as

$$\begin{aligned} \bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau)) &:= \sum_{a=1}^{M}\sum_{b=1}^{M} \mathrm{u}_e^a \mathrm{u}_f^b \bar{\mathrm{C}}_{ab}(\check{\mathbf{z}}(\tau)) \\ &= \sum_{a=1}^{M}\sum_{b=1}^{M} g_i^a g_j^b g_k^b \bar{\mathrm{C}}_{ab}(\check{\mathbf{z}}(\tau)) \\ &= \frac{1}{2}\frac{\check{x}_{ijk}(\tau) - \check{x}_i(\tau)\check{x}_{jk}(\tau)}{N} + O\left(\frac{1}{N^2}\right), \end{aligned} \tag{2.31}$$

which measures the scaled joint variability between the number of mutants at locus i and double-mutants at loci j and k.

Here $\check{x}_{ijk}(\tau)$ denotes the triple mutant frequency obtained by the transformations (2.21) and (2.22) with

$$\check{x}_{ijk}(\tau) := x_{ijk}(\lfloor N\tau \rfloor), \quad \tau \geq 0. \tag{2.32}$$

The $\frac{L(L-1)}{2} \times L$ lower left sub-matrix is just the transpose of the $L \times \frac{L(L-1)}{2}$ upper right matrix. Similarly, the entries of the bottom right sub-matrix of the diffusion matrix $\bar{C}(\check{\mathbf{x}}(\tau))$, i.e., $L < e \leq R$ and $L < e \leq R$, are given as =

$$\begin{aligned}\bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau)) &:= \sum_{a=1}^{M}\sum_{b=1}^{M} \mathrm{u}_e^a \mathrm{u}_f^b \bar{\mathrm{C}}_{ab}(\check{\mathbf{z}}(\tau)) \\ &= \sum_{a=1}^{M}\sum_{b=1}^{M} g_i^a g_j^a g_k^b g_l^b \bar{\mathrm{C}}_{ab}(\check{\mathbf{z}}(\tau)) \\ &= \frac{1}{2}\frac{\check{x}_{ijkl}(\tau) - \check{x}_{ij}(\tau)\check{x}_{kl}(\tau)}{N} + O\left(\frac{1}{N^2}\right), \end{aligned} \tag{2.33}$$

which measures the scaled joint variability between the number of double-mutants at loci i and j, and double-mutants at loci k and l, with $\check{x}_{ijkl}(\tau)$ denoting the quadruple mutant frequency with

$$\check{x}_{ijkl}(\tau) := x_{ijkl}(\lfloor N\tau \rfloor), \quad \tau \geq 0. \tag{2.34}$$

Note that while the diffusion matrix also depends on the dynamics of the triple and quadruple mutant frequencies, we only explicitly show the dependence on the single and double mutant frequencies for simplicity of notation.

We can show that the allele-level drift vector is a linear transformation of the genotype drift vector $\bar{d}_a(\check{\mathbf{z}}(\tau))$ defined in (2.18). Recalling (2.21), (2.22), and noting that we can express h_a in (3.3) as

$$h_a = \sum_{e=1}^{R} \mathrm{u}_e^a \mathrm{s}_e, \tag{2.35}$$

and the eth element of the allele-level drift vector is defined with a linear transformation of the genotype drift vector as

$$\begin{aligned}
\bar{\mathrm{d}}_e(\check{\mathbf{x}}(\tau)) &:= \sum_{a=1}^{M} \mathrm{u}_e^a \bar{d}_a(\check{\mathbf{z}}(\tau)) \\
&= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\check{z}_a(\tau) \left(\bar{h}_a - \sum_{b=1}^{M} \bar{h}_b \check{z}_b(\tau) \right) + \\
&\quad \bar{\mu} \left(\sum_{b=1,d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \check{z}_a(\tau) \right) - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\left((\tau)\right)\big) \Bigg) \\
&= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\check{z}_a(\tau)\big(1 - \check{z}_a(\tau)\big)\bar{h}_a - \check{z}_a(\tau) \sum_{b=1,b\neq a}^{M} \bar{h}_b \check{z}_b(\tau) + \\
&\quad \bar{\mu} \left(\sum_{b=1,d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \check{z}_a(\tau) \right) - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\big((\tau)\big)\big) \Bigg) \\
&= \check{\mathrm{x}}_e(\tau)\left(1 - \check{\mathrm{x}}_e(\tau)\right)\bar{\mathrm{s}}_e + \sum_{f\neq e} \bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau))\bar{\mathrm{s}}_f + \bar{\mu}\mathrm{v}_e(\check{\mathbf{x}}(\tau)) + \bar{r}\ \eta_e(\check{\mathbf{x}}(\tau)).
\end{aligned} \tag{2.36}$$

The transformation of the third and the fourth terms on the right hand side of (2.36), referred to here as the mutation term $\mathrm{v}_e(\check{\mathbf{x}}(\tau))$ and the recombination term $\eta_e(\check{\mathbf{x}}(\tau))$ respectively, is non-trivial and requires some algebraic manipulation which we detail below. We note here that the first L entries of $\bar{\mathrm{d}}_e(\check{\mathbf{x}}(\tau))$ constitute the drift vector of [155]. While the transformation of the first L entries mutation and recombination terms were derived in the Supplementary Information of [155], we reproduce these here as they aid in understanding the notation and subsequent derivation of remaining entries $L < e \leq R$ of the mutation and recombination terms.

Here, we show the computations involved with the mutation term in going from the second last line of (2.36) to the last line of (2.36). First consider the case $1 \leq e \leq L$,

$$\begin{aligned}
\mathrm{v}_e(\check{\mathbf{x}}(\tau)) &= \sum_{a=1}^{M} \mathrm{u}_e^a \left(\sum_{b=1, d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1, d_{ab}=1}^{M} \check{z}_a(\tau) \right) \\
&= \sum_{a=1}^{M} g_i^a \left(\sum_{b=1, d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1, d_{ab}=1}^{M} \check{z}_a(\tau) \right) \\
&= \sum_{a=1}^{M} \left(\sum_{b=1, d_{ab}=1}^{M} g_i^a (1 - g_i^b) \check{z}_b(\tau) + \sum_{b=1, d_{ab}=1}^{M} g_i^a g_i^b \check{z}_b(\tau) - \right. \\
&\qquad \left. \sum_{b=1, d_{ab}=1}^{M} g_i^a (1 - g_i^b) \check{z}_a(\tau) - \sum_{b=1, d_{ab}=1}^{M} g_i^a g_i^b \check{z}_a(\tau) \right) \\
&= \sum_{a=1}^{M} \sum_{b=1, d_{ab}=1}^{M} g_i^a (1 - g_i^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau) \right) + \sum_{a=1}^{M} \sum_{b=1, d_{ab}=1}^{M} g_i^a g_i^b \left(\check{z}_b(\tau) - \check{z}_a(\tau) \right).
\end{aligned} \tag{2.37}$$

Where the second last line above follows from noting that the first summation on the right side of the third last line can be decomposed into two parts. The first where genotypes a and b differ only at locus i, and hence mutation of genotype b to genotype a changes the mutant allele frequency at locus i. The second is where the two genotypes differ from each other at a locus other than i and hence a mutation from genotype b to a does not effect the mutant allele frequency at locus i. Similarly, the second summation in the third last line can also be split into two parts. Now, note that

$$\sum_{a=1}^{M} \sum_{b=1, d_{ab}=1}^{M} g_i^a g_i^b \left(\check{z}_b(\tau) - \check{z}_a(\tau) \right) = 0,$$

as this quantity represents the mutation of those genotypes b to genotype a where both a and b have the mutant allele at locus i, while

$$\sum_{a=1}^{M} \sum_{b=1, d_{ab}=1}^{M} g_i^a (1 - g_i^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau) \right) = 1 - 2\check{x}_i(\tau),$$

which represents the flow of mutational probabilities between the WT and the mutation allele, i.e., the mutational flux. Substituting the above two equations back in (2.37) yields

$$\mathrm{v}_e(\check{\mathbf{x}}(\tau)) = 1 - 2\check{x}_i(\tau) \qquad \text{for} \quad 1 \leq e \leq L. \tag{2.38}$$

Now consider the case when $L < e \leq R$ where we have

$$\begin{aligned}
\mathrm{v}_e(\check{\mathbf{x}}(\tau)) &= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\sum_{b=1,d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \check{z}_a(\tau) \Bigg) \\
&= \sum_{a=1}^{M} g_i^a g_j^a \Bigg(\sum_{b=1,d_{ab}=1}^{M} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \check{z}_a(\tau) \Bigg) \\
&= \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b)(1-g_j^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) + \\
&\quad \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b) g_j^b \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) + \\
&\quad \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b (1-g_j^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) + \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b g_j^b \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right).
\end{aligned} \tag{2.39}$$

Here, the summations on the right side of the last line above represent the net mutational flow to genotypes that contain alleles $(1,1)$ at locus-pair (i,j), from those genotypes that have alleles $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ respectively at locus-pair (i,j). We note that

$$\begin{aligned}
\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b)(1-g_j^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) &= 0 \\
\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b) g_j^b \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) &= \left(\check{x}_{ij}^{01}(\tau) - \check{x}_{ij}^{11}(\tau)\right) \\
\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b (1-g_j^b) \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) &= \left(\check{x}_{ij}^{10}(\tau) - \check{x}_{ij}^{11}(\tau)\right) \\
\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b g_j^b \left(\check{z}_b(\tau) - \check{z}_a(\tau)\right) &= 0
\end{aligned}$$

where we use the notation $\check{x}_{ij}^{10}(\tau)$ to refer to mutant with alleles $(1,0)$ and locus-pair (i,j). The first equation equals zero as the probability of more than one mutation in a sequence is negligibly small $O(\frac{1}{N^2})$ under the diffusion approximation. While the last equation equals zero as it represents the mutational flow of all those genotypes where both genotypes a and b contain the alleles $(1,1)$ at locus-pair (i,j). Substituting the above in (2.39) we have

$$\begin{aligned}
\mathrm{v}_e(\check{\mathbf{x}}(\tau)) &= \check{x}_{ij}^{01}(\tau) + \check{x}_{ij}^{10}(\tau) - 2\check{x}_{ij}^{11}(\tau) \\
&= \check{x}_{ij}^{01}(\tau) + \check{x}_{ij}^{11}(\tau) + \check{x}_{ij}^{10}(\tau) + \check{x}_{ij}^{11}(\tau) - 4\check{x}_{ij}^{11}(\tau) \\
&= \check{x}_{i}^{1}(\tau) + \check{x}_{j}^{1}(\tau) - 4\check{x}_{ij}^{11}(\tau) \\
&= \check{x}_{i}(\tau) + \check{x}_{j}(\tau) - 4\check{x}_{ij}(\tau), \qquad (2.40)
\end{aligned}$$

where we have dropped the superscripts in the last line. Thus, from (2.38) and (2.40) we have

$$\mathrm{v}_e(\check{\mathbf{x}}(\tau)) = \begin{cases} 1 - 2\check{x}_i(\tau) & 1 \leq e \leq L \\ \check{x}_i(\tau) + \check{x}_j(\tau) - 4\check{x}_{ij}(\tau) & L < e \leq R, \end{cases} \qquad (2.41)$$

where i and j are the subscript indices corresponding to the eth element of $\check{\mathbf{x}}(\tau)$.

Next we show the computations involved with the recombination term in going from the second last line of (2.36) to the last line of (2.36). First consider the case $1 \leq e \leq L$, for which

$$\begin{aligned}
\eta_e(\check{\mathbf{x}}(\tau)) &= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg((L-1)\big(\check{z}_a(\tau) - \psi_a\big((\tau)\big)\Bigg) \\
&= \sum_{a=1}^{M} g_i^a \Bigg((L-1)\big(\check{z}_a(\tau) - \psi_a\big((\tau)\big)\Bigg) \\
&= (L-1)\check{x}_i(\tau) - (L-1)\sum_{a=1}^{M} g_i^a \psi_a\big((\tau)\big) \\
&= (L-1)\check{x}_i(\tau) - (L-1)\sum_{a=1}^{M} g_i^a \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}\check{z}_c(\tau)\check{z}_d(\tau),
\end{aligned} \tag{2.42}$$

where the second line follows from (2.25), and the last line follows by substituting the definition of $\psi_a\big((\tau)\big)$ from (2.7). To further simplify, let

$$\theta_i^{cd} := \sum_{a=1}^{M} g_i^a R_{a,cd} \tag{2.43}$$

which is the probability that genotypes c and d recombine to form a genotype that has a mutation at locus i. For the bi-allelic model considered here, there are four possible scenarios for a recombination event: both genotypes c and d have allele 1 at their respective i-th locus, one of the genotypes has allele 1 while the other has allele 0, or both genotypes have allele 0 at the i-th locus.

We partition the summation term on the right side of (2.42) into these four recombination scenarios as follows

$$\begin{aligned}\sum_{a=1}^{M} g_i^a \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}\check{z}_c(\tau)\check{z}_d(\tau) &= \sum_{c=1}^{M}\sum_{d=1}^{M} \theta_i^{cd}\check{z}_c(\tau)\check{z}_d(\tau) \\ &= \sum_{c=1}^{M}\left(\sum_{d=1}^{M} g_i^c g_i^d \theta_i^{cd}\check{z}_c(\tau)\check{z}_d(\tau) + \sum_{d=1}^{M} g_i^c(1-g_i^d)\theta_i^{cd}\check{z}_c(\tau)\check{z}_d(\tau)\right) + \\ &\quad \sum_{c=1}^{M}\left(\sum_{d=1}^{M}(1-g_i^c)g_i^d\theta_i^{cd}\check{z}_c(\tau)\check{z}_d(\tau) + \sum_{d=1}^{M}(1-g_i^c)(1-g_i^d)\theta_i^{cd}\check{z}_c(\tau)\check{z}_d(\tau)\right).\end{aligned} \tag{2.44}$$

Note that

$$\begin{aligned} g_i^c g_i^d \theta_i^{cd} &= g_i^c g_i^d \\ g_i^c(1-g_i^d)\theta_i^{cd} &= \frac{1}{2} g_i^c(1-g_i^d) \\ (1-g_i^c)g_i^d\theta_i^{cd} &= \frac{1}{2}(1-g_i^c)g_i^d \\ (1-g_i^c)(1-g_i^d)\theta_i^{cd} &= 0 \end{aligned} \tag{2.45}$$

where the factor of $\frac{1}{2}$ arises because there is a 50% chance that genotype c (d) with a mutant at locus i and genotype d (c) with a wildtype at locus i will recombine to a genotype with a mutant at locus i. Hence, we can further write (2.44) as

$$\begin{aligned}\sum_{a=1}^{M} g_i^a \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}\check{z}_c(\tau)\check{z}_d(\tau) &= \sum_{c=1}^{M}\left(\sum_{d=1}^{M} g_i^c g_i^d \check{z}_c(\tau)\check{z}_d(\tau) + \frac{1}{2}\sum_{d=1}^{M} g_i^c(1-g_i^d)\check{z}_c(\tau)\check{z}_d(\tau)\right) \\ &+ \frac{1}{2}\sum_{c=1}^{M}\sum_{d=1}^{M}(1-g_i^c)g_i^d\check{z}_c(\tau)\check{z}_d(\tau) \\ &= \check{x}_i^2(\tau) + \frac{1}{2}\check{x}_i(\tau)(1-x_i(\tau)) + \frac{1}{2}\check{x}_i(\tau)(1-\check{x}_i(\tau)) \\ &= \check{x}_i(\tau).\end{aligned} \tag{2.46}$$

Substituting (2.46) back into (2.42), after some simple mathematical operations and simplification, we see that

$$\eta_e(\check{\mathbf{x}}(\tau)) = 0 \qquad \text{for} \quad 1 \leq e \leq L. \tag{2.47}$$

Now consider the case when $L < e \leq R$. Developing as in (2.42), we get

$$\eta_e(\check{\mathbf{x}}(\tau)) = (L-1)\check{x}_{ij}(\tau) - (L-1)\sum_{a=1}^{M} g_i^a g_j^a \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}\check{z}_c(\tau)\check{z}_d(\tau). \tag{2.48}$$

To simplify, we define $\theta_{ij}^{cd} := \sum_{a=1}^{M} g_i^a g_j^a R_{a,cd}$ where θ_{ij}^{cd} is the probability that genotypes c and d recombine to form a genotype which has a double mutant at locus-pair (i, j). We thus have

$$\sum_{a=1}^{M} g_i^a g_j^a \sum_{c=1}^{M}\sum_{d=1}^{M} R_{a,cd}\check{z}_c(\tau)\check{z}_d(\tau) = \sum_{c=1}^{M}\sum_{d=1}^{M} \theta_{ij}^{cd}\check{z}_c(\tau)\check{z}_d(\tau)\ . \tag{2.49}$$

To proceed, it is convenient to first recognize that $R_{a,cd}$, and thus θ_{ij}^{cd}, depend on the number of breakpoints occurring in the recombination event. However, under the small r assumption (2.13), it is sufficient to consider only a single breakpoint since the probability of more than one breakpoint is $O(\frac{1}{N^2})$ (see (2.15)). By noting that $1 = 1 - g_i^c + g_i^c$, we proceed by dividing the two summations in $\sum_{c=1}^{M}\sum_{d=1}^{M} \theta_{ij}^{cd}\check{z}_c(\tau)\check{z}_d(\tau)$ into 16 summations, corresponding to whether there are mutations at loci i and j in genotypes c and d. Specifically, these 16 summations correspond to the 16 possible allele-pairs in genotypes c and d, shown in the first and second columns of Table 2.1. We define the 'event' A_{ij}^{cd}, as the event that recombination of genotype c and d results in the locus-pair (i, j) both having mutant alleles. Similar to (2.44), we may thus decompose (2.49) as

$$
\begin{aligned}
\sum_{c=1}^{M}\sum_{d=1}^{M}\theta_{ij}^{cd}\check{z}_c(\tau)\check{z}_d(\tau) = {} & \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)(1-g_j^c)(1-g_i^d)(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)(1-g_j^c)(1-g_i^d)g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)(1-g_j^c)g_i^d(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)(1-g_j^c)g_i^dg_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)g_j^c(1-g_i^d)(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)g_j^c(1-g_i^d)g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)g_j^cg_i^d(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})(1-g_i^c)g_j^cg_i^dg_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^c(1-g_j^c)(1-g_i^d)(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^c(1-g_j^c)(1-g_i^d)g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^c(1-g_j^c)g_i^d(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^c(1-g_j^c)g_i^dg_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^cg_j^c(1-g_i^d)(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^cg_j^cg_i^d(1-g_j^d)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c}\sum_{d}\Pr(A_{ij}^{cd})g_i^cg_j^cg_i^dg_j^d\check{z}_c(\tau)\check{z}_d(\tau).
\end{aligned}
\tag{2.50}
$$

Table 2.1: Probabilities of recombination events in (2.51). The factor of $\frac{1}{2}$ arises because there is 50% chance of choosing genotypes c or d in the recombination process. The denominator is $L-1$ as there are only $L-1$ possible locations along the sequence length where a breakpoint (bpt.) can occur.

Genotype c locus-pair (i,j)	Genotype d locus-pair (i,j)	P_{1i}	P_{ij}	P_{jL}
00	00	0	0	0
00	01	0	0	0
00	10	0	0	0
00	11	$\frac{1}{2}\frac{i-1}{L-1}$	0	$\frac{1}{2}\frac{L-j}{L-1}$
01	00	0	0	0
01	01	0	0	0
01	10	0	$\frac{1}{2}\frac{j-i}{L-1}$	0
01	11	$\frac{1}{2}\frac{i-1}{L-1}$	$\frac{1}{2}\frac{j-i}{L-1}$	$\frac{1}{2}\frac{L-j}{L-1}$
10	00	0	0	0
10	01	0	$\frac{1}{2}\frac{j-i}{L-1}$	0
10	10	0	0	0
10	11	$\frac{1}{2}\frac{i-1}{L-1}$	$\frac{1}{2}\frac{j-i}{L-1}$	$\frac{1}{2}\frac{L-j}{L-1}$
11	00	$\frac{1}{2}\frac{i-1}{L-1}$	0	$\frac{1}{2}\frac{L-j}{L-1}$
11	01	$\frac{1}{2}\frac{i-1}{L-1}$	$\frac{1}{2}\frac{j-i}{L-1}$	$\frac{1}{2}\frac{L-j}{L-1}$
11	10	$\frac{1}{2}\frac{i-1}{L-1}$	$\frac{1}{2}\frac{j-i}{L-1}$	$\frac{1}{2}\frac{L-j}{L-1}$
11	11	$\frac{i-1}{L-1}$	$\frac{j-i}{L-1}$	$\frac{L-j}{L-1}$

Now using the Total Probability Theorem, we have

$$\Pr\left(A_{ij}^{cd}\right) = P_{1i} + P_{ij} + P_{jL} \tag{2.51}$$

where $P_{1i} = \Pr\left(A_{ij}^{cd} | 1 < \text{bpt.} < i\right) \times \Pr\left(1 < \text{bpt.} < i\right)$, $P_{ij} = \Pr\left(A_{ij}^{cd} | i < \text{bpt.} < j\right) \times \Pr\left(i < \text{bpt.} < j\right)$, $P_{jL} = \Pr\left(A_{ij}^{cd} | j < \text{bpt.} < L\right) \times \Pr\left(j < \text{bpt.} < L\right)$, bpt. stands for breakpoint., $\Pr\left(i < \text{bpt.} < j\right)$ is the probability that the breakpoint lies between loci i and j, with $i < j$, and $\Pr\left(A_{ij}^{cd} | i < \text{bpt.} < j\right)$ is the conditional probability that event A_{ij}^{cd} occurs. These probabilities are given in Table 2.1, from which we have

$$
\begin{aligned}
\sum_{c=1}^{M}\sum_{d=1}^{M}\theta_{ij}^{cd}\check{z}_c(\tau)\check{z}_d(\tau) = {} & \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)\left(1-g_i^c\right)\left(1-g_j^c\right)g_i^d g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c g_j^c\left(1-g_i^d\right)\left(1-g_j^d\right)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)\left(1-g_i^c\right)g_j^c g_i^d g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c g_j^c\left(1-g_i^d\right)g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)\left(1-g_i^c\right)g_j^c g_i^d\left(1-g_j^d\right)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c\left(1-g_j^c\right)\left(1-g_i^d\right)g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c\left(1-g_j^c\right)g_i^d g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c g_j^c g_i^d\left(1-g_j^d\right)\check{z}_c(\tau)\check{z}_d(\tau)+ \\
& \sum_{c=1}^{M}\sum_{d=1}^{M}\Pr\left(A_{ij}^{cd}\right)g_i^c g_j^c g_i^d g_j^d\check{z}_c(\tau)\check{z}_d(\tau)+ \\
= {} & 2\times\frac{L-j-i-1}{2(L-1)}\left(\check{x}_{ij}(\tau)-\check{x}_i(\tau)\check{x}_{ij}(\tau)-\check{x}_j(\tau)\check{x}_{ij}(\tau)+\check{x}_{ij}^2(\tau)\right)+ \\
& 2\times\frac{j-i}{2(L-1)}\left(\check{x}_i(\tau)\check{x}_j(\tau)-\check{x}_i(\tau)\check{x}_{ij}(\tau)-\check{x}_j(\tau)\check{x}_{ij}(\tau)+\check{x}_{ij}^2(\tau)\right)+ \\
& 2\times\frac{1}{2}\left(\check{x}_i(\tau)\check{x}_{ij}(\tau)-\check{x}_{ij}^2(\tau)\right)+\check{x}_{ij}^2(\tau)2\times\frac{1}{2}\left(\check{x}_j(\tau)\check{x}_{ij}(\tau)-\check{x}_{ij}^2(\tau)\right) \\
= {} & \check{x}_{ij}(\tau)-\frac{j-i}{L-1}\left(\check{x}_{ij}(\tau)-\check{x}_i(\tau)\check{x}_j(\tau)\right). \qquad (2.52)
\end{aligned}
$$

Substituting (2.52) together with (2.49) back into (2.48) we see that

$$\eta_e(\check{\mathbf{x}}(\tau)) = (j-i)\left(\check{x}_{ij}(\tau)-\check{x}_i(\tau)\check{x}_j(\tau)\right) \qquad \text{for} \quad L < e \leq R. \qquad (2.53)$$

Thus, from (2.47) and (2.53), by substituting terms and simplifying the expressions, we have

$$\eta_e(\check{\mathbf{x}}(\tau)) = \begin{cases} 0 & 1 \le e \le L \\ \\ (j-i)\big(\check{x}_{ij}(\tau) - \check{x}_i(\tau)\check{x}_j(\tau)\big) & L < e \le R, \end{cases} \tag{2.54}$$

where i and j are the subscript indices corresponding to the eth element of $\check{\mathbf{x}}(\tau)$.

Given the drift vector and diffusion matrix, we can directly apply [137, eq. 4.109] which gives the transition probability density over Δt generations, valid for small $\delta\tau\Delta t$ (equivalently large $\frac{N}{\Delta t}$), as

$$\phi(\check{\mathbf{x}}(\tau+\delta\tau\Delta t)|\check{\mathbf{x}}(\tau)) \approx \frac{\exp\left(-\frac{1}{4\delta\tau\Delta t}F_{diff}^{\mathrm{T}}\bar{C}(\check{\mathbf{x}}(\tau))^{-1}F_{diff}\right)}{(4\pi\delta\tau\Delta t)^{R/2}\sqrt{\det(\bar{C}(\check{\mathbf{x}}(\tau)))}} . \tag{2.55}$$

where,

$$F_{diff} = \check{\mathbf{x}}(\tau+\delta\tau\Delta t) - \check{\mathbf{x}}(\tau) - \bar{d}(\check{\mathbf{x}}(\tau))\delta\tau\Delta t \tag{2.56}$$

Thus, the transition probability for a single generation of the original discrete-time discrete-frequency WF process can (for large $\frac{N}{\Delta t}$) be approximated by

$$P(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)) \approx \phi(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)) \prod_{e=1}^{R} \mathrm{d}x_e(t_{k+1}) \tag{2.57}$$

where the dx_e represent small frequency differences accounting for the quantization of the continuous eth marginal allele frequency space at each time point.

The probability of observing a trajectory of mutant allele frequencies $(\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_K))$ conditioned on $\mathbf{x}(t_0)$ is then given by

$$\begin{aligned} P\left((\mathbf{x}(t_k))_{k=1}^{K} | \mathbf{x}(t_0) \right) &= \prod_{k=0}^{K-1} P(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)) \\ &\approx \prod_{k=0}^{K-1} \left[\frac{1}{\sqrt{\det C(\mathbf{x}(t_k))}} \left(\frac{N}{2\pi \Delta t_k} \right)^{R/2} \prod_{e=1}^{R} \mathrm{dx}_e(t_{k+1}) \right] \times \\ &\qquad \exp\left(-\frac{N}{2} S\left((\mathbf{x}(t_k))_{k=0}^{K} \right) \right) \end{aligned} \tag{2.58}$$

where

$$\begin{aligned} S\left((\mathbf{x}(t_k))_{k=0}^{K} \right) = \sum_{k=0}^{K-1} \frac{1}{\Delta t_k} \sum_{e=1}^{R} \sum_{f=1}^{R} \left[\mathrm{x}_e(t_{k+1}) - \mathrm{x}_e(t_k) - \mathrm{d}_e(\mathbf{x}(t_k)) \Delta t_k \right] \times \\ \left(\mathrm{C}^{-1}(\mathbf{x}(t_k)) \right)_{ef} \left[\mathrm{x}_f(t_{k+1}) - \mathrm{x}_f(t_k) - \mathrm{d}_f(\mathbf{x}(t_k)) \Delta t_k \right], \end{aligned}$$

which is the desired path integral representation. Note we have defined $\mathrm{d}_e(\mathbf{x}(t_k)) := \frac{\bar{\mathrm{d}}_e(\mathbf{x}(t_k))}{N}$ and $(\mathrm{C}(\mathbf{x}(t_k)))_{ef} := 2\left(\bar{\mathrm{C}}(\mathbf{x}(t_k)) \right)_{ef}$.

2.2.5 Marginal path likelihood (MPL) estimator with epistasis

The MPL parameter estimates are obtained by adopting a Bayesian approach. Specifically, we use the maximum a posteriori (MAP) criterion to find the most likely selection coefficients and epistasis terms given the measured single, double, triple and quadruple mutant frequencies at each sampling time point, along with knowledge of evolutionary parameters N, μ and r. For the purpose of developing an efficient inference approach, we assume that the observed frequencies are equal to the true frequencies in the population.

The MPL estimate of the selection coefficients and epistasis terms can thus be obtained by solving

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; \mu, r, N, \left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right) P^{\text{prior}}(\mathbf{s}), \tag{2.59}$$

where $P^{\text{prior}}(\mathbf{s})$ is the assumed (conjugate) prior

$$P^{\text{prior}}(\mathbf{s}) = \frac{1}{(2\pi\sigma^2)^{R/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\right),$$

with mean zero and variance $\sigma^2 > 0$, and the likelihood of the selection coefficients and epistasis terms, $\mathbf{s}$, given the observed data can be expressed as

$$\begin{aligned}\mathfrak{L}\left(\mathbf{s}; N, r, \mu, (\mathbf{x}(t_k))_{k=0}^{K}\right) &= P\left((\mathbf{x}(t_k))_{k=1}^{K} |\mathbf{x}(t_0), N, r, \mu, \mathbf{s}\right) \\ &= \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k), N, r, \mu, \mathbf{s}\right).\end{aligned} \tag{2.60}$$

While it is challenging to calculate the likelihood (2.60) exactly, the task is simplified by using the path integral approach outlined in the previous section with some modifications to account for time-samples drawn from non-unit time intervals, $\Delta t_k = t_{k+1} - t_k$. The likelihood of the selection coefficients and epistasis terms (2.60) could be expanded as

$$\begin{aligned}\mathfrak{L}\left(\mathbf{s}; \mu, r, N, \left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right) = &\left(\prod_{k=0}^{T-1} \frac{1}{\sqrt{\det C(\mathbf{x}(t_k))}} \left(\frac{N}{2\pi\Delta t_k}\right)^{R/2} \prod_{i=1}^{R} \mathrm{d}x_i(t_{k+1})\right) \times \\ &\prod_{k=0}^{K-1} \exp\left(-\frac{N}{2} S\left(\left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right)\right)\end{aligned} \tag{2.61}$$

where

$$P^{\text{prior}}(\mathbf{s}) = \frac{1}{(2\pi\sigma^2)^{R/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\right) \tag{2.62}$$

is the assumed prior with $\sigma^2 \in \mathbb{R}$. For convenience, we work with the natural logarithm of the above, i.e.,

$$\ln\left(\mathfrak{L}\left(\mathbf{s};\mu,r,N,\left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right)\right)+\ln\left(P^{\text{prior}}(\mathbf{s})\right)=\ln c_1-\frac{N}{2}\sum_{k=0}^{K-1}S\left(\left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right)+\ln c_2-\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}, \tag{2.63}$$

where c_1 and c_2 represent terms that are independent of $\mathbf{s}$. Next, we take the vector partial derivative with respect to $\mathbf{s}$ and equate it to zero to find the MAP estimate of $\mathbf{s}$. This gives

$$\begin{aligned}\mathbf{0}&=\frac{\partial}{\partial\mathbf{s}}\ln c_1-\frac{\partial}{\partial\mathbf{s}}\frac{N}{2}\sum_{k=0}^{K-1}S\left(\left(\mathbf{x}(t_k)\right)_{k=0}^{K}\right)+\frac{\partial}{\partial\mathbf{s}}\ln c_2-\frac{\partial}{\partial\mathbf{s}}\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\\&=\sum_{k=0}^{K-1}C(\mathbf{x}(t_k))\left[C(\mathbf{x}(t_k))\right]^{-1}[\mathbf{x}(t_{k+1})-\mathbf{x}(t_k)-\\&\qquad\Delta t_k C(\mathbf{x}(t_k))\mathbf{s}-\mu\Delta t_k\mathbf{v}(\mathbf{x}(t_k))-r\Delta t_k\eta(\mathbf{x}(t_k))+\gamma\ \mathbf{s},\end{aligned} \tag{2.64}$$

where $\gamma=1/N\sigma^2$. Solving the above yields the desired MPL estimator, i.e.,

$$\hat{\mathbf{s}}=\left[\sum_{k=0}^{K-1}\Delta t_k C(\mathbf{x}(t_k))+\gamma I\right]^{-1}\left[\mathbf{x}(t_K)-\mathbf{x}(t_0)-\mu\sum_{k=0}^{K-1}\Delta t_k\mathbf{v}(\mathbf{x}(t_k))-r\sum_{k=0}^{K-1}\Delta t_k\eta(\mathbf{x}(t_k))\right]. \tag{2.65}$$

We note that, in practice, it is not required to know the exact values of N or σ^2. Rather what is important is that their product γ has an appropriate strength, and this can be treated as a regularization parameter.

The MPL estimator (2.65) has an intuitive interpretation. It computes the observed change in the single and double mutant allele frequencies between the final and the initial time points, adjusts it by accounting for the (inward and outward) mutational flows of single and double mutant frequencies over time, and then applies a correction to the double mutant frequencies to account for the effect of recombination. Finally, it accounts for linkage effects through the inverse of the (regularized) integrated covariance matrix.

As shown in (2.65), significant changes in mutant frequencies – that is, ones that are substantially larger than those expected due to finite population size alone – that cannot readily be explained by mutation, recombination, or the effects of background mutations provide evidence of selection or epistatic interactions. For example, mutant alleles that are separated by a long distance on the genome and which remain strongly linked despite recombination would be evidence of a positive epistatic interaction.

2.2.6 Combining multiple independent observations

The inference framework may be applied to incorporate observations of mutant allele frequencies from multiple independent replicates. These replicates may be parallel evolutionary experiments or time-series data from distinct studies. Each replicate represents a unique evolutionary path that may have different initial conditions and/or sampling parameters, independent from the other replicates. Here we give the specific generalization for the bi-allelic model with symmetric mutation probabilities, as considered in 2.2. Further extension to multi-allele and asymmetric mutation probability models is straightforward.

For a scenario with Q replicates, the MAP estimate of the selection coefficients is the solution to

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; \mu, N, (\mathbf{x}^1(t_k^1))_{k=0}^{K_1}, \cdots, (\mathbf{x}^Q(t_k^Q))_{k=0}^{K_Q}\right) P^{\text{prior}}(\mathbf{s}), \tag{2.66}$$

where $\mathbf{x}^q(t_k^q) = (x_1^q(t_k^q), \cdots, x_L^q(t_k^q))$ is the observed mutant allele frequencies at generation t_k^q of replicate q. The likelihood function admits

$$\mathfrak{L}\left(\mathbf{s}; \mu, N, (\mathbf{x}^1(t_k^1))_{k=0}^{K_1}, \cdots, (\mathbf{x}^Q(t_k^Q))_{k=0}^{K_Q}\right) = \prod_{q=1}^{Q} \prod_{k=0}^{K_q-1} P\left(\mathbf{x}^q(t_{k+1}^q) | \mathbf{x}^q(t_k^q), N, \mu, \mathbf{s}\right) \tag{2.67}$$

and, as before, the prior is

$$P^{\text{prior}}(\mathbf{s}) = \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\right) . \tag{2.68}$$

Using (2.58), we obtain the path integral approximation to the likelihood function

$$\begin{aligned}&\mathfrak{L}\left(\mathbf{s}; \mu, N, (\mathbf{x}^1(t_k^1))_{k=0}^{K_1}, \cdots, (\mathbf{x}^Q(t_k^Q))_{k=0}^{K_Q}\right) \approx \\ &\prod_{q=1}^{Q} \left(\prod_{k=0}^{K_q-1} \frac{1}{\sqrt{\det C(\mathbf{x}^q(t_k^q))}} \left(\frac{N}{2\pi\Delta t_k^q}\right)^{L/2} \prod_{i=1}^{L} \mathrm{d}x_i^q(t_{k+1}^q) \right) \exp\left(-\frac{N}{2} S\left((\mathbf{x}^q(t_k^q))_{k=0}^{K_q}\right)\right) .\end{aligned} \tag{2.69}$$

Substituting this approximation in (2.66), we get the MPL estimator of multiple independent replicates as,

$$\begin{aligned}\hat{\mathbf{s}} = &\left[\sum_{q=1}^{Q} \sum_{k=0}^{K_q-1} \Delta t_k^q C(\mathbf{x}^q(t_k^q)) + \gamma I\right]^{-1} \times \\ &\left(\sum_{q=1}^{Q}\left[\mathbf{x}^q(t_{K_q}^q) - \mathbf{x}^q(t_0^q) - \mu \sum_{k=0}^{K_q-1} \Delta t_k^q \mathbf{v}(\mathbf{x}^q(t_k^q)) - r \sum_{k=0}^{K_q-1} \Delta t_k^q \eta(\mathbf{x}^q(t_k^q))\right]\right) ,\end{aligned} \tag{2.70}$$

where $C(\mathbf{x}^q(t_k^q))$ is the covariance matrix of the mutant allele frequencies at generation t_k^q of the qth replicate, $\gamma = 1/N\sigma^2$ as before, and $\Delta t_k^q = t_{k+1}^q - t_k^q$.

2.2.7 Asymmetrical mutation probabilities

So far we have assumed the forward and backward mutation probabilities are equal. The MPL framework can easily accommodate asymmetrical mutation probabilities as was also shown in [155] for the additive fitness model case. Here, we derive the expression of the drift vector of the allele-level diffusion process for a fitness model with pairwise epistasis terms. The diffusion matrix, being independent of the mutation probability, remains unchanged.

We begin by defining $\mu_{01,i}$ and $\mu_{10,i}$ as the mutation probabilities, at locus i, of the WT allele mutating to mutant allele and the mutant allele mutating to the WT allele respectively. Similar to (2.13), as $N \to \infty$

$$\mu_{01,i} = \frac{\bar{\mu}_{01,i}}{N} + O\Big(\frac{1}{N^2}\Big), \quad \mu_{10,i} = \frac{\bar{\mu}_{10,i}}{N} + O\Big(\frac{1}{N^2}\Big), \tag{2.71}$$

and consequently

$$\mu_{ab} = \frac{\bar{\mu}_{ab}}{N} + O\Big(\frac{1}{N^2}\Big), \tag{2.72}$$

where $\bar{\mu}_{01,i}$, $\bar{\mu}_{10,i}$, and $\bar{\mu}_{ab}$ are constants independent of N.

The ath entry of the drift vector $\bar{d}(\check{\mathbf{z}}(\tau))$ characterizing the genotype-level diffusion process in the case of equal forward and backward mutation probabilities was given by (2.18). In the scenario with locus specific unequal forward and backward mutation probabilities, the ath entry of the drift vector is given as

$$\bar{d}_a((\tau)) = \check{z}_a(\tau)\left(\bar{h}_a - \sum_{b=1}^{M} \bar{h}_b \check{z}_b(\tau)\right) + \left(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba}\check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab}\check{z}_a(\tau)\right) - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\left((\tau)\right)\big). \tag{2.73}$$

Following the same steps as before, the eth element of the allele-level drift vector is defined as

$$
\begin{aligned}
\bar{\mathrm{d}}_e(\check{\mathbf{x}}(\tau)) &:= \sum_{a=1}^{M} \mathrm{u}_e^a \bar{d}_a(\check{\mathbf{z}}(\tau)) \\
&= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\check{z}_a(\tau) \Bigg(\bar{h}_a - \sum_{b=1}^{M} \bar{h}_b \check{z}_b(\tau) \Bigg) + \\
&\qquad \Bigg(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} \check{z}_a(\tau) \Bigg) - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\left((\tau)\right) \big) \\
&= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\check{z}_a(\tau)\big(1-\check{z}_a(\tau)\big)\bar{h}_a - \check{z}_a(\tau) \sum_{b=1,b\neq a}^{M} \bar{h}_b \check{z}_b(\tau) + \\
&\qquad \Bigg(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} \check{z}_a(\tau) \Bigg) - \bar{r}(L-1)\big(\check{z}_a(\tau) - \psi_a\big((\tau)\big) \big) \Bigg) \\
&= \check{\mathrm{x}}_e(\tau)\left(1-\check{\mathrm{x}}_e(\tau)\right)\bar{\mathrm{s}}_e + \sum_{f\neq e} \bar{\mathrm{C}}_{ef}(\check{\mathbf{x}}(\tau))\bar{\mathrm{s}}_f + \Omega_e + \bar{r}\ \eta_e(\check{\mathbf{x}}(\tau)),
\end{aligned}
\tag{2.74}
$$

where $\eta_e(\check{\mathbf{x}}(\tau))$ is given by (2.54) and Ω_e is the mutation term in the asymmetrical mutation probabilities scenario. As in case of symmetrical mutation probabilites, we first consider the case $1 \leq e \leq L$, for which

$$
\begin{aligned}
\Omega_e &= \sum_{a=1}^{M} \mathrm{u}_e^a \Bigg(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} \check{z}_a(\tau) \Bigg) \\
&= \sum_{a=1}^{M} g_i^a \Bigg(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} \check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} \check{z}_a(\tau) \Bigg) \\
&= \sum_{a=1}^{M} \Bigg(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} g_i^a (1-g_i^b) \check{z}_b(\tau) + \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba} g_i^a g_i^b \check{z}_b(\tau) - \\
&\qquad \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} g_i^a (1-g_i^b) \check{z}_a(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab} g_i^a g_i^b \check{z}_a(\tau) \Bigg) \\
&= \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a (1-g_i^b) \left(\bar{\mu}_{ba} \check{z}_b(\tau) - \bar{\mu}_{ab} \check{z}_a(\tau) \right) + \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_i^b \left(\bar{\mu}_{ba} \check{z}_b(\tau) - \bar{\mu}_{ab} \check{z}_a(\tau) \right),
\end{aligned}
\tag{2.75}
$$

where the second last line above follows from noting that the first summation on the right side of the third last line can be decomposed into two parts. The first where genotypes a and b differ only at locus i, and hence mutation of genotype b to genotype a changes the mutant allele frequency at locus i. The second is where the two genotypes differ from each other at a locus other than i and hence a mutation from genotype b to a does not effect the mutant allele frequency at locus i. Similarly, the second summation in the second last line can also be split into two parts. Now, note that

$$\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_i^b \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right) = 0,$$

as this quantity represents the mutation of those genotypes b to genotype a where both a and b have the mutant allele at locus i, while

$$\sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a (1 - g_i^b) \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right) = \bar{\mu}_{01,i}\left(1 - \check{x}_i(\tau)\right) - \bar{\mu}_{10,i}\check{x}_i(\tau), \qquad (2.76)$$

which represents the flow of mutational probabilities between the WT and the mutation allele, i.e., the mutational flux. Here $\mu_{\alpha\beta,i}$ is the per generation probability of allele α mutating to allele β at locus i.

Now consider the case when $L < e \leq R$ where we have

$$\begin{aligned}
\Omega_e &= \sum_{a=1}^{M} \mathrm{u}_e^a \left(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba}\check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab}\check{z}_a(\tau) \right) = \sum_{a=1}^{M} g_i^a g_j^a \left(\sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ba}\check{z}_b(\tau) - \sum_{b=1,d_{ab}=1}^{M} \bar{\mu}_{ab}\check{z}_a(\tau) \right) \\
&= \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1 - g_i^b)(1 - g_j^b) \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right) + \\
&\quad \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1 - g_i^b) g_j^b \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right) + \\
&\quad \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b (1 - g_j^b) \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right) + \sum_{a=1}^{M} \sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b g_j^b \left(\bar{\mu}_{ba}\check{z}_b(\tau) - \bar{\mu}_{ab}\check{z}_a(\tau)\right).
\end{aligned} \qquad (2.77)$$

Here, the summations on the right side of the second last line above represent the net mutational flow to genotypes that contain alleles $(1,1)$ at locus-pair (i,j), from those genotypes that have alleles $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ respectively at locus-pair (i,j). We note that

$$\begin{gathered}
\sum_{a=1}^{M}\sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b)(1-g_j^b)\left(\bar{\mu}_{ba}\check{z}_b(\tau)-\bar{\mu}_{ab}\check{z}_a(\tau)\right)=0\\
\sum_{a=1}^{M}\sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a (1-g_i^b)g_j^b\left(\bar{\mu}_{ba}\check{z}_b(\tau)-\bar{\mu}_{ab}\check{z}_a(\tau)\right)=\bar{\mu}_{01,i}\check{x}_{ij}^{01}(\tau)-\bar{\mu}_{10,i}\check{x}_{ij}^{11}(\tau)\\
\sum_{a=1}^{M}\sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b(1-g_j^b)\left(\bar{\mu}_{ba}\check{z}_b(\tau)-\bar{\mu}_{ab}\check{z}_a(\tau)\right)=\bar{\mu}_{10,j}\check{x}_{ij}^{01}(\tau)-\bar{\mu}_{10,j}\check{x}_{ij}^{11}(\tau)\\
\sum_{a=1}^{M}\sum_{b=1,d_{ab}=1}^{M} g_i^a g_j^a g_i^b g_j^b\left(\bar{\mu}_{ba}\check{z}_b(\tau)-\bar{\mu}_{ab}\check{z}_a(\tau)\right)=0.
\end{gathered}$$

Substituting the above in (2.77) we have

$$\begin{aligned}
\Omega_e &= \bar{\mu}_{01,i}\check{x}_{ij}^{01}(\tau)-\bar{\mu}_{10,i}\check{x}_{ij}^{11}(\tau)+\bar{\mu}_{10,j}\check{x}_{ij}^{01}(\tau)-\bar{\mu}_{10,j}\check{x}_{ij}^{11}(\tau)\\
&= \bar{\mu}_{01,i}\left(\check{x}_j^1(\tau)-\check{x}_{ij}^{11}(\tau)\right)-\bar{\mu}_{10,i}\check{x}_{ij}^{11}(\tau)+\bar{\mu}_{01,j}\left(\check{x}_i^1(\tau)-\check{x}_{ij}^{11}(\tau)\right)-\bar{\mu}_{10,j}\check{x}_{ij}^{11}(\tau). \qquad (2.78)
\end{aligned}$$

Dropping the superscripts from $\check{x}_i(\tau)$ and $\check{x}_{ij}(\tau)$ as before, and from (2.76) and (2.78), we have

$$\Omega_e = \bar{\mu}_{01,i}\mathrm{v}_e^{'}(\check{\mathbf{x}}(\tau))-\bar{\mu}_{10,i}\mathrm{v}_e^{''}(\check{\mathbf{x}}(\tau))+\bar{\mu}_{01,j}\mathrm{w}_e^{'}(\check{\mathbf{x}}(\tau))-\bar{\mu}_{10,j}\mathrm{w}_e^{''}(\check{\mathbf{x}}(\tau)) \qquad (2.79)$$

$$\mathrm{v}_e^{'}(\check{\mathbf{x}}(\tau)) = \begin{cases} 1 - \check{x}_i(\tau) & 1 \le e \le L \\ \\ \check{x}_j(\tau) - \check{x}_{ij}(\tau) & L < e \le R \end{cases}$$

$$\mathrm{v}_e^{''}(\check{\mathbf{x}}(\tau)) = \begin{cases} \check{x}_i(\tau) & 1 \le e \le L \\ \\ \check{x}_{ij}(\tau) & L < e \le R \end{cases}$$

$$\mathrm{w}_e^{'}(\check{\mathbf{x}}(\tau)) = \begin{cases} 0 & 1 \le e \le L \\ \\ \check{x}_i(\tau) - \check{x}_{ij}(\tau) & L < e \le R, \end{cases}$$

and

$$\mathrm{w}_e^{''}(\check{\mathbf{x}}(\tau)) = \begin{cases} 0 & 1 \le e \le L \\ \\ \check{x}_{ij}(\tau) & L < e \le R. \end{cases}$$

Here (2.79) is the mutational term in the asymmetrical mutation probabilities case.

Following similar steps as before, one can derive the MPL estimate with asymmetrical mutation probabilities as

$$\begin{aligned} \hat{\mathrm{s}}_e = \sum_{f=1}^{R} \left[\sum_{k=0}^{K-1} \Delta t_k \mathrm{C}(\mathbf{x}(t_k)) + \gamma I \right]_{ef}^{-1} \times \\ \Bigg[\mathrm{x}_f(t_K) - \mathrm{x}_f(t_0) - \mu_{01,i} \sum_{k=0}^{K-1} \Delta t_k \mathrm{v}_f^{'}(\mathbf{x}(t_k)) + \mu_{10,i} \sum_{k=0}^{K-1} \Delta t_k \mathrm{v}_f^{''}(\mathbf{x}(t_k)) - \\ \mu_{01,j} \sum_{k=0}^{K-1} \Delta t_k \mathrm{w}_f^{'}(\mathbf{x}(t_k)) + \mu_{10,j} \sum_{k=0}^{K-1} \Delta t_k \mathrm{w}_f^{''}(\mathbf{x}(t_k)) - r \sum_{k=0}^{K-1} \Delta t_k \eta_f(\mathbf{x}(t_k)) \Bigg]. \end{aligned} \quad (2.80)$$

2.2.8 Equivalence with the genotype estimate

The MAP estimate of the allele selection coefficients and epistasis terms can also be obtained from the genotype path integral (2.20) by solving

$$\hat{\mathbf{s}} =_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; \mu, r, N, \left(\mathbf{z}(t_k)\right)_{k=0}^{K}\right) P^{\text{prior}}(\mathbf{s}) \ . \tag{2.81}$$

The prior probability, with $\sigma^2 \in \mathbb{R}$, is the same as in (2.68) and given below for convenience

$$P^{\text{prior}}(\mathbf{s}) = \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{s}^{\text{T}}\mathbf{s}\right) \ .$$

From (2.20), the approximate genotype path-likelihood is given by

$$\begin{aligned}\mathfrak{L}\left(\mathbf{s}; \mu, r, N, \left(\mathbf{z}(t_k)\right)_{k=0}^{K}\right) = &\left(\prod_{k=0}^{T-1} \frac{1}{\sqrt{\det C(\mathbf{z}(t_k))}} \left(\frac{N}{2\pi\Delta t_k}\right)^{R/2} \prod_{a=1}^{M} \mathrm{d}z_a(t_{k+1})\right) \times \\ &\prod_{k=0}^{K-1} \exp\left(-\frac{N}{2} S\left(\left(\mathbf{z}(t_k)\right)_{k=0}^{K}\right)\right) \ ,\end{aligned}$$

with

$$\begin{aligned}S\left(\left(\mathbf{z}(t_k)\right)_{k=0}^{K}\right) = \sum_{k=0}^{K-1} \frac{1}{\Delta t_k} \sum_{a=1}^{M}\sum_{b=1}^{M} &\left[z_a(t_{k+1}) - z_a(t_k) - d_a(\mathbf{z}(t_k))\Delta t_k\right] \times \\ &\left(C^{-1}(\mathbf{z}(t_k))\right)_{ab} \left[z_b(t_{k+1}) - z_b(t_k) - d_b(\mathbf{z}(t_k))\Delta t_k\right],\end{aligned}$$

where $\Delta t_k = t_{k+1} - t_k$ and

$$\begin{aligned}d_a(\mathbf{z}(t_k)) = z_a(t_k)\left(h_a - \sum_{b=1}^{M} h_b z_b(t_k)\right) + \mu\left(\sum_{b=1, d_{ab}=1}^{M} z_b(t_k) - \sum_{b=1, d_{ab}=1}^{M} z_a(t_k)\right) - \\ r(L-1)\big(z_a(t_k) - \psi_a\left(\mathbf{z}(t_k)\right)\big),\end{aligned}$$

with

$$C_{ab}(\mathbf{z}(t_k)) = \begin{cases} z_a(t_k)(1 - z_a(t_k)) & a = b \\ -z_a(t_k)z_b(t_k) & a \neq b. \end{cases}$$

Maximizing the natural logarithm of (eq.2.81) and taking the partial derivative of the vector gives

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{s}} \ln c_1 - \frac{\partial}{\partial \mathbf{s}} \frac{N}{2} \sum_{k=0}^{K-1} S\left(\left(\mathbf{z}(t_k) \right)_{k=0}^{K} \right) + \frac{\partial}{\partial \mathbf{s}} \ln c_2 - \frac{\partial}{\partial \mathbf{s}} \frac{1}{2\sigma^2} \mathbf{s}^{\mathrm{T}} \mathbf{s}\ . \tag{2.82}$$

We note that h_a can be expressed as

$$h_a = \sum_{e=1}^{R} \mathrm{u}_e^a \mathrm{s}_e,$$

and from (2.21) and (2.25), that the single and double mutant allele frequencies can be expressed

$$\mathrm{x}_e = \sum_{e=1}^{R} u_e^a z_a(t). \tag{2.83}$$

Substituting these transformation in (2.82) and using the results in (2.33), (2.36), (2.41), and (2.54), we obtain

$$\hat{\mathbf{s}} = \left[\sum_{k=0}^{K-1} \Delta t_k C(\mathbf{x}(t_k)) + \gamma I \right]^{-1} \left[\mathbf{x}(t_K) - \mathbf{x}(t_0) - \mu \sum_{k=0}^{K-1} \Delta t_k \mathbf{v}(\mathbf{x}(t_k)) - r \sum_{k=0}^{K-1} \Delta t_k \eta(\mathbf{x}(t_k)) \right]\ , \tag{2.84}$$

which is the same estimator as in (2.65) obtained from allele-level path integral.

2.2.9 Robustness

Numerical issues may arise in computing the estimate in (2.80) in scenarios with severe data limitations (low number of samples, large time between samples). These can be addressed by assuming the allele frequency trajectories are piecewise continuous and the covariance matrix, $C(\mathbf{x}(t_k))$, is also a piecewise continuous function. This allows to replace the summation over time in (2.80) with integration, which can then be computed analytically. To be more specifically, the diagonal terms of the integrated covariance matrix are

$$\frac{(3-2\,\mathrm{x}_e(t_{k+1}))\,(\mathrm{x}_e(t_k)+\mathrm{x}_e(t_{k+1}))}{6}-\frac{\mathrm{x}_e^2(t_k)}{3}, \tag{2.85}$$

and the off-diagonal terms of the integrated covariance matrix are

$$\frac{\mathrm{x}_{ef}(t_k)+\mathrm{x}_{ef}(t_{k+1})}{2}-\left(\frac{\mathrm{x}_e(t_k)\mathrm{x}_f(t_k)}{3}+\frac{\mathrm{x}_e(t_{k+1})\mathrm{x}_f(t_{k+1})}{3}+\frac{\mathrm{x}_e(t_k)\mathrm{x}_f(t_{k+1})}{6}+\frac{\mathrm{x}_e(t_{k+1})\mathrm{x}_f(t_k)}{6}\right), \tag{2.86}$$

where the same mapping holds for indices e and f in (2.85) and (2.86). The mutation terms are now given as

$$\mathrm{v}^{'}_f(.)=\begin{cases}1-\frac{x_i(t_k)+x_i(t_{k+1})}{2} & 1\leq f\leq L\\ \frac{x_j(t_k)+x_j(t_{k+1})}{2}-\frac{x_{ij}(t_k)+x_{ij}(t_{k+1})}{2} & L<f\leq R\end{cases}$$

$$\mathrm{v}^{''}_f(.)=\begin{cases}\frac{x_i(t_k)+x_i(t_{k+1})}{2} & 1\leq f\leq L\\ \frac{x_{ij}(t_k)+x_{ij}(t_{k+1})}{2} & L<f\leq R\end{cases}$$

$$\mathrm{w}^{'}_f(.)=\begin{cases}0 & 1\leq f\leq L\\ \frac{x_i(t_k)+x_i(t_{k+1})}{2}-\frac{x_{ij}(t_k)+x_{ij}(t_{k+1})}{2} & L<f\leq R,\end{cases}$$

and

$$\mathrm{w}^{''}_f(.)=\begin{cases}0 & 1\leq f\leq L\\ \frac{x_{ij}(t_k)+x_{ij}(t_{k+1})}{2} & L<f\leq R.\end{cases}$$

While the recombination term $\eta_f(.)$ is 0 for $1\leq e\leq L$ and given for $L<e\leq R$ by

$$(i-j)\left(\frac{x_{ij}(t_k)+x_{ij}(t_{k+1})}{2}-\left(\frac{x_i(t_k)x_j(t_k)}{3}+\frac{x_i(t_{k+1})x_j(t_{k+1})}{3}+\frac{x_i(t_k)x_j(t_{k+1})}{6}+\frac{x_i(t_{k+1})x_j(t_k)}{6}\right)\right) \tag{2.87}$$

2.2.10 Simulation setup

We generated evolutionary histories by running WF simulations consisting of a population of $N = 1000$ bi-allelic sequences evolving for T generations. We then randomly sampled n_s sequences every Δt generations, and used these sampled trajectories for inference of fitness parameters.

In simulations where it was required to control data variability in a population, we specified the number and the frequencies of the unique genotypes in the initial population (also known as founder sequences), and disallowed mutations and recombination. We refer to the all-zero genotype as the WT genotype. In simulations where the initial population contained more than one unique genotype, one of these was always the WT while the others were chosen from the set of remaining $2^L - 1$ possible genotypes at random, without replacement. All simulation results were computed over 1000 Monte Carlo runs. Unless stated otherwise, the initial frequency of each non-WT genotype was set to 5% of the population size, the sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, $T = 100$ generation were used for inference, and the regularization parameter, γ, was set to one.

2.2.11 Data Availability

Simulation data and a MATLAB (version R2017a) implementation of MPL used for producing results in the paper are freely available at `https://github.com/mssohail/epistasis-inference`.

2.3 Results

2.3.1 Accurate estimation of epistasis and selection coefficients

We first analyzed the performance of MPL on a two-locus bi-allelic system. We ran extensive simulations varying the selection strength, the composition of the initial population and different types of epistasis. The types of epistatic interactions we considered include positive epistasis, where the double mutant has a fitness higher than the sum of the individual fitness effects of each mutant allele; negative epistasis, where the fitness of the double mutant is lower than the sum of the individual fitness effects of each mutant allele; and sign epistasis, where the direction and the magnitude of the fitness effect of epistasis is opposite to and larger than the sum of the individual fitness effect of the two mutant alleles.

We found that MPL is typically able to accurately infer underlying fitness parameters. In the simulation shown in Figure 2.1, the initial population consisted of only the wild-type (WT) genotype.

MPL estimates of selection coefficients were accurate in each simulated scenario. Estimates of the epistasis terms were better in scenarios where both the selection coefficients were beneficial (Figure 2.1A) compared with the scenarios where both were deleterious (Figure 2.1B), regardless of the type of epistasis. This is because double mutants tend to appear very rarely in cases where both single mutants were less fit than WT, as the single mutants are rapidly purged from the population. In such cases (Figure 2.1B), the double mutant genotype never exceeded 4% of the population in our simulations.

A similar situation occurred in the positive sign epistasis scenario (Figure 2.1B *bottom* panel). Thus, data variability constrains the accuracy of the epistasis estimates, which is also reflected in the uncertainty of the inferred parameters (Figure 2.2).

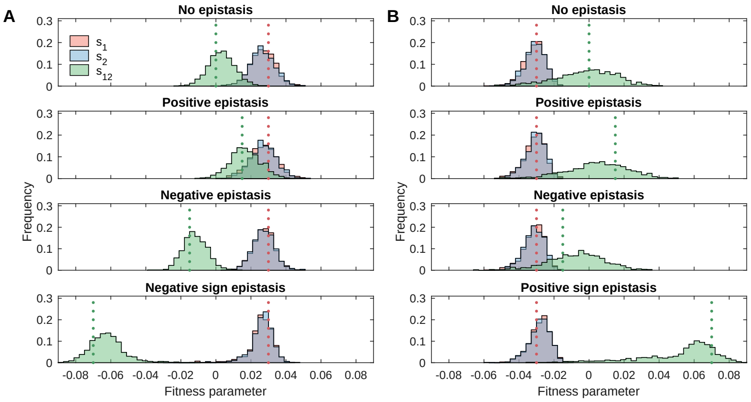

Figure 2.1: MPL can accurately infer selection coefficients and pairwise epistasis terms. (A) shows distribution of inferred selection coefficients and pairwise epistasis terms for various forms of epistasis when both selection coefficients are positive, while (B) shows the same for the case when both selection coefficients are negative. The results were obtained for a two-locus system with selection coefficients s_1, s_2 corresponding to the mutant alleles at locus 1 and 2 respectively, pairwise epistasis term s_{12}, per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$, and the initial population consisted of only the WT genotype (00). The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. The dashed lines represent the true selection coefficients (s_1 and s_2) and epistasis term (s_{12}). In these simulations, $s_1 = s_2$, hence the histograms of the estimates of the two have a significant overlap shown in grey color.

We further tested the ability of MPL to infer selection coefficients and epistasis terms under varying degrees of data variability by changing the composition of genotypes in the initial population. We found that the inference of these fitness parameters was quite accurate when all four genotypes appeared at high frequencies in the population, even when both single mutations were deleterious (*top left* panel of Figure 2.3). When some of the

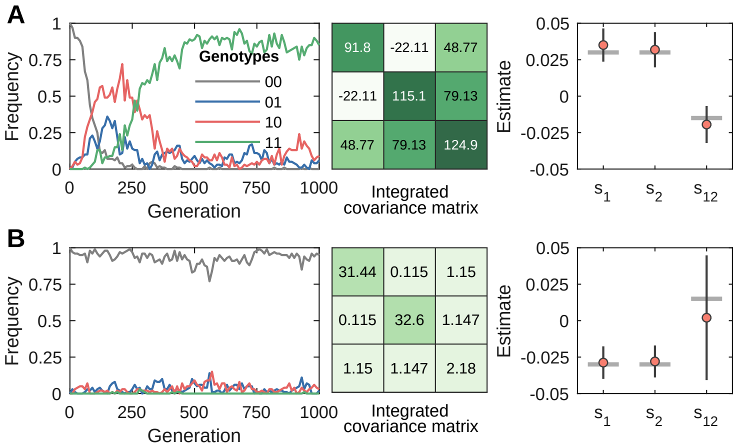

Figure 2.2: Higher data variability leads to more accurate inference. (A) shows a sample run (*left* panel) of a two-locus system (negative epistasis scenario) where all genotypes are well represented in the data, as indicated by the magnitude of the diagonal entries of the integrated covariance matrix (*center* panel). This leads to accurate estimation of the epistasis term and the selection coefficients (*right* panel). The vertical bars in the *right* panel indicate the 95% confidence intervals while the horizontal bars indicate the true selection coefficients and epistasis terms. (B) shows a sample run (*left* panel) of a two-locus system (positive epistasis scenario) where the double mutant genotype has limited variation, as indicated by the magnitude of the bottom right entry of the integrated covariance matrix (*center* panel). This leads to low accuracy in the estimate of the epistasis term. The selection coefficient estimates are still accurate because the single mutant genotypes, although present at low frequencies, are well represented in the data as indicated by the first two entries of the diagonal of the integrated covariance matrix (*center* panel). The results were obtained for a two-locus system with selection coefficients s_1, s_2 corresponding to the mutant alleles at locus 1 and 2 respectively, pairwise epistasis term s_{12}, per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$ and the initial population consisted of only the WT genotype. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference.

mutant genotypes are never present in the population, however, not all fitness parameters can be accurately inferred.

Based on patterns of variation in the time-series data, the estimated fitness parameters can be naturally classified into one of three categories: accessible, partially accessible,

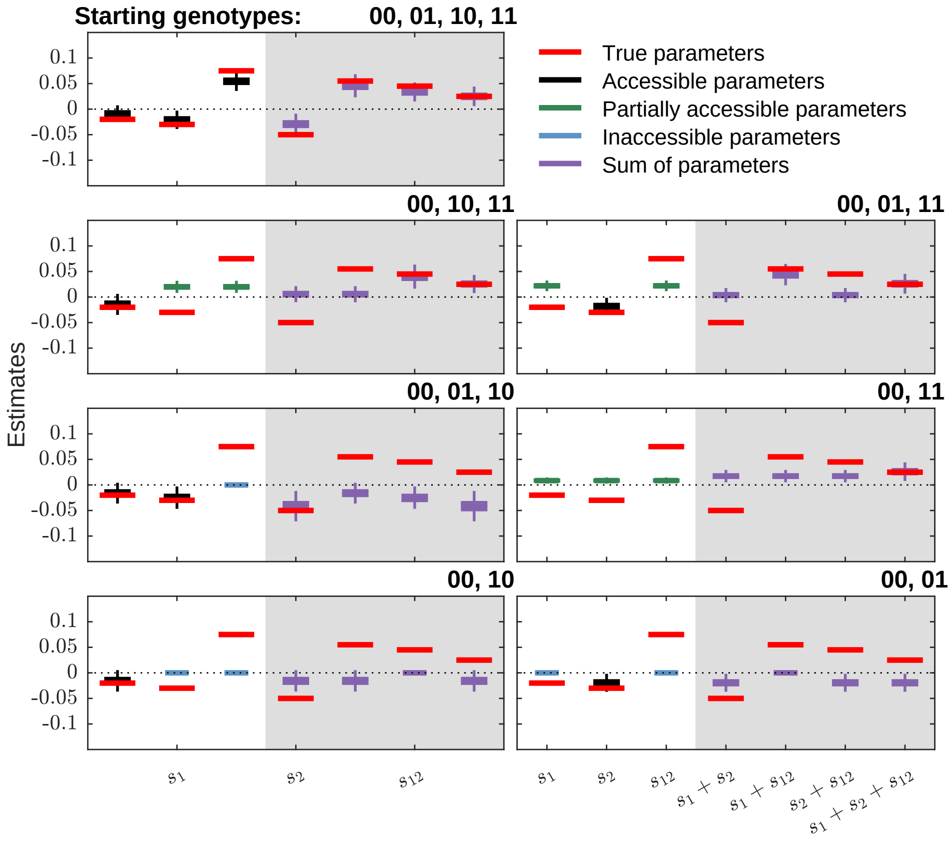

Figure 2.3: MPL can accurately estimate individual fitness parameters (selection coefficients and epistasis terms) and/or their sums depending on the variation present in the population. The results are for a two-locus system with positive sign epistasis (selection coefficients s_1, s_2 and pairwise epistasis term s_{12}). The boxplots of inferred selection coefficients and epistasis terms are shown on white background in each panel, while those of their sums are shown on grey background. The red bars indicate the true values of the respective terms. The boxplots show the standard data summary (minimum, first quartile, median, third quartile, maximum). In order to control data variability, both the per locus mutation probability and the per locus recombination probability were set to zero. The initial population contained the genotypes indicated above each panel, and the frequency of each non-WT genotype in the initial population was set to 10% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 150$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference.

or inaccessible, by examining the structure of the integrated covariance matrix used as part of the MPL estimator. Accessible fitness parameters are ones that could be independently estimated in principle (vice-versa for the inaccessible parameters), whereas partially accessible fitness parameters can only be estimated as part of a sum. Specifically, this is done by reducing the integrated covariance matrix to its reduced row-echelon form and checking the linear dependencies of its rows. The fitness parameters whose corresponding rows of the integrated covariance matrix are linearly independent are denoted as accessible. These can be estimated meaningfully. The fitness parameters corresponding to linearly dependent rows are classed as partially accessible. While these parameters cannot be meaningfully estimated individually, we can still estimate their sum. Finally, fitness parameters corresponding to the rows of the integrated covariance matrix with all zero entries are referred to as inaccessible as there is insufficient data to provide a meaningful estimate, either individually or as part of a sum, of these parameters. As an example, we can consider a population with two loci labeled 1 and 2 where only two genotypes are ever observed, one with both WT and one with both mutant alleles. Then the individual coefficients s_1, s_2, s_{12} cannot be independently inferred, but their sum $s_1 + s_2 + s_{12}$ can be estimated.

When the population consisted of all but one of the single mutant genotypes (*right* and *left* panels of second row of Figure 2.3), one of the selection coefficients was accessible (and thus accurately inferred) while the remaining two fitness parameters were partially accessible. In scenarios where the double mutant was absent from the population (*left* panel of third row of Figure 2.3), the selection coefficients were accessible, however there was no data to make any meaningful inference of the epistasis term. When the data contained only the WT and the double mutant genotypes (*right* panel of third row of Figure 2.3), all three fitness parameters were partially accessible as their inferred sum was accurate even though neither the selection coefficients nor the epistasis terms could be accurately inferred individually. Finally, in scenarios where only one of the two loci was polymorphic, and thus accessible, it was not possible to make a meaningful inference about the selection coefficient at the non-polymorphic locus or the pairwise epistasis term (*bottom left* and *bottom right* panels of Figure 2.3).

Additional tests demonstrated that the performance of MPL was consistent across a variety of landscapes, comprising of beneficial and/or deleterious selection coefficients and various forms of epistasis like positive, negative, positive sign and negative sign epistasis (Figure 2.4).

2.3.2 Analysis of a more complex five-locus epistatic fitness landscape

We ran further simulations on a more complex five-locus system to test the effects of data variability on the inference of MPL. Data variability in these simulations was controlled in two ways: (i) by specifying the number of unique genotypes in the initial population (Figure 2.8), and (ii) by combining data from multiple independent low-variability replicates (Figure 2.7). As expected, there was an increase in the fraction of accessible fitness parameters (Figures 2.8C, 2.8E, 2.7C and 2.7E) and better inference of the fitness landscape (Figures 2.8B and 2.7B) as the level of data variability increased. Our results show that for a given level of data variability, the fraction of accessible selection coefficients is higher than the fraction of accessible epistasis terms (Figure 2.6), i.e., higher data

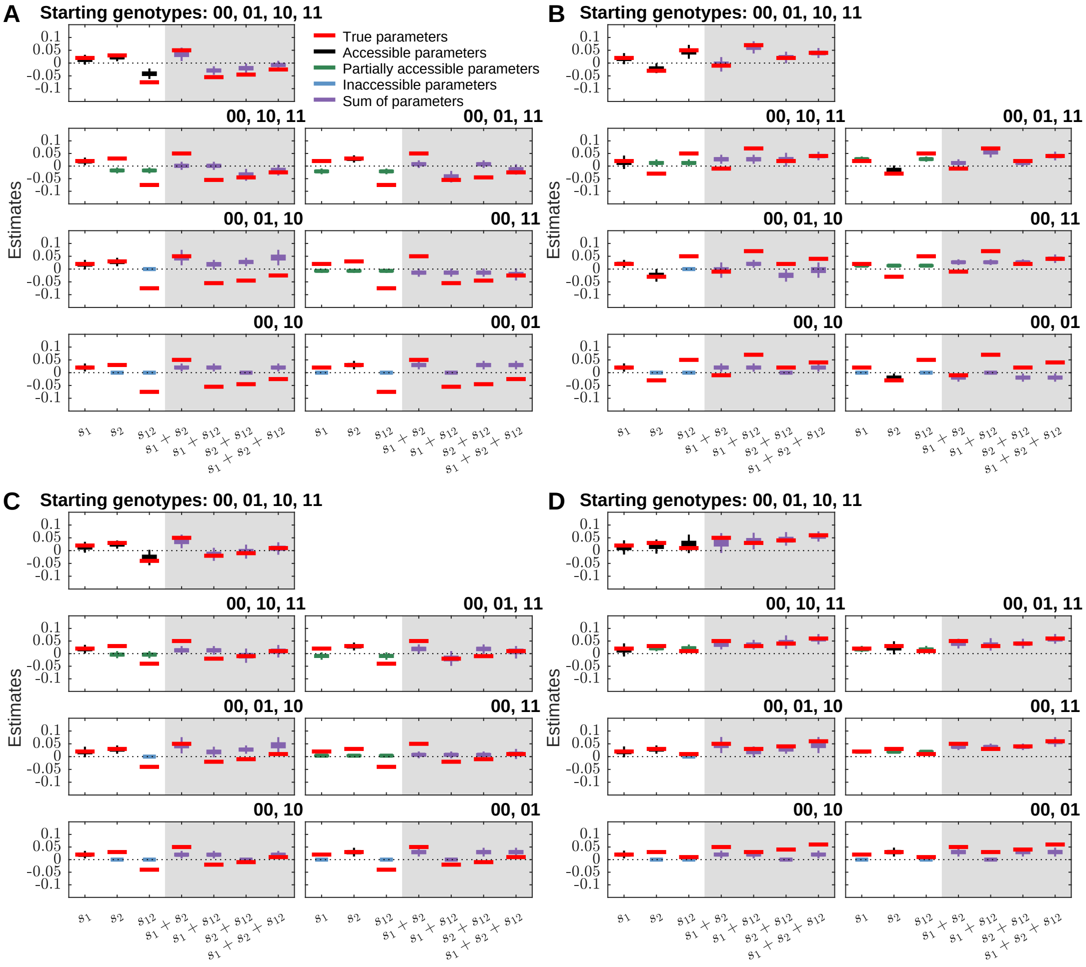

Figure 2.4: MPL can accurately estimate individual selection coefficients and pairwise epistasis terms and/or their sums depending on the variation present in the population. Results are for a two-locus system with (A) negative sign epistasis, (B) positive sign epistasis, (C) negative epistasis, (D) positive epistasis. The boxplots of inferred selection coefficients and epistasis terms are shown on white background in each panel, while those of their sums are shown on grey background. The red lines indicate the true values of the respective terms. The boxplots show the standard data summary (minimum, first quartile, median, third quartile, maximum). In order to control data variability, both the per locus mutation probability and the per locus recombination probability were set to zero. The initial population contained the genotypes indicated above each panel, and the frequency of each non-WT genotype in the initial population was set to 10% of the population size. The sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, with $T = 150$ generation used for inference.

variability is required for inference of epistasis than that required for inference of selection coefficients alone. This is because, for an epistasis term to be accessible, both corresponding selection coefficients must also be accessible.

We used area under the receiver operating characteristic curve (AUROC) as a performance metric to quantify the ability of MPL to classify beneficial and deleterious fitness parameters. When computed over all selection coefficients (*left* panels of Figures 2.8D and 2.7D) and all pairwise epistasis terms (*left* panels of Figures 2.8F and 2.7F), the results showed higher detection performance with increasing data variability. The poor performance at low variability was due to the large number of parameters that were either inaccessible or partially accessible, and thus cannot be meaningfully inferred due to lack of data. Computing the AUROC metric but restricted to *only* those selection coefficients classed as accessible revealed that the MPL estimator was able to correctly classify nearly all of such selection coefficients, under all scenarios considered (*right* panels of Figures 2.8D and 2.7D). The classification of accessible epistasis terms also showed good performance at moderate and high data variability (*right* panels of Figures 2.8F and 2.7F). Although none of the epistasis terms were accessible at low data variability, combining multiple replicates using (2.70) resulted in some epistasis terms becoming accessible (Figure 2.7E).

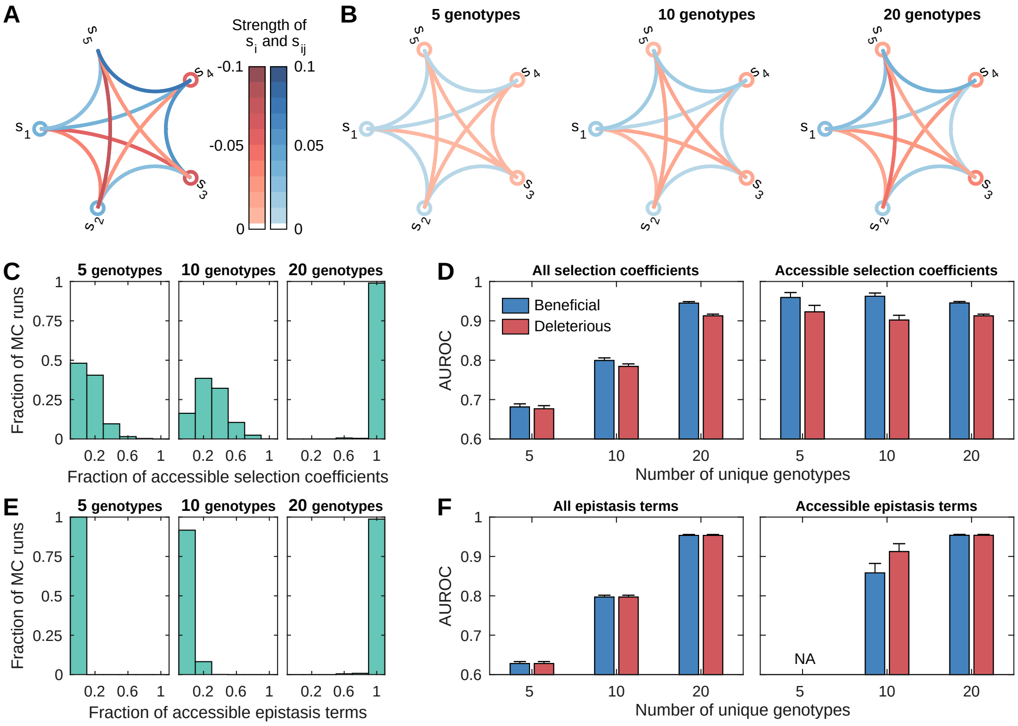

Figure 2.5: The fraction of selection coefficients and epistasis terms that are accessible depends on data variability. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by varying the number of unique genotypes in the initial population to either 5, 10 or 20). (C) shows the fraction of accessible selection coefficients as a function of data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs.

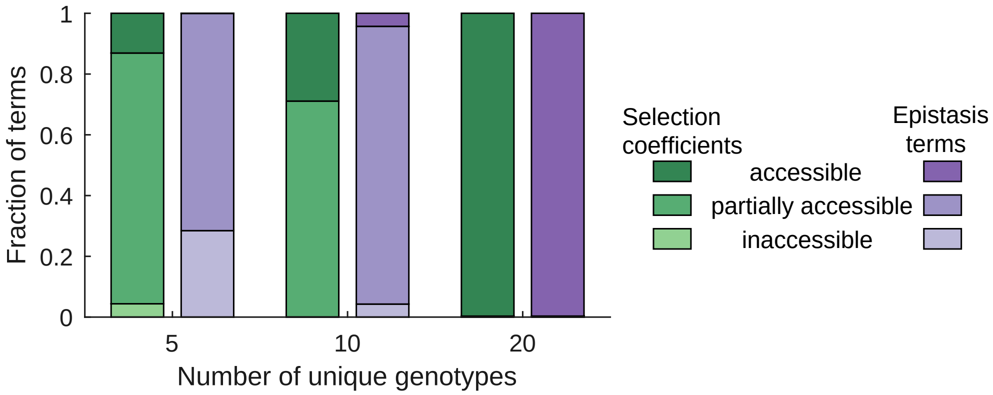

Figure 2.6: The average fraction of accessible selection coefficients and pairwise epistasis terms increases with increasing data variability (controlled by changing the number of unique genotypes in initial population to either five, ten or twenty). The selection coefficients and pairwise epistasis terms of a five-locus system were classified into three categories based on the reduced row-Echelon form (see 2.2) of the integrated covariance matrix in (2.65). Results are for a five-locus system (i.e., five selection coefficients and ten pairwise epistasis terms) simulated in Figure 2.8.

Similar results were obtained across a range of fitness landscapes differing in the degree of sparsity in their pairwise epistasis terms (Figure 2.9). These tests demonstrate that MPL has a very good ability to detect those fitness parameters for which there is sufficient data to enable inference and classification.

2.3.3 Robustness to sampling parameters

The accuracy of the estimator will depend on how well the underlying population dynamics is sampled. This includes how often the population is sampled in time and the number of samples measured at each time point. Here we test the robustness of the MPL method with respect to these sampling parameters. In general, one would expect

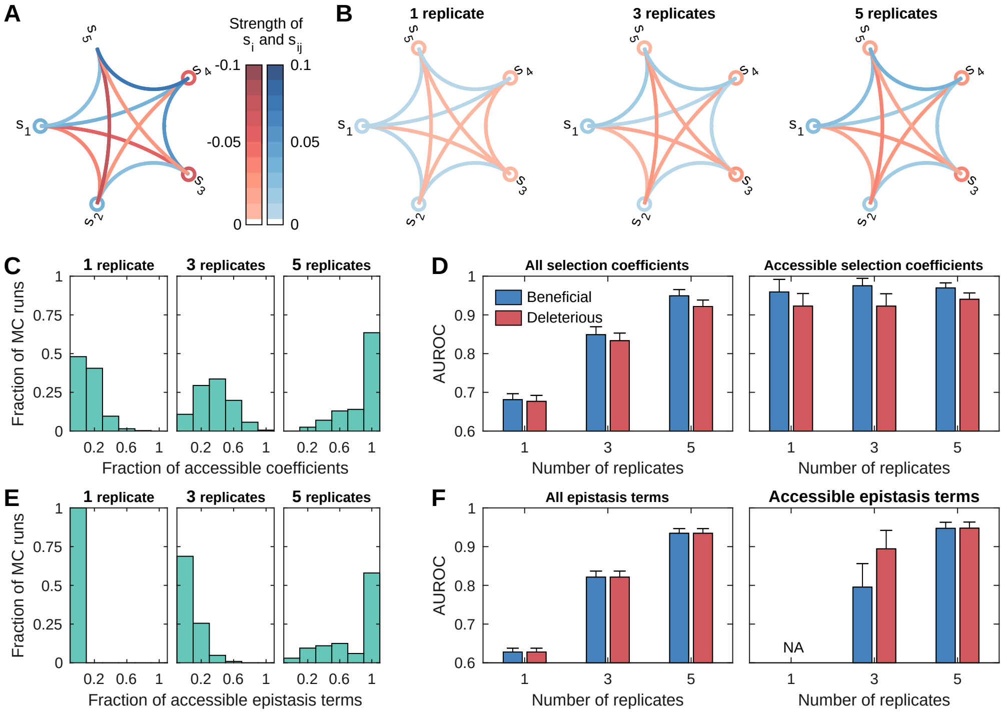

Figure 2.7: The fraction of selection coefficients and epistasis terms accessible in low data variability scenarios can be increased by combining multiple independent replicates. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by using either 1, 3 or 5 replicates for inference). (C) shows the fraction of accessible selection coefficients increases with the increase in data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. The initial population contained five unique genotypes. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs.

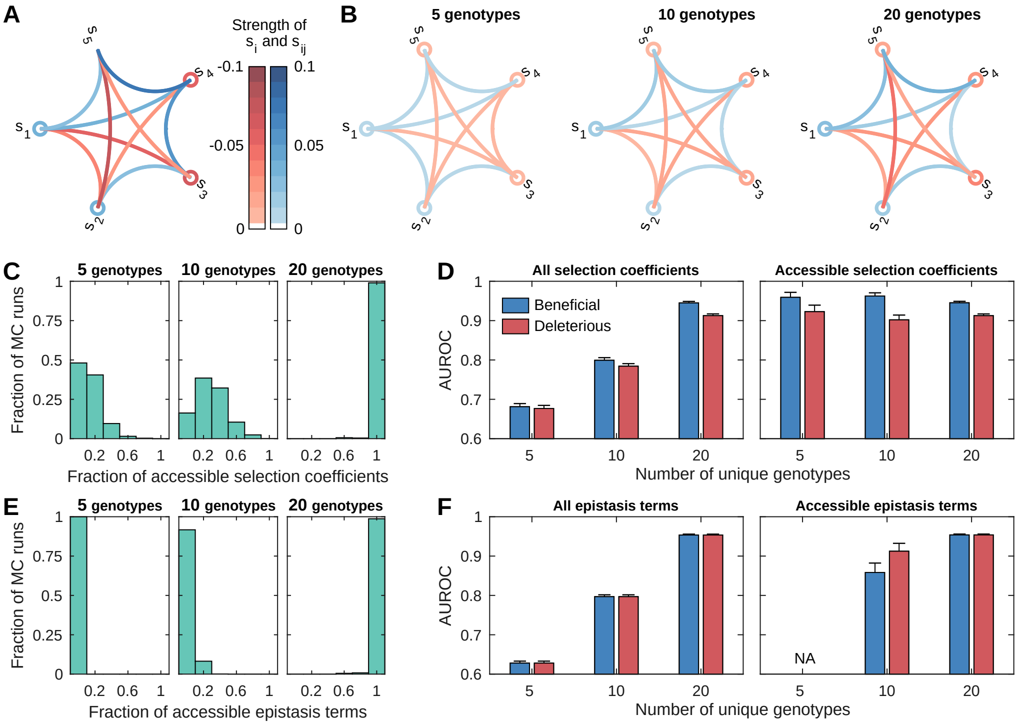

Figure 2.8: The fraction of selection coefficients and epistasis terms that are accessible depends on data variability. (A) shows the true fitness parameters of the five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of data variability (controlled by varying the number of unique genotypes in the initial population to either 5, 10 or 20). (C) shows the fraction of accessible selection coefficients as a function of data variability. The *left* and *right* panels of (D) show the classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of data variability. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control data variability. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs.

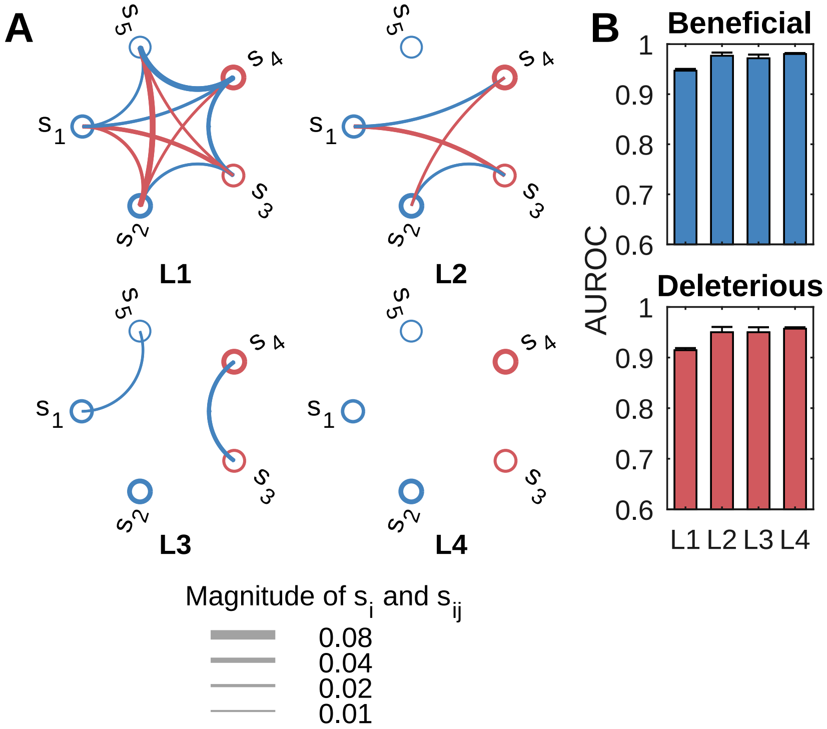

Figure 2.9: MPL performs well on dense as well as sparse fitness landscapes. (A) shows true (model) fitness landscapes with varying density of pairwise epistasis terms, while (B) shows the corresponding mean AUROC of detecting accessible beneficial and deleterious selection coefficients. Results are for a five-locus system where the initial population contained twenty unique genotypes, with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. The selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). Error bars indicate the standard error of the mean.

performance to degrade as samples are taken further apart in time, as less of the trajectory dynamics are captured. Moreover, taking limited samples at each time point would reduce the accuracy of the allele frequency estimates which may also compromise the accuracy of the MPL estimate.

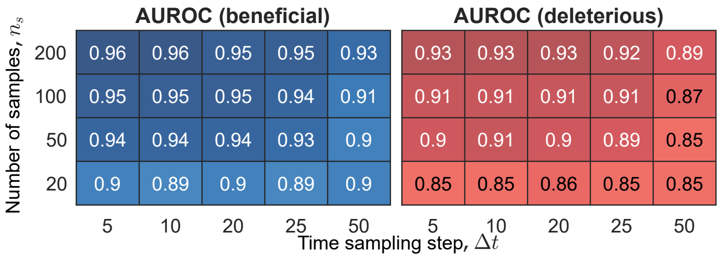

Figure 2.10: MPL is robust to variation in sampling parameters. The *left* and *right* panels show the mean AUROC performance of detecting accessible beneficial and deleterious selection coefficients, respectively. Results are for a five-locus system with the fitness landscape shown in Figure 2.8A. The initial population contained twenty unique genotypes with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$ and $T = 100$ generations were used for inference. All results were averaged over 1000 Monte Carlo runs.

To test the robustness of estimator, we ran extensive simulations under various sampling conditions. These simulations demonstrated that MPL can accurately detect both accessible selection coefficients and accessible epistasis terms for a range of sampling parameters (Figure 2.10 and Figure 2.11, respectively). MPL performed quite well even when the observed data consisted of a low number of samples, n_s, with only a few time samples (large time sampling step, Δt). For example, at $n_s = 50$ (from a population of $N = 1000$), the AUROC of detecting accessible beneficial selection coefficients (Figure 2.10) varied from 0.94 to 0.9 when the time sampling step was increased from $\Delta t = 5$ to $\Delta t = 50$ (corresponding to 21 and 3 time samples respectively). These results show that MPL estimator is robust to reasonable limitations in sampling depth and frequency.

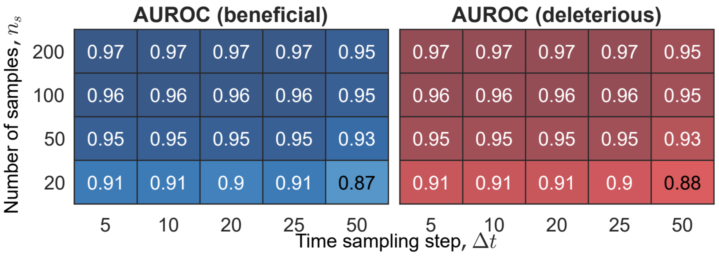

Figure 2.11: MPL is robust to variation in sampling parameters. The *left* and *right* panels show the mean AUROC performance of detecting beneficial and deleterious accessible epistasis terms respectively. Results are for a five-locus system with the fitness landscape shown in Figure 2.8A. The initial population contained twenty unique genotypes with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$.

2.3.4 Comparison with a model that does not account for epistasis

For fitness landscapes with epistasis, any inference model that does not explicitly account for epistasis will ascribe the effect of epistasis terms to individual selection coefficients, thereby over- or under-estimating them.

To test this, we ran simulations to compare the performance of the MPL method, which accounts for both linkage and epistasis, with the one we proposed previously, which accounts only for linkage and considers a first order fitness model with no epistasis [155]. Here, we term this variant as 'MPL (without epistasis)'. We generated numerous fitness landscapes, with similar magnitudes of selection coefficients and pairwise epistasis terms as the fitness landscape in Figure 2.8A, but differing in the density of epistasis terms from a purely additive landscape (no epistasis terms) to a highly epistatic landscape (with all pairwise epistasis terms being non-zero). We grouped these landscapes on the basis of number of non-zero pairwise epistasis terms. Our results demonstrate that the two estimators have similar performance when the underlying fitness landscape is additive or has low density of pairwise epistasis terms, while the estimator accounting for epistasis has superior performance when the fitness landscape is highly epistatic (Figure 2.12). Interestingly, our simulations showed that even in scenarios where none of the epistasis terms were accessible (Figure 2.6), MPL still showed a marked improvement in performance over MPL (without epistasis) in classifying accessible selection coefficients (Figure 2.13). Overall, our approach enabled us to disentangle the confounding effects of linkage and epistasis from data, resulting in more accurate inference of fitness parameters.

2.3.5 Computational complexity

The closed-form nature of the MPL estimate (2.65) makes it potentially computationally efficient. The two most computationally-intensive steps in the algorithm are (i) calculating the triple and quadruple mutant allele frequencies from the data, and (ii) inversion of the regularized integrated covariance matrix. The number of triple and quadruple mutant frequencies required for computing the inverse term in (2.65) increases as L^4, where L is the number of loci. However, this number can be reduced following variability in the data. For instance, for any locus-pair (i, j) whose double mutant frequency is zero, it follows that any three tuple (i, j, k) involving the same pair will have a triple mutant frequency of zero and hence its calculation can be avoided. Similarly the number of quadruple mutant frequencies that need to be computed can also be reduced.

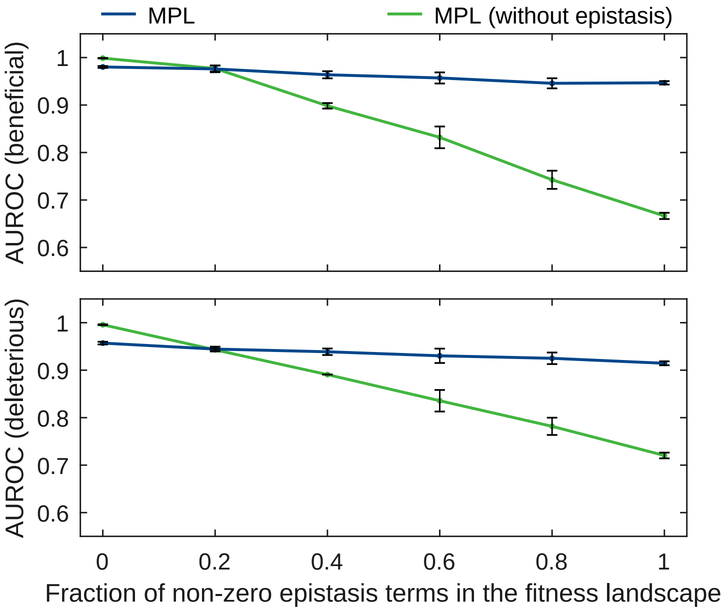

Figure 2.12: Ability of MPL to accurately identify selection coefficients is robust to the density of non-zero epistasis terms in the fitness landscape. Comparison of MPL with MPL (no epistasis) method of [155] shows that both methods have similar performance when the underlying fitness landscape has few epistatic interactions. As the density of non-zero pair-wise epistasis terms increases in the underlying fitness landscape, there is a continuous degradation in the performance of MPL (no epistasis) while performance of MPL remains robust. The *top* and *bottom* panels show the mean AUROC performance of detecting beneficial and deleterious selection coefficients from the rest respectively. Error bars indicate the standard error of the mean. Results are for a five-locus system where the initial population contained 20 unique genotypes, with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. The sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, with $T = 100$ generation used for inference. All simulation results were computed over 1000 Monte Carlo runs.

The computations required for computing the inverse term can also be reduced by considering only the polymorphic loci $L_p < L$, instead of the whole sequence, leading to $R_p = L_p(L_p + 1)/2$ parameters to be estimated. The inverse would then require $\mathcal{O}(R_p^3)$ computations, with $R_p \ll R$ in practice for realistic data sets.

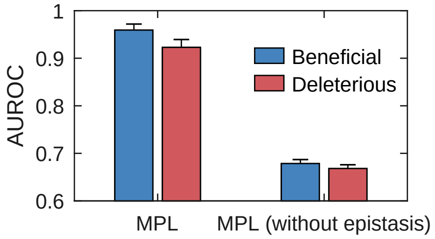

Figure 2.13: MPL outperforms the MPL (without epistasis) method in classification of beneficial and deleterious accessible selection coefficients even on data with low variability. The figure shows the mean AUROC performance of the two methods. Error bars indicate the standard error of the mean. Results are for a five-locus system with a fully connected fitness landscape. The initial population contained five unique genotypes, with both the per locus mutation probability μ and per locus recombination probability r set to zero.

2.4 Discussion

Epistasis is a pervasive phenomenon that can strongly shape the evolution. Genetic time-series data provides an opportunity to detect and estimate epistatic contributions to fitness. However, developing methods that can efficiently yield accurate inferences has remained a challenge. Here we proposed a method to address this challenge. Our approach is a physics-based approach that builds upon a framework that we recently introduced for non-epistatic models [155]. Through simulations, we demonstrated that our method can accurately infer both pairwise epistasis effects and individual selection coefficients, provided sufficient variation exists in the data.

Moreover, the method systematically reveals necessary conditions on genetic variation in the data in order for accurate inferences to be possible, and for the separate contributions of epistasis and allele selection coefficients to be inferrable.

MPL uses a path-integral to approximate the likelihood of a set evolutionary parameters (including epistasis), given an observed time-series of allele frequencies and their correlations. This framework can also be adapted for different evolutionary scenarios. In recent work, it was applied together with epidemiological models to infer the transmission effects of mutations from genomic surveillance data, and to study the evolution of SARS-CoV-2 [103]. The data input to MPL, under a fitness model with pairwise epistasis terms, consists of the single, double, triple and quadruple mutant allele frequencies. While these are readily available from long-read sequencing data, the double and higher mutant allele frequencies cannot be computed extensively for short-read data. More work is required to develop methods that can accurately detect or infer selection and epistasis for such data sets. However, the trend toward longer read lengths in third-generation sequencing technologies [130] suggests that higher order mutant frequencies will be more readily available in future data sets. While fitness models with higher-order epistasis involving more than two mutant alleles are also possible [181], here we restricted our analysis to a fitness model with pairwise epistasis terms. In principle, the MPL framework can be extended to account for higher-order epistasis terms by explicitly modeling the evolution of higher-order mutant allele frequencies. However, the contribution of epistasis terms to fitness typically decline with order [180] and, at least in some scenarios, the gain achieved by modeling higher-order epistasis beyond pairwise terms appear to be minimal [110].

MPL, like all inference methods, requires sufficient diversity to enable parameter inference. For a fitness model with pairwise epistasis terms, the number of model parameters to be inferred increases quadratically with the sequence length. As such, data with insufficient variation may lead to a situation where most of the model parameters are partially accessible or inaccessible (Figure 2.6). This is not intrinsically a limitation of our specific method, but rather of a lack of exploratory power in the data.

The current approach infers a fitness landscape with epistasis terms between every pair of mutant alleles, in contrast to an additive fitness landscape inferred in [155]. One can also consider selecting the most likely fitness model, given the data, from a reduced set of models with different densities of epistasis terms using a model selection approach. However, it may only be feasible to pursue selection approaches for moderate sized systems due to the exponential increase in the number of possible models with increasing system size. An alternative approach can be to apply a sparsity constraint on the epistasis terms. Future work on this problem can leverage sparsity inducing techniques such as the least absolute shrinkage and selection operator (LASSO) regression family of methods [169, 188], to come up with a computationally efficient algorithm suitable for systems with hundreds or thousands of segregating mutations.

Chapter 3

Inferring mutational effects with deep mutational scanning data

3.1 Traditional mutagenesis experiments

3.1.1 Alanine mutagenesis experiments

In the field of molecular biology, alanine scanning is a targeted mutagenesis method employed to investigate the role of a particular residue in the stability or function of a protein of interest. This technique involves substituting the target residue with alanine, a non-bulky and chemically inert amino acid. The choice of alanine is based on its methyl functional group, which mimics the secondary structure preferences observed in many other amino acids. In certain cases where it is necessary to maintain the size of the mutated residue, bulky amino acids such as valine or leucine may be utilized. Alanine scanning enables researchers to assess the significance of specific amino acid residues in protein structure and function. By systematically replacing target amino acids with alanine, a small and neutral amino acid, scientists can gain valuable insights into the structure, stability, and function of proteins. [121]

The alanine scanning technique is not only useful in determining the contribution of specific residues to protein stability or function but also for assessing the involvement of side chains in bioactivity. This can be achieved through site-directed mutagenesis or the creation of a PCR library with random mutations. Additionally, computational methods have been developed to estimate thermodynamic parameters by considering theoretical alanine substitutions.

One notable advantage of this technique is its speed, as it allows for the simultaneous analysis of multiple side chains, eliminating the need for extensive protein purification and biophysical analysis. With its widespread use in biochemical fields, the technique has become highly matured and well-established. The obtained data can be further validated using various methods such as infrared spectroscopy (IR), nuclear magnetic resonance (NMR) spectroscopy, mathematical analyses, or bioassays [183, 150, 79, 64].

Initially, the target residue is identified based on existing knowledge or predictions. Primers are designed to introduce the desired alanine mutation and surround the target residue. These primers are used in PCR amplification to create a mutated DNA construct, which is then sequenced to confirm the presence of the desired mutation. The mutant protein is produced by expressing the mutated DNA in suitable expression systems and purified using established protein purification techniques. A range of biochemical and biophysical assays are performed to compare the properties of the mutant protein with the wild-type version. By analyzing the results, researchers can assess the effects of the alanine mutation on the protein's structure, stability, enzymatic activity, and other functional aspects, thereby gaining insights into the role of the mutated residue in protein function.[189]

The next crucial step involves functional analysis of the alanine mutant protein. Researchers employ a range of experimental approaches to assess the impact of the alanine substitution on the protein's properties [94, 120, 132, 102, 20]. This may include enzymatic activity assays, binding studies with ligands or substrates, structural analyses using techniques like X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, or other relevant assays depending on the protein's function and research objectives. By comparing the behavior and characteristics of the mutant protein to that of the wild-type protein, researchers can discern the role played by the specific amino acid residue in question.

The information obtained through alanine mutagenesis experiments can be highly valuable in multiple areas of research. For instance, it can contribute to our understanding of protein structure-function relationships[149], shed light on the mechanisms underlying protein-ligand interactions or enzymatic activities[163], and aid in the design of novel therapeutics or protein engineering strategies[98, 33]. By systematically mutagenizing multiple amino acid residues within a protein, researchers can construct detailed maps of the protein's functional hotspots and identify critical regions responsible for its activity or binding specificity.

In summary, alanine mutagenesis is a versatile and widely employed technique that allows researchers to probe the functional significance of specific amino acids within proteins. By systematically introducing alanine substitutions, scientists can gain valuable insights into protein structure, stability, and activity. This approach has revolutionized our understanding of protein function and has far-reaching implications in numerous fields, from basic molecular biology to applied biotechnology and medicine.

3.1.2 Shortcomings of Alanine mutagenesis experiments

While alanine mutagenesis is a valuable tool in protein research, it also has some limitations and potential shortcomings. Although alanine-scanning mutagenesis is a valuable technique for mapping protein-binding interfaces, it can be a labor-intensive process. The method requires the production and purification of numerous mutant proteins, with each mutant needing individual evaluation of its structural integrity and binding affinity.[183] And alanine is a small, non-polar amino acid, and its introduction may not fully capture the effects of the original residue, particularly if it possesses unique properties or contributes to specific interactions within the protein. Thus, mutaional effects of other genetic variants on mutated sites would not be observed. Alanine mutagenesis only provides information about the specific role of individual residues within a protein, but it may not capture the intricate network of interactions and cooperative effects that can exist within protein structures. Proteins are complex systems, and altering a single residue may not fully reflect the consequences of multiple mutations occurring simultaneously.

Furthermore, alanine mutagenesis experiments typically focus on the effects of substitutions at specific positions, which may not encompass the entire protein sequence. This approach may overlook residues that are distant from the targeted positions but still play critical roles in the protein's structure or function. While alanine mutagenesis can provide insights into the impact of mutations on protein stability and function, it does not provide direct information on the underlying molecular mechanisms involved. Additional experiments and complementary techniques are often required to fully understand the functional consequences of alanine substitutions.

Despite these limitations, alanine mutagenesis remains a valuable tool in protein research, particularly when combined with other approaches and used in a complementary manner. It is important for researchers to be aware of these limitations and consider them when interpreting the results obtained from alanine mutagenesis experiments.

3.2 Deep mutational scanning (DMS) experiments

3.2.1 Introduction of DMS experiments

Deep mutational scanning (DMS) experiment is a powerful approach used to systematically explore the functional consequences of amino acid substitutions across a protein sequence. These experiments combine high-throughput mutagenesis with next-generation sequencing to generate large libraries of mutant variants and measure their effects on protein function, stability, or other desired properties. DMS allows for the comprehensive analysis of mutational effects, providing valuable insights into protein structure-function relationships, evolution, and disease-related variants.

The general workflow of a deep mutational scanning experiment involves several key steps, seen Fig. 3.1: 1. Library Generation: A diverse library of mutant variants is generated by introducing random or targeted mutations in the DNA sequence encoding

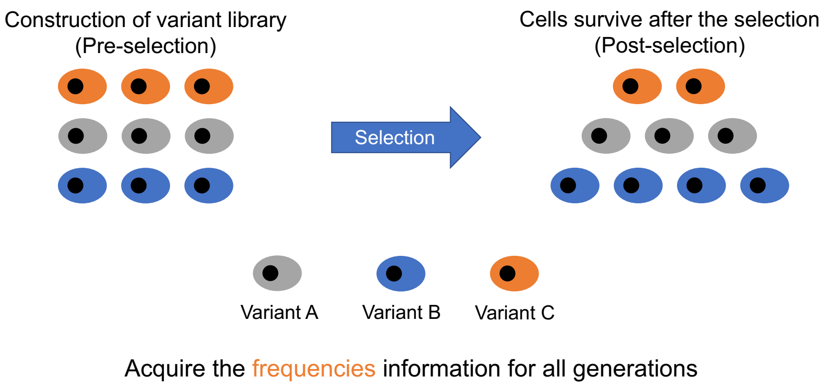
Construction of variant library
(Pre-selection)
Selection
Cells survive after the selection
(Post-selection)
Variant A
Variant B
Variant C
Acquire the frequencies information for all generations

Figure 3.1: Main steps in Deep mutational scanning experiments.

the protein of interest. This can be achieved through techniques such as error-prone PCR, saturation mutagenesis, or synthetic DNA oligonucleotide pools. The library should cover a wide range of amino acid substitutions at each position in the protein sequence.

2. Library Expression and Selection: The generated mutant library is expressed in a suitable host system, such as bacteria, yeast, or mammalian cells. The expressed proteins are subjected to a selection or screening process that assesses the functional consequences of the introduced mutations. This can involve various assays, such as enzyme activity assays, binding assays, growth-based assays, or fluorescence-based readouts.

3. High-Throughput Sequencing: Next-generation sequencing technologies are employed to quantify the abundance of each mutant variant in the library before and after selection. The DNA or RNA from the initial library and the selected pool of mutants is sequenced, allowing for the determination of the mutation frequency at each position in the

protein sequence. This provides a quantitative measurement of the effects of each mutation on protein function or other desired properties.

4. Analysis and Interpretation: The sequencing data is analyzed to identify mutations that have a significant impact on protein function. Statistical methods, such as enrichment analysis or machine learning algorithms, can be applied to determine the effects of individual mutations or identify functionally important residues or regions. Comparison with control datasets or reference sequences can further aid in the interpretation of the data.

3.2.2 Advantages of DMS experiments and comparison with traditional mutagenesis techniques

DMS experiments offer distinct advantages over traditional mutagenesis experiments when it comes to quantitative assessment of mutational effects. DMS experiments utilize high-throughput sequencing technologies to quantify the abundance of each mutant variant in the library before and after selection. This allows for the precise measurement of the effects of individual mutations across the entire protein sequence. In contrast, traditional mutagenesis experiments often rely on qualitative or semi-quantitative assessments, such as visual screening or limited sampling of mutant clones, which may not provide comprehensive quantitative data on mutational effects.

And DMS experiments generate large libraries of mutant variants, covering a wide range of amino acid substitutions at each position in the protein sequence. This enables the exploration of a comprehensive mutational landscape, providing a more accurate assessment of the effects of individual mutations. Traditional mutagenesis experiments, especially those involving site-directed mutagenesis, typically focus on specific mutations or a limited number of variants, which may not capture the full range of mutational effects.

More quantitatively, DMS experiments employ statistical methods to analyze the sequencing data and determine the effects of individual mutations. These analyses allow for the identification of mutations that significantly impact protein function and the quantification of their effects. In contrast, traditional mutagenesis experiments often rely on visual inspection or qualitative assessments, making it challenging to precisely quantify mutational effects or perform statistical comparisons between different mutants. DMS experiments enable the estimation of fitness or functional scores for individual mutant variants. By comparing the abundance of each mutant in the starting library to its abundance in the selected pool, researchers can infer the impact of each mutation on protein function. This quantitative assessment provides a more precise understanding of the relative effects of different mutations. In traditional mutagenesis experiments, it is generally more challenging to obtain quantitative fitness or functional scores due to the limited scope of mutational exploration.

To compare with alanine mutagenesis experiments, while alanine mutagenesis experiments focus on replacing specific amino acids with alanine, which is a small, non-polar amino acid, DMS allows for the simultaneous analysis of multiple amino acid substitutions at every position in a protein sequence. This comprehensive coverage enables the identification of a broader spectrum of functional variants and provides a more thorough understanding of the effects of mutations. DMS leverages high-throughput techniques and statistical analysis to generate large datasets and derive meaningful insights. The use of high-throughput sequencing allows for the analysis of thousands or even millions of variants simultaneously, providing a more comprehensive exploration of mutational effects. Statistical methods applied to DMS data enable robust identification of functionally important residues, estimation of fitness landscapes, and comparisons between mutants. In contrast, alanine mutagenesis experiments typically involve a smaller scale of mutational analysis and may lack the statistical power necessary for comprehensive characterization. Moreover, DMS allows for the exploration of chemical diversity beyond alanine substitutions. Alanine, being a non-polar residue, may not adequately mimic the chemical and structural properties of the original amino acid being replaced. In contrast, DMS enables the investigation of a wide range of amino acid substitutions, including polar, charged, and aromatic residues. This broader exploration of chemical space provides a more realistic assessment of how diverse amino acid changes impact protein function and structure.

Overall, DMS represents a powerful and advanced approach compared to traditional alanine mutagenesis. Its ability to comprehensively explore sequence space, encompass chemical diversity, provide quantitative measurements, identify functional hotspots, and leverage high-throughput techniques and statistical analysis make it the preferred choice for many researchers studying protein function, structure, and evolution. While traditional mutagenesis still has its utility in targeted investigations, DMS offers a more comprehensive and nuanced understanding of mutational effects, facilitating deeper insights into protein biology.

3.3 Related data sets resources

We analyzed DMS experiments from two main sources: Multiplex assays of variant effect database (MaveDB) and data shared from the lab of Prof. Jesse Bloom at the Fred Hutchinson Cancer Research Center. The main difference between MaveDB data and Bloom's data is about the independent mutations sequencing and full genetic sequence reads. Data sets in MaveDB always have full genetic sequencing reads, while Bloom's lab data always only have independent mutation sequencing, especially the single site variant reads. People used different analytical tools to analyze experimental data generated by different resources, however, our method, popDMS, could help with inferring the mutational effects, including single-site selection coefficients or pairs epistasis, in multiple experimental data types systematically.

3.3.1 Multiplex assays of variant effect database (MaveDB)

The exploration of genetic variation through experimental methods has played a pivotal role in unveiling the intricate mechanisms governing gene functionality and enhancing our comprehension of the clinical implications of human genetic diversity. The advent of multiplex assays of variant effect (MAVEs) capitalizes on high-throughput DNA sequencing to significantly expand the scope of variants that can be subjected to experimental scrutiny.

A MAVE experiment furnishes a collection of scores that elucidate the functional impact of numerous variants within a coding sequence, promoter, enhancer, or other genetic components when compared to a reference sequence. MAVEs have garnered rapid adoption across both fundamental research and clinical applications. As a result, the cumulative count of variants furnished with functional data through MAVE experiments was projected to surpass 200,000 by the conclusion of 2018 [178].

Despite the significance of MAVE data for both fundamental research and clinical applications, there lacks a standardized resource for the discovery and dissemination of this information. In response to this gap, they have developed MaveDB (accessible at `https://www.mavedb.org`) [46], a publicly accessible repository tailored for housing large-scale measurements of sequence variant impact. This repository is designed to be interoperable with applications aimed at interpreting these datasets.

Furthermore, they have introduced the inaugural application, named MaveVis, which serves to retrieve, visualize, and provide context to variant effect maps. By combining the functionality of the database and these applications, the scientific community gains the capability to efficiently extract insights from these potent datasets. This initiative not only addresses the need for centralized storage of MAVE data but also equips researchers with the tools necessary to harness the wealth of information contained within these datasets effectively.

3.3.2 Bloom's lab DMS data

The Bloom Lab (`https://github.com/jbloomlab`), led by Professor Jesse D. Bloom at the University of Washington's Department of Genome Sciences, is a pioneering research group renowned for its contributions to evolutionary biology, virology, and computational biology. A significant focus of the lab's work revolves around Deep Mutational Scanning (DMS) experiments, a groundbreaking technique that delves into the effects of genetic mutations on proteins and their functions.

DMS experiments, a cornerstone of the Bloom Lab's research, involve creating libraries of mutated protein variants, with each variant containing a single amino acid mutation. These libraries are then subjected to high-throughput assays that measure the impact of each mutation on the protein's properties, such as its stability, binding affinity, or enzymatic activity. By analyzing these large-scale datasets, researchers gain insights into the structural and functional consequences of individual mutations.

The DMS experiments conducted in the Bloom Lab are pivotal in deciphering how genetic changes shape the behavior of viruses, including influenza and HIV. These experiments shed light on critical questions, such as how viruses evolve to evade the immune system, how drug resistance emerges, and how viral proteins interact with host cells [186, 168, 72, 68, 17, 104, 71, 40, 43, 140, 5, 153, 77, 42]. DMS data are crucial for identifying functionally important amino acids, understanding the effects of mutations on protein structure, and predicting how viruses might evolve under different selection pressures.

3.4 State-of-the-art analytical methods inferring mutational effects with DMS data

The primary objective of a deep mutational scanning experiment commonly involves gauging the impact of individual amino-acid mutations on a protein's functionality. As an example, the intention might be to evaluate the influence of each mutation to a viral protein on the virus's capacity to replicate within cell cultures. The primary objective of a DMS experiment is to comprehensively assess the effects of various amino acid mutations on a protein's behavior, particularly its function and structure.

Analytical tools for DMS experimeents play a crucial role in advancing our understanding of how genetic mutations impact protein functionality. These tools aid in organizing and efficiently managing the vast volume of data, enabling accessibility and streamlined analysis. Moreover, analytical tools facilitate pattern identification by unveiling trends, correlations, and clusters of mutations within the data, elucidating functional regions in the protein. Robust statistical analysis is imperative for accurate interpretation of DMS data, and analytical tools provide the necessary framework for applying appropriate statistical tests to determine the statistical significance of observed effects. In cases where DMS experiments introduce biases or variations, analytical tools offer normalization techniques to rectify technical factors and ensure precise assessment of mutation effects. Visualization of intricate DMS datasets is simplified through these tools, as they assist in creating visual representations like heatmaps, scatter plots, and mutation landscapes, enabling researchers to comprehend overarching trends. Additionally, analytical tools facilitate integration of experimental data with protein structural information, a vital aspect of many DMS experiments aiming to correlate mutation effects with 3D protein structures. These tools aid in prediction validation, enabling the comparison of computational predictions with experimental DMS data, ultimately enhancing the accuracy of computational models. Furthermore, analytical tools provide valuable biological insight by helping researchers interpret the broader implications of observed mutation effects, including evolutionary contexts and disease relevance. The high-throughput nature of DMS experiments, involving simultaneous testing of numerous mutations, is efficiently managed by analytical tools, which enable the processing of large datasets and extraction of valuable insights. Lastly, in the collaborative landscape of DMS research involving data sharing, analytical tools play a pivotal role by standardizing data formats, facilitating the seamless exchange and comparison of results across diverse research groups.

3.4.1 *dms_tools* and *dms_tools*2, ratio method

One approach to quantifying the mutation effects is to express them as the propensity of each position within the protein for each potential amino acid. In essence, amino acids that are highly favored tend to become more prevalent (or maintain their existing frequency) following functional selection, while less favored amino acids tend to decrease in occurrence during this selection process.

dms_tools and *dms_tools*2[16] are software packages specifically designed to support the analysis of data generated from Deep Mutational Scanning (DMS) experiments. These tools assist researchers in handling and interpreting the large datasets produced by DMS experiments, which involve assessing the effects of mutations on protein function.

dms_tools is a software package that provides a suite of computational tools and algorithms for the analysis of DMS data. These tools are designed to process, visualize, and interpret the effects of mutations on protein properties. *dms_tools* typically includes functionalities such as data pre-processing, statistical analysis, visualization of mutation effects, and potentially integration with protein structure information. It helps researchers extract insights from complex DMS data sets and understand how mutations influence protein behavior.

*dms_tools*2 is an updated or enhanced version of the original *dms_tools*. It could include additional features, improved algorithms, and better support for the evolving needs of researchers working with DMS data. Given that software tools in scientific research are often refined and updated over time, *dms_tools*2 could signify a continuation of the original software with enhancements based on user feedback and technological advancements.

An essential element of deep mutational scanning involves the analysis of the data, encompassing several steps. Initially, the raw reads obtained from deep sequencing need to undergo processing to quantify mutations both before and after the selection process. Subsequently, these mutation counts are utilized to deduce the biological impacts of the mutations. The enrichment ratio $E_{r,x}$ of variant r at site x is calculated as following:

$$E_{r,x} = \frac{f_{post}^{r,x}}{f_{pre}^{r,x}}, \tag{3.1}$$

where $f_{post}^{r,x}$ is the post-selection frequency of variant r at site x, while $f_{pre}^{r,x}$ is the pre-selection frequency of variant r at site x. Than the enrichment ratio is normalized by site to have the measurement of preference:

$$\pi_{r,x} = \frac{E_{r,x}}{\sum_i E_{i,x}}, \tag{3.2}$$

dms_tools and *dms_tools*2 employ a Bayesian methodology to deduce site-specific preferences from DMS experimental data. The algorithm operates by evaluating the probabilities of observed counts, considering the unknown preferences of amino acids at specific positions, as well as mutation and error rates. Plausible prior distributions are assigned to these unobserved parameters. These priors reflect the underlying assumption that all potential identities, such as different amino acids, possess equal preferences. Additionally, the mutation and error rates for each position are assumed to match the overall average rates for the entire gene.

The algorithm employs Markov Chain Monte Carlo (MCMC) techniques to compute the posterior probability of preferences, given the observed mutation counts. This approach helps refine our understanding of the preferences by incorporating both the data and the prior assumptions, ultimately providing a more comprehensive assessment of the site-specific effects.

3.4.2 *Enrich* and *Enrich*2, ratio-regression method

In the DMS experiments, the absence of a standardized method for calculating scores introduces challenges in comparing studies, while existing tailored techniques fail to suit the diverse range of current experimental designs. Additionally, no established approach measures the uncertainty surrounding each score, thus limiting the data usefulness. For example, an impactful application of deep mutational scanning involves annotating variants within human genomes to improve variant interpretation[158]. In this context, estimating the uncertainty linked to each measurement within a shared framework holds vital significance. Current practices, at best, resort to ad hoc filtering of potentially low-quality scores, often relying on manually determined read-depth thresholds.

To address these concerns, a solution emerges in the form of *Enrich*/*Enrich*2[59, 142]. These advanced and user-friendly computational tools present a comprehensive statistical model for the analysis of deep mutational scanning data. Especially, *Enrich*2 encompasses scoring techniques that can be flexibly applied to deep mutational scans, regardless of the number of time points involved. Unlike existing methods, *Enrich*2 not only delivers variant scores but also estimates standard errors, which consider both sampling discrepancies and consistency among replicates. The findings underscore that *Enrich*2's scoring methods surpass current approaches in a variety of experimental scenarios. By effectively eliminating noisy variants and improving the detection of subtle effect variants, *Enrich*2 facilitates robust statistical comparisons between different variants. Importantly, *Enrich*2's platform independence and user-friendly graphical interface cater to experimental biologists with limited experience in bioinformatics, making it highly accessible.

When dealing with experimental designs involving two or more time points, *Enrich*2 is utilized by the researchers to calculate a score for each variant through weighted linear least squares regression. These time points exhibit potential variation in their spacing; they might be irregularly spaced, such as in cases where samples are collected at different intervals during a yeast selection, or they can be uniformly spaced to represent discrete rounds or bins, as observed in successive rounds of a phage selection. The method is built upon the premise that the selection pressure remains relatively constant throughout the selection process. Operating within this framework, the score allocated to each variant corresponds to the slope of the regression line. For every time point within the selection, including the initial time point, a logarithmic ratio is computed, indicating the frequency of the variant relative to the wild-type's frequency at the corresponding time point. These ratios are subsequently subjected to regression analysis against time.

To ensure precise results, regression weights are computed for each variant at each time point, leveraging the Poisson variance of the variant's count. A standard error for each score is estimated using the weighted mean square of the residuals around the fitted regression line. To evaluate the statistical significance of each score, p-values are calculated using the z-distribution, operating under the assumption of a null hypothesis wherein the variant's behavior mirrors that of the wild-type (i.e., it exhibits a slope of 0). If the slope of one variant is larger than 0, it means this variant is beneficial than the wild-type variant and smaller than 0 means it is a deleterious variant comparing with wild-type residue.

3.4.3 Drawbacks of popular analytical tools inferring mutational effects

Analytical tools have demonstrated remarkable success in deep mutational scanning (DMS) experiments, revolutionizing our understanding of protein structure, function, and evolution. Their success is evident through various advancements and discoveries in fields such as molecular biology, biotechnology, and medicine. However, there are some drawbacks in current tools:

1. No support theoretical support and lack of hitchhiking effect solution: Unlike strict inference of MPL[154, 155], there is no underlying theoretical math support for preference or other ratio-based methods. Although the variant frequency increment or decrement could be an indicator whether this variant is adaptable to the selection environment, these ratio based methods lack of explanation about why the frequency ratio or slope of the frequency trajectory can describe how much frequency of particular variant is accumulated or killed. Also, the frequency increment of genetic variants is not always caused by beneficial properties, but maybe from the hitchhiking effects[76]. MPL is an inference method quantifying mutational effects by measuring how much variants survive or die during the selection stage. The selection coefficients or epistasis can be used to quantify how much of genetic variants will be changed during the evolution. And MPL is a reliable inference tool resolving the genetic linkage effects.

2. Sensitive to the small frequencies: Because of ratio calculation in the current analytical tools inferring mutational effects, the small frequencies from finite sampling is a big problem in mutational effects inference. Finite sampling in sequencing experiments refers to the limitation posed by the number of sequences that can be analyzed or obtained from a given sample. In molecular biology and genetics, sequencing experiments involve determining the order of nucleotides or amino acids in a DNA, RNA, or protein sample. However, due to technological and resource constraints, it's often not possible to sequence every single molecule in a sample, resulting in a subset of sequences being obtained and analyzed. Such small frequencies cause sensitive calculation of preference or functional scores. In MPL, utilizing Bayesian techniques to incorporate the uncertainties arising from finite sampling in variant frequency trajectories has the potential to enhance the resilience of our methodology. When dealing with limited data due to finite sampling, Bayesian approaches offer several benefits that help mitigate the uncertainties and biases arising from small sample sizes. Bayesian methods allow people to incorporate prior knowledge or beliefs about the phenomenon people are studying. This prior information can provide valuable context, helping to guide the analysis when data are sparse. By combining prior knowledge with observed data, MPL can yield more robust and accurate estimates and predictions. Also, MPL involve adding regularization terms to the analysis. This helps prevent overfitting, a common issue when working with small data sets. Regularization ensures that model complexity is controlled, which can improve the generalizability of results beyond the limited sample.

3.5 popDMS infers mutation effects with DMS experiments data

Understanding the relationship between protein sequence and phenotype is a central question in evolution and protein engineering. In recent years, a new family of experimental methods, commonly referred to as deep mutational scanning (DMS) or multiplexed assays for variant effects (MAVEs), have been developed to directly measure the functional effects of large numbers of mutations simultaneously [57, 63]. DMS experiments generally work by generating a vast library of protein variants that are then passed through rounds of selection that favor functional variants while eliminating inactive ones [60]. One can then compare variant frequencies in the pre- and post-selection libraries to estimate the functional effects of mutations. This approach has been successfully applied in a wide variety of contexts, from studying the function of enzymes [139] and tRNAs [107] to measuring the mutational tolerance of influenza [168, 104, 43] and HIV-1 [71, 40, 72] surface proteins.

Despite the success of DMS experiments, current approaches for analyzing DMS data yield surprisingly modest correlation between the inferred functional effects of mutations in experimental replicates, leaving a significant amount of variance in the data unexplained. Popular methods use the ratio between post- and pre-selection variant frequencies to estimate mutational effects [59, 75, 16]. Ratio-based methods are particularly sensitive to noise when variant counts are low, a common occurrence in DMS experiments. Regression-based approaches [3, 158, 118, 136, 142] provide better performance, but substantial uncertainty remains.

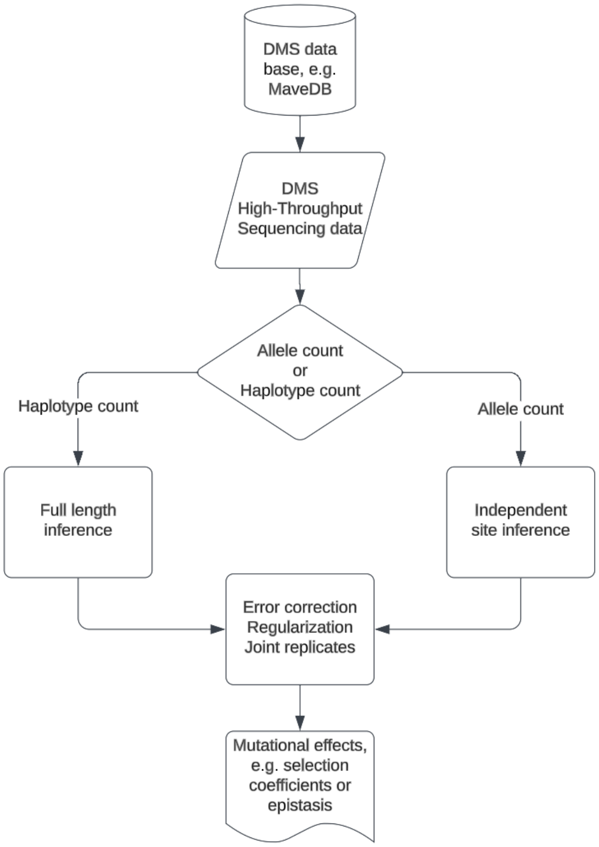
DMS data base, e.g. MaveDB
DMS High-Throughput Sequencing data
Allele count or Haplotype count
Haplotype count
Allele count
Full length inference
Independent site inference
Error correction Regularization Joint replicates
Mutational effects, e.g. selection coefficients or epistasis

Figure 3.2: popDMS pipeline workflow.

Here we introduce a new inference pipeline, popDMS, to estimate the functional effects of mutations in DMS experiments using statistical methods from population genetics. Pipeline workflow refers to Fig. 3.2. In our approach, we view rounds of phenotypic selection in experiments as analogous to rounds of reproduction in natural populations. We then use sequence data to estimate the effects of mutations on fitness in the experiments. Simulations demonstrate that estimates from popDMS are more robust to noise than alternative methods. In tests on 25 DMS data sets [3, 157, 54, 158, 23, 42, 77, 153, 5, 72, 140, 39, 107], we find that popDMS always yields higher correlations between experimental replicates than the current state-of-the-art. Data sets details seen in Table 3.1 and Table 3.2. Codes of popDMS are accessible at `https://github.com/bartonlab/paper-DMS-inference`.

Our approach uses a probabilistic model, Marginal Path Likelihood (MPL) method, enabling us to combine statistical power across multiple replicates instead of averaging the results of independent experiments. popDMS is also capable of estimating epistatic interactions between mutations when appropriate data is available. Overall, popDMS is fast, robust to noise, easily extensible, and delivers far more reliable inferences of mutational effects than existing methods.

3.5.1 Inferring mutation effects by popDMS

We model the evolution of variant frequencies in DMS experiments following population genetics. In this model, the effect of each mutation i is quantified by a selection coefficient s_i, which describes the relative advantage or disadvantage of the mutation for surviving selection in the DMS experiment. Using recently-developed computational methods[155], one can then quantify the probability of experimentally observed variant frequency change as a function of the selection coefficients and infer the coefficients that best explain the data. Our approach also includes error adjustment if error correction data provided.

We built a general pipeline, popDMS, inferring mutational effects from DMS experimental data by the popDMS method for both human and virus proteins. The inferred selection coefficients require single and double allele frequencies data for single allele selection coefficients (in addition to triple and quadruple allele frequencies for epistatic inference, if applicable). During the experiment, the counts of all genotypes or genetic variants are

Table 3.1: Bloom's lab data sets information

Reference paper	Sample name	Inference method
Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans	A/Green-winged Teal/Ohio/175/1986_PB2_CCL141 A/Green-winged Teal/Ohio/175/1986_PB2_A549	dms_tools2
Mapping mutational effects along the evolutionary landscape of HIV envelope	BG505.W6M.C2.T332N BF520.W14M.C2	dms_tools2
Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants	A/Perth/16/2009 (H3N2)	dms_tools2
Deep Mutational Scanning Comprehensively Maps How Zika Envelope Protein Mutations Affect Viral Growth and Antibody Escape	ZIKV_MR766_int_GFP/MR766	dms_tools2
Identification of HIV-1 Envelope Mutations that Enhance Entry Using Macaque CD4 and CCR5	BF520.W14M.C2_hu BF520.W14M.C2_rhm	dms_tools2
Deep Mutational Scan of the Highly Conserved Influenza A Virus M1 Matrix Protein Reveals Substantial Intrinsic Mutational Tolerance	A/WSN/1933 (H1N1) Aichi/68(H3N2)	dms_tools2
Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA	A/Aichi/2/1968(H3N2)_NP_MxA A/Aichi/2/1968(H3N2)_NP_MxAneg A/Aichi/2/1968(H3N2)_NP_MS	dms_tools
Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs	A/PR/1934 (H1N1) A/Aichi/1968 (H3N2)	dms_tools

Table 3.2: MaveDB data sets information

Reference paper	**Sample name**	**Inference method**
A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function	hYAP65 WW domain	linear regression of logarithm of enrichment ratio
Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis	T7-longE4BU	Enrich, enrichment ratio
Massively Parallel Functional Analysis of BRCA1 RING Domain Variants	BRCA1-RING_Yeast-two-hybrid-based BRCA1-RING_E3-ligase	Enrich, enrichment ratio
Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning	TpoR-MPL TpoR-MPL_S505N	Enrich2, linear regression of logarithm of enrichment ratio
Saturation editing of genomic regions by multiplex homology-directed repair	BRCA1-DBR1	logarithm of enrichment score

collected at each time point for each replicate with different experimental conditions. The counts of each genetic variant are then normalized as frequencies by sites. The selection coefficients of each variant are inferred by popDMS with the frequencies of all the variants from each generation. The selection coefficients of observed variants on all sites could be summed up to obtain the fitness of the corresponding genotype by the additive assumption of individual fitness in the evolutionary model. We also generate some visualization Figures to examine the fitness inference in detailed.

Given the observed allele frequency trajectories, we derive an analytical expression to estimate the selection coefficients and epistasis terms,

$$\hat{\mathbf{s}} = \left[\mathrm{C}_{\mathrm{int}} + \gamma I\right]^{-1} \times \left[\Delta\mathbf{x} - \mu\ \mathrm{v}_{\mathrm{int}}\right]\ .$$

$\hat{\mathbf{s}}$ is the vector of estimated selection coefficients and pairwise epistatic interaction. $\mathrm{C}_{\mathrm{int}}$ denotes the integrated covariance matrix of single and double mutant allele frequencies along the evolution. γ is a regularization parameter, and I is the identity matrix. $\Delta\mathbf{x}$ gives the single and double mutant allele frequencies difference between the last and first time points. μ is the per-locus per-generation mutation rate. $\mathrm{v}_{\mathrm{int}}$ is the integrated single and double mutant allele frequencies over time. We explain the details about each term in the next section.

In detail, to examine the evolutionary dynamics, we simulated the evolutionary trajectories of genetic variant frequencies with selection by Wright-Fisher (WF) model. We considered the evolution of N individuals with selection effects. Each individual has one sequence with length L, and it results in $M = q^L$ genotypes, where q is the genetic variant number, $q = 2$ (either 0 for wildtype or 1 for mutant) for the bi-allelic case, $q = 4$ for nucleotides (A, T, C, G) and $q = 20$ for amino acids. The bi-allelic Wrightian fitness of ath genotype f_a is defined by the additive rules with single site selection coefficients s_i and pairwise epistatic selection coefficients s_{ij} as

$$f_a = 1 + \sum_{i=1}^{L} s_i g_i^a + \sum_{i=1}^{L} \sum_{j=i+1}^{L} s_{ij} g_i^a g_j^a, \tag{3.3}$$

where g_i^a represents either 0 (wildtype) or 1(mutant). For the generation t, the genotype population frequency vector is $\mathbf{z}(t) = (z_1(t), \cdots, z_M(t))$, where $z_a(t) = n_a(t)/N$ and $n_a(t)$

is the number of individuals out of total N individuals that having the genotype a at generation t.

With WF model the probability of observing genotype frequencies $\mathbf{z}(t+1)$ at generation $t+1$ given the genotype frequency $\mathbf{z}(t)$ at generation t,

$$P\left(\mathbf{z}(t+1)\middle|\mathbf{z}(t)\right) = N! \prod_{a=1}^{M} \frac{\left(p_a(\mathbf{z}(t))\right)^{Nz_a(t+1)}}{(Nz_a(t+1))!}, \tag{3.4}$$

where

$$p_a(\mathbf{z}(t)) = \frac{z_a(t)f_a + \sum_{b\neq a}\left(\mu_{ba}f_b - \mu_{ab}f_a\right)}{\sum_{b=1}^{M} z_b(t)f_b}, \tag{3.5}$$

and μ_{ba} is the probability of genotype b mutating to genotype a. To simplify the explanation, we assume that the probability of a mutation occurring from a wild-type allele to a mutant allele is the same as the probability of a mutation occurring from a mutant allele to a WT allele. We use the symbol μ to represent this probability.

The probability of genotype frequency vector follows the transition probability between the neighbouring generations by discrete time Markov Chain with initial genotype frequency vector $\mathbf{z}(t_0)$,

$$P\left((\mathbf{z}(t_k))_{k=1}^{K} \left|\mathbf{z}(t_0)\right.\right) = \prod_{k=0}^{K-1} P\left(\mathbf{z}(t_{k+1})|\mathbf{z}(t_k)\right). \tag{3.6}$$

The allele frequency could be represented by the linear combination of genotype frequencies as following:

$$\begin{aligned} x_i(t) &= \sum_{a=1}^{M} g_i^a z_a(t), & x_{ij}(t) &= \sum_{a=1}^{M} g_i^a g_j^a z_a(t), \\ x_{ijk}(t) &= \sum_{a=1}^{M} g_i^a g_j^a g_k^a z_a(t), & x_{ijkl}(t) &= \sum_{a=1}^{M} g_i^a g_j^a g_k^a g_l^a z_a(t), \end{aligned} \tag{3.7}$$

where $x_i(t)$, $x_{ij}(t)$, $x_{ijk}(t)$ and $x_{ijkl}(t)$ are the single, double, triple, and quadruple mutant allele frequencies at site i, site-pair (i,j), site-triplet (i,j,k) and site-quartet (i,j,k,l) respectively at generation t. And the frequency vector of the single site is $\mathbf{x}(t) = (x_1(t), \cdots, x_L(t))$, while the frequency vector of pairing site is $\mathbf{x}(t) = \big(x_1(t), \cdots, x_L(t), x_{12}(t), x_{13}(t), \cdots, x_{(L-1)L}(t)\big)$ $= (x_1(t), \cdots, x_L(t), x_{L+1}(t), \cdots, x_R(t))$, where L is the sequence length and $R = L(L{+}1)/2$.

Similar to genotype level with WF model, the probability of observing allele frequencies $\mathbf{x}(t+1)$ at generation $t+1$ given the allele frequency $\mathbf{x}(t)$ at generation t,

$$P\Big(\mathbf{x}(t+1)\Big|\mathbf{x}(t)\Big) = N! \prod_{a=1}^{M} \frac{\big(p_a(\mathbf{x}(t))\big)^{Nx_a(t+1)}}{(Nx_a(t+1))!}, \tag{3.8}$$

where

$$p_a(\mathbf{x}(t)) = \frac{x_a(t) f_a}{\sum_{b=1}^{M} x_b(t) x_b} \tag{3.9}$$

The probability of having allele frequency vector $\mathbf{x}(t_k)$ follows the transition probability between the neighbouring generations by discrete time Markov Chain with initial allele frequency vector $\mathbf{x}(t_0)$,

$$P\Big((\mathbf{x}(t_k))_{k=1}^{K} \,|\mathbf{x}(t_0)\Big) = \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)\right). \tag{3.10}$$

Referring to the MPL inference framework[155], given the observed frequency path of genetic variants $(\mathbf{x}(t_0), \mathbf{x}(t_1), \cdots, \mathbf{x}(t_K))$ at generations t_k, with $k \in \{0, 1, \ldots, K\}$, we can use Bayesian inference to estimate the most possible selection coefficients corresponding to the data. We approximate (3.10) by a path integral. To begin with, the WF process is approximated by a diffusion process. This approximation enables the transition probabilities to be approximated by the transition probability density of a diffusion process, multiplied by a constant scaling term. While numerical integration techniques can be employed to solve the diffusion equations and approximate $P\left(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k)\right)$, this approach is computationally demanding and yields complex expressions that are challenging to analyze, even at the single locus level. Alternatively, we use the path integral approach, which involves discretizing the transition probability density for small time steps, to compute efficiently.

By assuming a Gaussian prior for the selection coefficients and epistasis parameters and applying the maximum a posteriori criterion, we derive an analytical expression for the estimates of the selection coefficients and epistasis terms, which are combined into a vector $\hat{\mathbf{s}}$ based on the observed allele frequency trajectories by maximizing the likelihood of the selection coefficients and epistasis interaction $\mathbf{s}$,

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; \left(\mathbf{x}(t_k)\right)_{k=0}^{K}, \theta\right) P^{\text{prior}}(\mathbf{s}), \tag{3.11}$$

where $P^{\text{prior}}(\mathbf{s})$ is the prior probability function of selection coefficient $\mathbf{s}$ with the mean as zero and the variance as $\sigma^2 > 0$ in a Gaussian form,

$$P^{\text{prior}}(\mathbf{s}) = \frac{1}{(2\pi\sigma^2)^{\frac{R}{2}}} e^{\left(-\frac{1}{2\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\right)},$$

and the likelihood function given the observed frequency path can be written as the posterior probability function of allele frequencies path given the initial allele frequencies with some other evolutionary parameters θ as following:

$$\begin{aligned} \mathfrak{L}\left(\mathbf{s}; \left(\mathbf{x}(t_k)\right)_{k=0}^{K}, \theta\right) &= P\left(\left(\mathbf{x}(t_k)\right)_{k=1}^{K} |\mathbf{x}(t_0), \theta\right) \\ &= \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k), \theta\right). \end{aligned} \tag{3.12}$$

By simplifying the maximum a posterior with the path integral, the estimator of selection coefficients could be expressed in the following way:

$$\hat{s}_e = \sum_{r=1}^{R} \left[\sum_{k=0}^{K-1} \Delta t_k \mathbf{C}(\mathbf{x}(t_k)) + \gamma I \right]_{er}^{-1} \times \left[x_r(t_K) - x_r(t_0) - \mu \sum_{k=0}^{K-1} \Delta t_k v_r(\mathbf{x}(t_k)) \right],$$

for $e = 1, ..., R$, where γ is the regularization strength and I is the identity matrix. $\mathbf{C}(\mathbf{x}(t_k))$ is the covariance matrix of mutant allele frequencies describing the covariance of the allele frequencies at generation t_k, which is symmetric and of size R×R, quantifying the speed of evolution and linkage effects:

$$C_{ij}(\mathbf{x}(t)) = \begin{cases} x_i(t)(1 - x_i(t)) & i = j \\ x_{ij}(t) - x_i(t)x_j(t) & i \neq j\,. \end{cases} \tag{3.13}$$

where i and j are the combination of single site index and pair sites index $(1, 2, ..., L$, $12, 13, ..., (L-1)L)$. And we also defined the mutational flux,

$$v_e(\mathbf{x}(t_k)) = \begin{cases} 1 - 2x_i(t_k) & 1 \leq e \leq L \\ x_i(t_k) + x_j(t_k) - 4x_{ij}(t_k) & L < e \leq R\,. \end{cases} \tag{3.14}$$

The potential adjustment by accounting for the flux-in and flux-out single and double mutant frequencies along evolution with a further possible correction to the recombination effect of double genetic variants could be included in the numerator term of selection coefficient estimation.

3.5.2 Sequencing error correction

Sequencing error correction is a critical step in genomic and molecular analysis workflows that aims to identify and rectify inaccuracies introduced during the process of

DNA or RNA sequencing. These errors can arise from various sources, and correcting them is crucial to ensure the reliability and accuracy of downstream analyses.

Most sequencing technologies have inherent limitations[161, 128, 112]. Errors can occur due to limitations in base-calling accuracy, variations in signal intensity, or chemical reactions during sequencing. And errors can be introduced during library preparation steps, including DNA fragmentation, adapter ligation, and PCR indexing. Signal intensity variations in sequencing technologies can alos lead to misinterpretations of base calls, particularly in regions with complex sequences or secondary structures.

Error correction is crucial to ensure the accuracy of subsequent analyses. Mistakes in the sequences can lead to incorrect conclusions in studies involving genetic variation, gene expression, or genome assembly. In genomic studies, identifying single-nucleotide variations (SNVs) and insertions/deletions (indels) is a common task. Accurate correction reduces false positives and negatives in variant calling, improving the reliability of genetic variant discovery. Also, corrected sequences are vital for accurate functional annotation. In studies of gene expression or epigenetic modifications, errors can mislead interpretations of biological functions. Additionally, accurate sequencing data is foundational for reproducibility. Errors can lead to irreproducible results, undermining the credibility of scientific research.

popDMS has an error correction module, which is functional when the error control sequencing data, existed. Normally, these control measures are derived from extensively sequencing a wildtype library. In this scenario, any observable mutations are presumed to result from deep sequencing inaccuracies rather than authentic mutations. These advanced methodologies calculate a parameter that gradually converges towards the preferences estimated using the straightforward ratio method when dealing with substantial counts and an absence of sequencing errors. For example in Bloom's lab data sets, the rates of sequencing errors can be quantified by control sequencing of the unmutated gene. Such error rates would be included in the frequencies collection process at each generation.

3.5.3 Regularization optimization

The covariance matrix $\mathbf{C}(\mathbf{x}(t_k))$ is expected to be calculated using allele frequencies of all individuals in the population at time t_k, while the reality is that we only have access to a sample covariance matrix that is computed from a subset of the population. As a result, our ability to accurately compute covariance matrix is limited by the effects of finite sampling. Regularization techniques are often employed to mitigate the impact of noisy input in inference algorithms. Simply, we add an weighted identity matrix by regularization magnitude to integrated allele frequency covariance matrix.

To optimize the regularization used in the inference pipeline, we applied different magnitudes of regularization when inferring the selection coefficients. We compared the correlation coefficient of selection coefficients infeerred by popDMS across different replicates with different regularization strength. The optimized regularization would be determined when the decrement of correlation coefficients of selection coefficients across replicates larger than the threshold by weakening a strong regularization to weaker ones (Fig. 3.3).

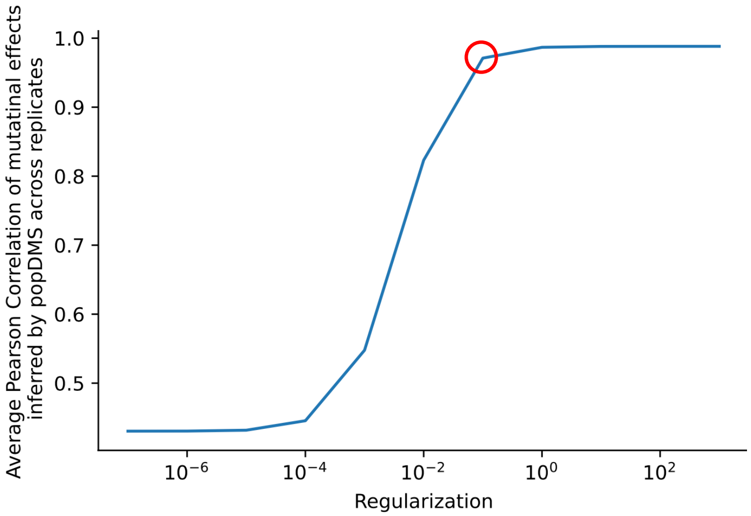

Figure 3.3: **Regularization optimization sample of TpoR data set.** popDMS will run the inference with different regularization magnitudes, depending on the least frequency of each data set. With each regularization, the pipeline will calculate the average Pearson correlation of mutational effects inferred by popDMS across replicates. Then the pipeline will choose the least regularization having the correlation drawdown less than 10% of overall correlation range. The red circle is the optimized regularization magnitude for further analysis.

3.5.4 Simulation shows the ratio methods are sensitive to sampling noise.

The popular measurement of mutational effects, enrichment ratios, has several shortcomings. One of the concerned problems is how to deal with finite sampling. Finite sampling is a common problem in statistical experiments, which also happens when the evolutionary population is not well sampled. It might cause inference inconsistency across different experimental replicates, especially for low-frequency sensitive methods. Statistically, the bias of ratio-based methods increases when the population counts are getting decrease.

The enrichment ratio takes the counts of variants out of all populations before the selection divided by the frequency of the same variants after the selection. Most of variants of the DMS experiment have less than 100 reads and even sometimes less than 10 reads comparing with more than millions of total sequence reads, which is easily affected by the stochastic noise in sequencing measurement and finite sampling. And small fluctuations in the sequencing process would affect the mutational effects inferred by the ratio-based methods for low count variants. For example,

Here, we tested the robustness of selection coefficients inference by limited sampling space and found that popDMS inference was more robust than enrichment ratios in the evolutionary simulation. To simulate the evolutionary process with finite sampling, we selected the HIV envelope mutagenesis experimental data as our sampling pool. The initial variant distribution could be re-sampled without losing generality by the Bayesian statistical inference tool, Stan [27]. The input of Stan was the variant frequency of real data, assuming it follows the multinomial distribution. Stan then output the multinomial parameters. We re-sampled the initial frequencies of variants from inferred distribution of variants by Stan.

To evaluate the performance of popDMS, we generate the simulated data by implementing the Wright-Fisher evolutionary model with discrete generations and finite-sampling in Python. In order to reproduce the finite-sampling scenario of the experimental data, we used the initial generation sequence frequency data in one of the real data sets (data set HIV Env BG505) as the input of the simulation. We assume that the target sequence is having only one site with 21 variants, including stop codon. By assuming that the allele frequency distribution at one genetic site is following the multinomial distribution, we re-sample the population distribution with Monte Carlo Markov Chain implemented by PyStan, python API of Stan package. With the initial population distribution, the population undergoes a multinomial sampling process to determine the number of variants $x_a(t)$ by the previous generation population distribution, and the drawing probability of $p_a(x_a(t))$ is represented as the following,

$$p_a(x(t)) = \frac{x_a(t-1)f_a}{\sum_{b=1}^{M} x_b(t-1)x_b} \tag{3.15}$$

where the fitness f_a is sampled by normal distribution with mean as 0 and variance as 0.1. The sampling parameters were set to the true population size as 10^8, the finite-sampling size as $5 * 10^4$, total generations as 30, the number of replicates as 100. The optimized regularization of popDMS in this simulation is 10^{-2}. We compared the correlation coefficients of mutational effects across replicates with different generations as 2, 5 and 10 generations. The counts of each variant would be randomly selected by multinomial distribution with its fitness weights at each time point(Fig. 3.4a).

We applied popDMS and ratio/regression based methods to infer the mutational effects respectively (Fig. 3.4b). The enrichment ratio is calculated by dividing the pre-selection allele frequency by the corresponding post-selection allele frequency. The logarithm scaled enrichment ratio is taking the logarithm of the enrichment ratio, slightly reducing the finite sampling problem but still being sensitive to the low-frequency data. The logarithm regression method is based on the slope of the regression by each allele frequency trajectory. Within the expectation, the ratio-based methods are less consistent across the replicates than the popDMS. With the same evolutionary period and finite sampling size, the popDMS outperforms the ratio-based methods. The ratio-based methods lack theoretical support and they are sensitive to low-frequency data. Although the logarithm regression method is more stable than the other ratio-based methods, because of the high repeatability of regression itself, the MPL is still more consistent across simulation replicates in different population size levels by the regularization control.

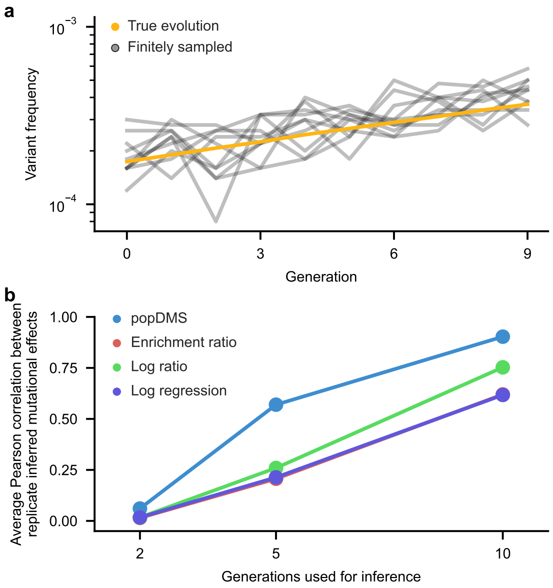

Figure 3.4: **MPL is a more consistent method than the other mutational effects inference methods in the population genetic simulation with finite sampling. a**, Allele frequency trajectories of one beneficial mutant in the simulation. The red curve is the true evolution trajectory of this mutant with large population size, while the grey curves are the trajectories of finite sampling of this mutant. The stochastic trajectories of finite sampling have high variance when apply the inference methods of mutational effects. **b**, Performance comparison of different mutational effects inference methods with finite sampling. As the generations used for inference increased, the Pearson correlation coefficients increase for all the methods, and MPL always has the higher Pearson correlation coefficients than the other inference methods.

3.5.5 Inferences by popDMS are more consistent than preferences across different replicates in real DMS experiments.

Based on the evolutionary simulation with finite sampling, the popDMS is more consistent than the enrichment ratio. The ratio based method is sensitive to the finite-sampling. The variance of the measured enrichment ratios is large for the small allele frequencies in the mutagenesis experiments, especially having the small frequencies before selection and finite-sampling. popDMS provides more consistent inference than the ratio based method by constraining the inference with optimized regularization from prior distribution. To show how consistent the different methods is, we calculate the Pearson's correlation coefficients of selection coefficients across different replicates in various data sets. To test the performance of methods with real data, we implemented our method for a total of 25 different experiments from 15 research data sets.[3, 157, 54, 158, 23, 42, 77, 153, 5, 72, 140, 39]. The popDMS only requires simple data pre-processing by providing the counts of single or multiple alleles in each generation. Among all the mutational effects inferences of a single variant, the selection coefficient inferences are more consistent than preferences across different replicates in the same experiments, regardless of the different protein kinds or diverse selection types(Fig. 3.5a). Although the regression method has higher consistency than the simple ratio form inference and it performs decently in the real data(such as BRCA1-RING and hYAP65 WW domain data sets), the popDMS has even higher correlations in these data sets. Also, the popDMS method has the regularization term to help with reducing the noise in the experiments, providing a stable factor for the popDMS inference framework. Our inference pipeline, the regularization would be automatically optimized by searching grids of different magnitudes of regularization. The regularization originally comes from the prior distribution of selection coefficients. The regularization is chosen by the criterion of “high correlation but weak regularization”. Although the regularization could help with decreasing the sampling errors, too strong regularization always underrates the real mutational effects. But regardless of how strong the regularization is, the selection coefficients of highly beneficial variants would not be inferred as deleterious and vice versa.

The fitness of the genotype is always emphasized in genetic research. Based on the underlying evolutionary model of the popDMS framework, the WF model defines the fitness of a genotype by the additive rule, such that the total fitness of one genotype is the summation over the selection coefficients of each genetic variant. In the ratio-based methods, the definition of the fitness of a genotype is the ratio between the frequency of the corresponding genotype after the selection and before the selection, still having similar drawbacks mentioned before. Additionally, the additive fitness from the WF model could also be used as the prediction of potential genotypes which are not observed in the evolution, while the ratio-based methods only output the fitness of the genotypes that already exist during the evolutionary process.

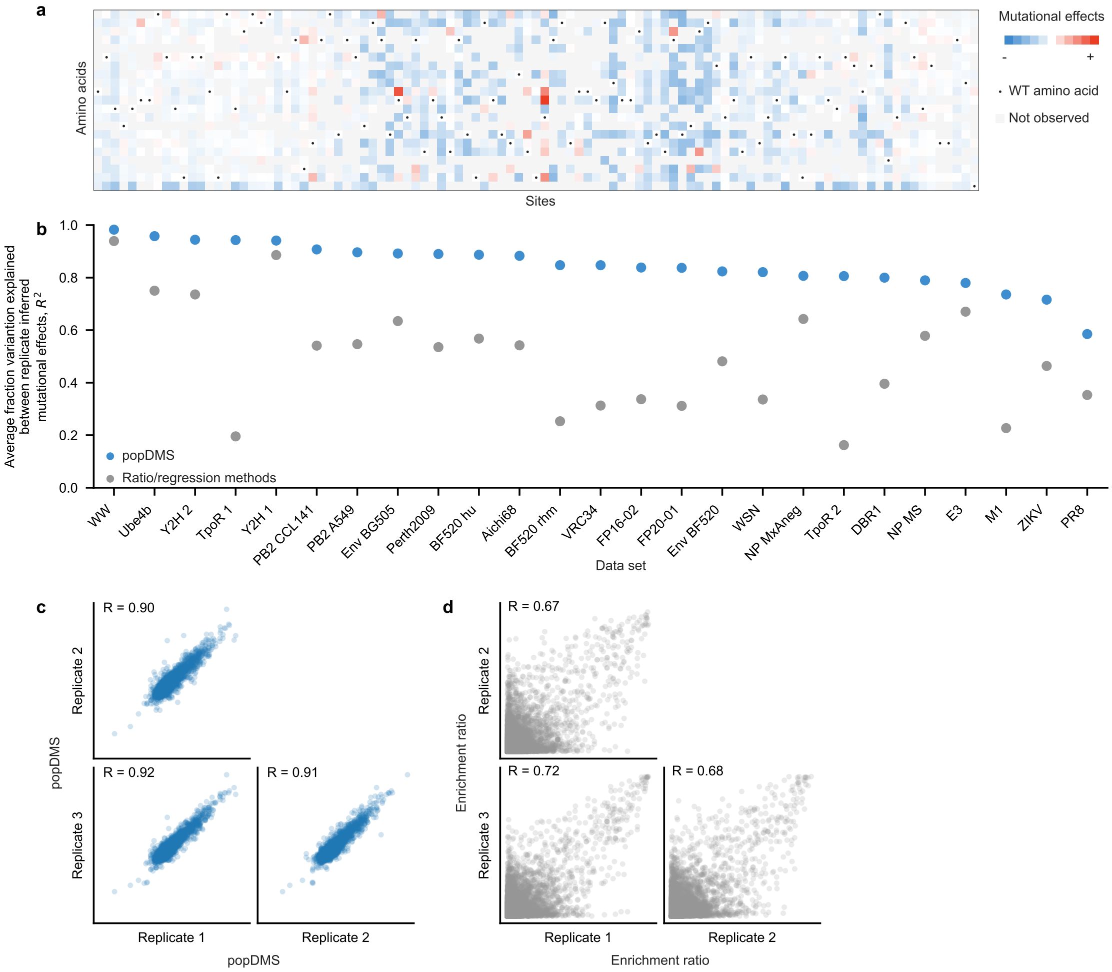

Figure 3.5: **popDMS overview.** **a**, Mutational effects heatmap example of protein T7-longE4BU. Blue, white, and red boxes represent genetic variants that were deleterious, neutral, or beneficial, respectively, during the selection process; grey represents not observed variants; and dotted entries represent the wild-type residue.**b**, Inference consistency comparison of total 25 different experiments in 15 research data sets of human and virus proteins. popDMS infers more consistent mutational effects than the ratio/regression methods within all DMS experiments. **c**, Scatter plots of HIV envelop BF520 mutational effects inferred by popDMS across three experimental replicates. The correlation coefficients are high and don't come from some outliers misleading the correlations. **d**, Scatter plots of HIV envelop BF520 mutational effects inferred by dms_tool2 package across three experimental replicates.

3.5.6 popDMS with multiple replicates

Although taking the average of preferences calculated from different replicates reduces experimental errors, popDMS can collect the observations of frequencies from different replicates to obtain more accurate mutational effects(cite to epistasis paper). Though the ratio-based methods take the average of the mutational effects across the replicates for each variant, by which the errors introduced by experimental measurements could be improved, it doesn't provide more accurate inference with more evolutionary information, only reducing the systematic errors. In contrast to ratio-based methods, combining multiple replicates data, popDMS could infer the accurate mutational effects with higher confidence than single replicate inference. Each replicate is treated as an additional "consecutive" evolutionary path of the previous evolution, which means the multiple replicates of inference by popDMS can not only reduce the errors from the experiments by extending the time-series data but also infer more reliable results with more evolutionary data.

It is natural to incorporate data from multiple independent replicates into this inference approach. These replicates may represent independent evolutionary paths with varying sampling parameters and starting conditions. And popDMS could link all the replicates together to have the joint inference with all available replicate data. Denote $t_1^q, \ldots, t_{Kq}^q$ as the sampling times of the qth replicate, and $\mathbf{x}_i^q(t_k^q)$ and $\mathbf{x}_{ij}^q(t_k^q)$ represent the frequencies vector of the single and double mutant alleles at the ith locus and (i,j)th pairs of loci at generation t_k^q. The observed trajectory of the single and double mutant allele frequencies of the qth replicate can be denoted as $\mathbf{x}_i^q(t_k^q)$=$(x_1^q(t_k^q)$, $\ldots$, $x_L^q(t_k^q)$, $x_{12}^q(t_k^q)$, $x_{13}^q(t_k^q)$, $\ldots$, $x_{(L-1)L}^q(t_k^q))$.

And the estimator of selection coefficients with multiple replicates data could be expressed in the following way:

$$
\begin{aligned}
\hat{s}_e = \sum_{r=1}^{R} & \left[\sum_{q=1}^{Q} \sum_{k=0}^{K_q-1} \Delta t_k^q \mathbf{C}(\mathbf{x}^q(t_k^q)) + \gamma I \right]_{er}^{-1} \\
\times \sum_{q=1}^{Q} & \left[x_r^q(t_{K_q}^q) - x_r^q(t_0^q) - \mu \sum_{k=0}^{K_q-1} \Delta t_k^q v_r(\mathbf{x}^q(t_k^q)) \right] ,
\end{aligned}
\tag{3.16}
$$

where Q is the number of independent replicates observed, $\Delta t_k^q = t_{k+1}^q - t_k^q$, and $\mathbf{C}(\mathbf{x}^q(t_k^q))$ is the covariance matrix of the allele frequencies at generation t_k^q of the qth replicate.

3.5.7 Main differences between popDMS and existed inference methods in single variant mutational effects inference.

After the inference of selection coefficients of each genetic variant on the genetic sequence from DMS experiments data by popDMS, we could generate some visualizations of mutational effects, such as logo-plot (Fig. 3.6a), for single variant mutational effects. The logo-plot is an alternative form of sequence logo, one graphical representation of the mutational effects of nucleotides or amino acids on genetic sequences. Logo-plot consists of several stacks of letters(A, T, C, and G for nucleotides or one-letter abbreviation of amino acids), and each stack represents each site on the sequence. The height of each letter is exponentially proportional to the mutational effect for the certain amino acids at that site and all the stacks of letters would be normalized to 1. We take the exponential form of selection coefficient $e^{\beta s_i} / \sum_j e^{\beta s_j}$ as the height of the genetic variant letter, where β is the scaling parameter. Higher letter variants mean that they have more beneficial mutational effects during evolution and lower ones are more deleterious.

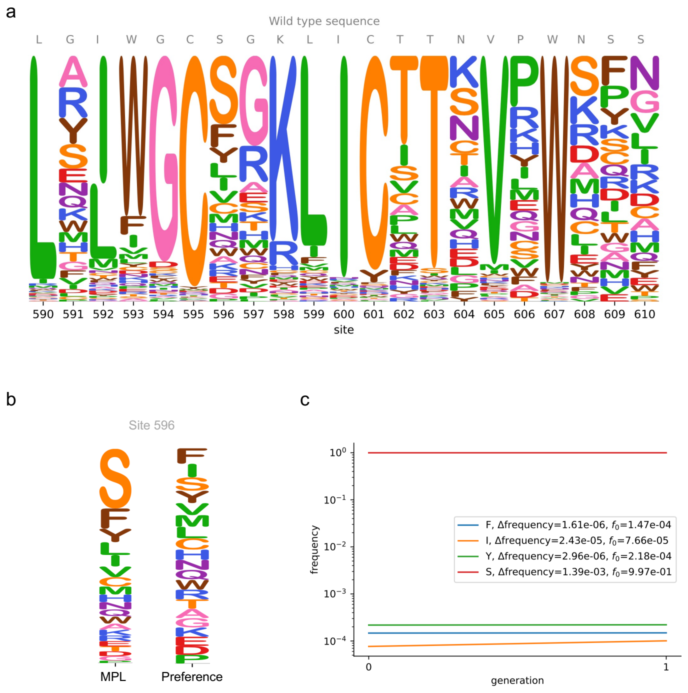

Figure 3.6: **Visualization of single variant mutational effects and comparison of single variants mutational effects inference. a,** Logo plot of mutational effects in HIV Env BG505 example. **b,** Comparison of HIV Env BG505 mutational effects inferred by popDMS and preference method. Example site is 596, left logo plot is by popDMS, and right logo plot is by preference. **c,** Example frequency trajectories of top selection coefficients inferred by popDMS at site 596. $\triangle frequency$ is the frequency change between the post-selection stage and pre-selection stage, f_0 is the initial frequency of each variant.

In the existed mutational effect inference methods, such as in the paper of WW domain [3], the mutational effects come from the frequency ratio, which is in linear scale, and such methods only care about how fast the variant grows. If one mutation has small initial frequency, it is easier to have a large preference only with a small amount of increment comparing with "large" increment of wild type variant. In contrast, popDMS cares about how much the frequency will change even the initial frequencies vary a lot of different variants. No matter how large or small the initial frequency of particular variants would be, if the variants are accumulated a lot during the evolution, the popDMS would be more possible to have the beneficial selection coefficients for these variants. But the mutaional effects inferred by ratio based method would be different for large initial frequency variants and small initial frequency variants with similar frequency changes. Smaller initial frequency will always makes the mutational effects of corresponding variants larger than the larger initial frequency variants, but the mutational effects should not be dependent on how much the initial frequencies the genetic variants have, but on how well such variants would be survive under the selection pressure, which could be indicated by the frequency change flux instead of frequency ratio trend.

In the example of HIV Env BG505 data set, we can observe how popDMS makes more sense in mutational effects inference. At the site 596 of BG505, the mutational effects of most variants at this site are similar across the inference methods, but according to the mutational effects inferred by popDMS and preference, amino acid S is the most preferred amino acid inferred by popDMS, while amino acid F is the most preferred amino acid inferred by preference (Fig. 3.6b). Although the frequency ratios of non-WT variants at site 596 are larger than wild type amino acid S because of small frequencies and considerable frequency increments, the frequency change of amino acid S is much larger than the rest non-WT variants at this site (Fig. 3.6c). More beneficial variants will have higher probability to survive under selection pressure, so the amino acid S, which is the wild type amino acid, should be the most healthy variant at site 596.

3.5.8 Inference of epistatic interactions with popDMS and consensus gauge

Although the inference of mutational effects of single alleles from the finite sampling mutagenesis is already a challenging task, popDMS can still output the mutational effects with higher consistency and accuracy. The problem becomes much more complicated when the population contains multiple simultaneously mutational alleles and these co-evolving alleles will affect the mutational effects not only by itself but also the interaction between the other variants on the sequence. Then the fitness differences between different genotypes in the population, in addition to the selection coefficients of single alleles, may also have an interaction term, or named epistatic selection coefficients.

popDMS can also measure the epistatic effects from time-series genetic sequence data[154]. With the definition of enrichment ratios, the enrichment epistasis is proportional to the difference between the preference of the double mutants and the multiplication of preferences of the single mutants.[3] However, as mentioned before, the measurement of single variants preference is unstable with the finite sampling, and the similar problem occurs when the epistatic interaction is introduced.

We compared the epistasis inference by popDMS and one regression-based inference method[3]. The popDMS epistasis are transformed by consensus gauge[138] to remove wildtype variants epistasis. The detailed consensus transformation is by following:

$$s_{i,j}(a,b) \to s_{i,j}(a,b) - s_{i,j}(c_i,b) - s_{i,j}(a,c_j) + s_{i,j}(c_i,c_j), \tag{3.17}$$

where $s_{i,j}(a,b)$ is the epistasis interaction of state a at site i and state b at site j. And c is the wild type state. After consensus gauge, all epistatic interactions with wild types become zero. And the updated epistatic interactions of non wild type variant pairs are compared with epistasis score used in current epistatic mutational effects paper [3].

3.5.9 Comparison of epistatic interaction between popDMS and functional score methods.

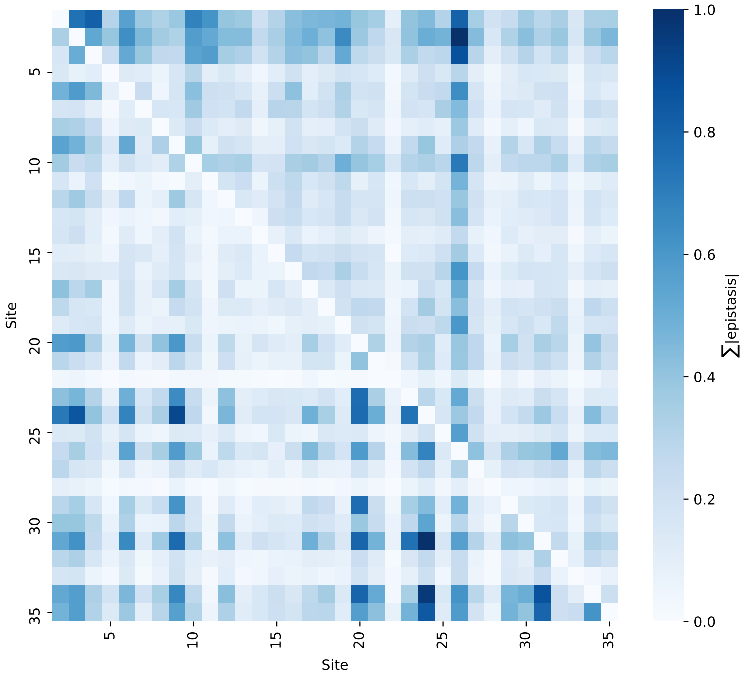

Figure 3.7: **Heatmap of epistatic strength across methods for WW domain data set.** Upper right is the absolute epistatic strength across sites inferred by ratio method, while the lower left is inferred bu popDMS. The similar pattern of strong epistatic interaction is observed in this heatmap by two methods listed. The colorbar indicates how strong the asbolute epistatic strength is.

To compare epistatic interactions inferred by popDMS and other state-of-art inference methods, we used consensus gauged epistasis inferred by popDMS and scaled functional scores by wild-type genetic background[3]. By observing the scatter plot of consensus gauge

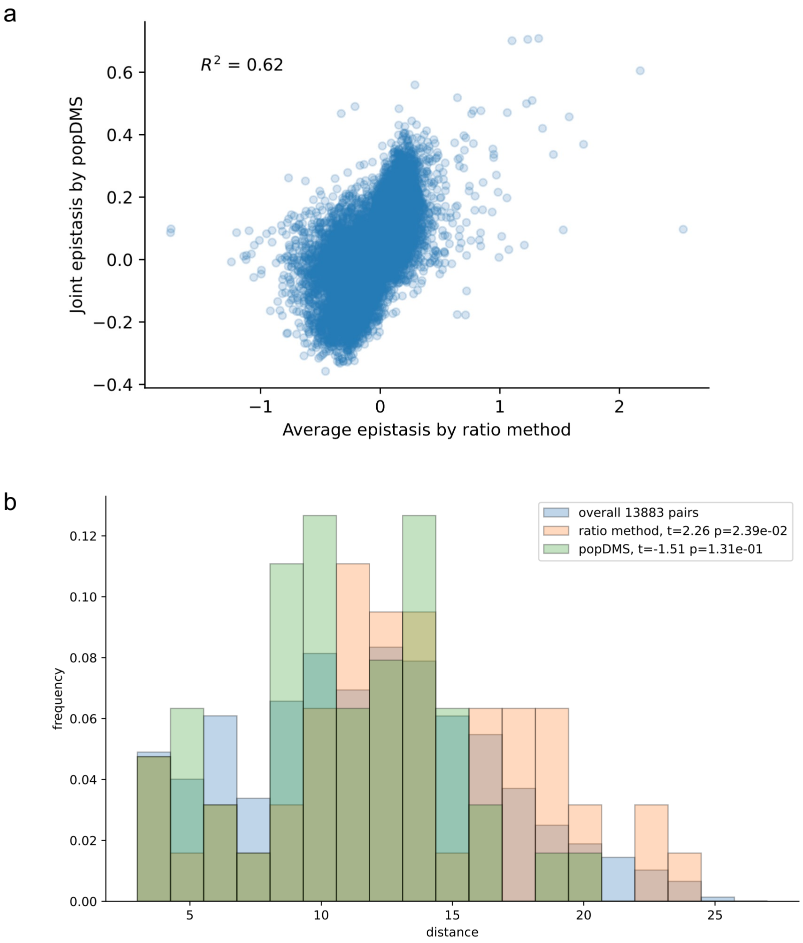

Figure 3.8: **Comparison of epistatic interaction inferred by popDMS vs. ratio method.** **a,** Scatter plot of element-wise epistasis inference by popDMS and ratio method. **b,** Histogram of pairs distance. Overall histogram is the distance distribution of all possible site pairs. The distances of popDMS and ratio method are the site distance of the site pairs with top 50 epistasis inference by different methods. The distance unit is angstrom (Å). t-statistics and p-value suggest whether the average distance of top epistasis inferred is significantly different from the average of overall pair sites distance.

epistasis from popDMS vs. epistasis score from ratio-based method, we do find a big effect from wild type genetic background.

In Fig. 3.7, the inferred epistasis by popDMS and regression-based method have some similar patterns of epistasis strength. The epistasis strength shows that how interactive the site pairs would be, both beneficial and deleterious interaction contribute positively to the epistasis strength. To check the similarities and difference between methods, the raw epistasis scatter plot is in Fig. 3.8a. By definition of two methods, the consensus gauged epistasis inferred by popDMS should be comparable with the epistasis score inferred by the ratio method[3].

Epistasis refers to the phenomenon where the effects of one gene on a trait are modified by the presence of one or more other genes. It can involve genetic sequences that are physically close to each other on the same chromosome or even on different chromosomes. The strength of epistatic interactions can vary widely and is influenced by multiple factors, including the proximity of the genetic sequences involved[159]. So we also checked the site pairs distance distribution for inferred top epistatic interactions in Fig. 3.8b. The distance between site pairs is the atomic-level coordinate distance with the unit as angstrom (Å). Commonly, closer site pairs have stronger interaction. In the distance distribution, the site pairs with top epistasis inferred by popDMS have the distance mean smaller than the overall distance mean significantly, while the regression-based method has the larger distance mean. And also, the average site distance of site pairs with top epistasis indicated by popDMS is significantly smaller than the average site distance of site pairs indicated by ratio method, with p-value as 0.0053 and t-statistics as -2.85 angstrom (Å).

3.6 Discussion

We have presented a MPL based inference pipeline, popDMS, inferring the mutational effects from DMS data. Unlike lack of theoretical support in enrichment ratio based method, MPL is originally derived from the Wright-Fisher model. By revealing the genetic linkage effect, the MPL is one more reliable method than purely empirical inference method in quantifying the mutational effects in complex evolving context. And because of its closed form calculation, popDMS is computational efficiently.

Not only in the simulation, but also in DMS experimental data sets, popDMS has higher consistency than enrichment ratio based method. In the simulation part, it is obvious that popDMS can handle with the finite sampling issue and outperform the other state of the art methods, from original enrichment ratio to logarithm regression method. The reason of the consistency comes from that popDMS utilizes the allele frequency changes and prior distribution of selection coefficients, which provide more stable factors to the inference.

In the DMS experiments, the fluctuations of allele frequencies are from the finite sampling and measurement errors. The finite sampling introduces small frequencies that makes the inference of mutational effects more sensitive when apply the ratio based methods. With allele frequency change calculation, some systematic errors could be canceled out, which is difficult to be adjusted with enrichment ratio method. If error corrected raw data provided, MPL could infer the mutational effects more accurately.

MPL takes advantages of genetic linkage in inferring mutational effects. The enrichment ratio based methods ignore the linkage effects between genetic variants, which is not the optimal way explaining the mutational effects with such complex evolutionary context, because the genetic linkage could strongly bias the inferred mutational effects. And in the epistasis inference, not only the genetic linkage between two locus in included, but also triple and quadruple locus data are taken into account, by which all the genetic linkage information is included in the mutational effects inference. By considering linkage effects and wild type genetic background, we can estimate the mutational effects more accurately than other simple ratio inference methods.

In some cases, genetic sequences that are physically close to each other on the same chromosome might exhibit stronger epistatic interactions. This could be due to the fact that physically neighboring genes are more likely to interact with each other during processes like DNA replication, transcription, and recombination. Additionally, if genes are physically close, they may be more likely to share regulatory elements or participate in common biochemical pathways, leading to greater potential for interaction. However, it's important to note that the strength of epistatic interactions is not solely determined by physical proximity. Genes located on different chromosomes or farther apart on the same chromosome can also exhibit strong epistasis if they are functionally related or involved in the same biological pathways. Epistatic interactions can be highly context-dependent and might involve complex networks of genetic and molecular interactions.

Chapter 4

Conclusions

MPL leverages genetic linkage to its advantage in mutational effect inference. Enrichment ratio methods overlook these genetic linkage effects between variants, which isn't ideal for explaining mutational effects in intricate evolutionary contexts. Genetic linkage can significantly bias inferred mutational effects, which MPL aims to address. Additionally, our approach extends to epistasis inference, considering not only linkage between two loci but also triple and quadruple locus data, thus incorporating comprehensive genetic linkage information into mutational effect inference. Furthermore, we've made the code for our method available, allowing for reproducibility and adaptation in various contexts.

Although fitness models with higher-order epistasis involving more than two mutant alleles are possible, our analysis focuses on pairwise epistasis terms. In principle, the MPL framework can be extended to account for higher-order epistasis terms by explicitly modeling the evolution of higher-order mutant allele frequencies. However, the contribution of epistasis terms to fitness typically diminishes with their order, and modeling higher-order epistasis beyond pairwise terms may offer minimal gains in some scenarios. With the help of MPL, the epistasis could be inferred with more confidence.

It is essential to note that MPL, like all inference methods, requires sufficient genetic diversity in the data to enable accurate parameter inference. In the case of a fitness model with pairwise epistasis terms, the number of model parameters to be inferred increases quadratically with the sequence length. Consequently, data with insufficient variation may result in most model parameters being partially accessible or inaccessible. However, when multiple low-diversity independent replicates are available, MPL provides a systematic approach to combine them and overcome this limitation.

In contrast to previous inference framework only taking the single selection coefficients into account, which inferred an additive fitness landscape, the current approach infers a fitness landscape with epistasis terms between every pair of mutant alleles. Model selection approaches could be considered to select the most likely fitness model from a reduced set of models with different densities of epistasis terms. However, due to the exponential increase in the number of possible models with increasing system size, model selection might only be feasible for moderate-sized systems. Alternatively, a sparsity constraint on the epistasis terms can be applied. Future work may explore the use of sparsity-inducing techniques to develop computationally efficient algorithms suitable for systems with hundreds or thousands of segregating mutations.

MPL employs a path integral to approximate the likelihood of various evolutionary parameters, including epistasis, based on observed time-series data of allele frequencies and their correlations. This framework can also be adapted for different evolutionary scenarios.

For the input data to MPL under a fitness model with pairwise epistasis terms, we use the frequencies of single, double, triple, and quadruple mutant alleles, which can be readily obtained from long-read sequencing data. However, computing double and higher mutant allele frequencies extensively for short-read data requires further development of accurate detection or inference methods. Nevertheless, with the increasing trend towards longer read lengths in third-generation sequencing technologies, higher-order mutant frequencies are expected to become more accessible in future data sets.

Regarding future directions, our approach could benefit from accounting for more complex fitness functions, as seen in recent literature. In DMS experiments, fluctuations in allele frequencies arise from finite sampling and measurement errors. MPL's allele frequency change calculation can help mitigate systematic errors that are harder to correct with enrichment ratio methods. With access to error-corrected raw data, MPL's ability to accurately infer mutational effects would be further enhanced.

Inferring mutational effects holds pivotal importance across multiple scientific domains, ranging from genetics and evolutionary biology to biomedicine and pharmaceutical research. This process involves comprehending the consequences of genetic variations, such as mutations, on an organism's traits, fitness, and overall functioning. We have introduced an MPL-based approach and inference package, popDMS, for inferring mutational effects from different resources data, addressing issues present in traditional enrichment ratio methods that lack theoretical grounding. Unlike enrichment ratio related methods, which lack support from the Wright-Fisher model, MPL is derived from this model and thus offers enhanced reliability for quantifying mutational effects in complex evolving scenarios. Thanks to its closed-form calculation, MPL is also computationally efficient. Although the computation of the linear equations is still heavy, the framework could be accelerated by some approximation algorithms.

It's worth noting that certain data sets might lack known correlations, necessitating a degree of estimation. However, our method still exhibits superior consistency to enrichment ratio methods in both simulation and DMS experimental data sets. In simulations, MPL excels at managing finite sampling challenges, surpassing other contemporary techniques such as the original enrichment ratio method and logarithm regression. This consistency stems from MPL's utilization of allele frequency changes and prior distribution of selection coefficients, providing more stable factors for inference.

Epistasis is a widespread phenomenon that significantly influences evolution. Genetic time-series data offer valuable opportunities to identify and estimate epistatic effects on fitness. Nevertheless, developing methods capable of efficiently and accurately making inferences remains a challenge. In this study, we propose a physics-based approach to tackle this challenge. Building upon a previously introduced framework for non-epistatic models, our pipeline utilizes simulations to demonstrate its accuracy in inferring pairwise epistasis effects and individual selection coefficients, provided sufficient data variation exists.

More analyses could be done with epistasis inference. Epistasis scores can help predict the phenotypic outcome of certain genetic combinations. By understanding which combinations of genetic variants lead to stronger or weaker interactions, researchers can gain insights into the underlying biological mechanisms and pathways that contribute to the trait of interest. And epistasis scores can be used to assess the combined effect of multiple genetic variants on disease risk. Some diseases are known to result from complex interactions between multiple genes. By analyzing the epistatic interactions, people can identify individuals who might be at higher risk due to specific genetic combinations. Also, in pharmacogenomics, which studies how genes affect an individual's response to drugs, epistasis scores can help identify individuals who might respond differently to a drug based on their genetic interactions. This information can guide personalized medicine approaches, leading to more effective and safer drug prescriptions.

It's important to note that interpreting and utilizing epistasis scores can be complex, especially when dealing with traits influenced by numerous genetic and environmental factors. Additionally, the reliability and accuracy of epistasis scores depend on the quality of the data and the statistical methods used to calculate them. As the field of genomics advances, the use of epistasis scores is likely to become more refined and integrated into various applications in genetics, medicine, and biology.

Bibliography

[1] Syed Faraz Ahmed, Ahmed A Quadeer, David Morales-Jimenez, and Matthew R McKay. Sub-dominant principal components inform new vaccine targets for HIV Gag. *Bioinformatics*, 35(20):3884–3889, 2019.

[2] Carlos L. Araya and Douglas M. Fowler. Deep mutational scanning: Assessing protein function on a massive scale. *Trends in Biotechnology*, 29, 2011.

[3] Carlos L. Araya, Douglas M. Fowler, Wentao Chen, Ike Muniez, Jeffery W. Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 2012.

[4] Brian J Arnold, Michael U Gutmann, Yonatan H Grad, Samuel K Sheppard, Jukka Corander, Marc Lipsitch, and William P Hanage. Weak epistasis may drive adaptation in recombining bacteria. *Genetics*, 208(3):1247–1260, 2018.

[5] Orr Ashenberg, Jai Padmakumar, Michael B. Doud, and Jesse D. Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS Pathogens*, 13, 2017.

[6] Claudia Bank, Gregory B Ewing, Anna Ferrer-Admettla, Matthieu Foll, and Jeffrey D Jensen. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, 30(12):540–546, 2014.

[7] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the National Center for Biotechnology Information. *Journal of Virology*, 82(2):596–601, 2008.

[8] Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, 461(7268):1243–1247, 2009.

[9] John P Barton, Nilu Goonetilleke, Thomas C Butler, Bruce D Walker, Andrew J McMichael, and Arup K Chakraborty. Relative rate and location of intra-host HIV

evolution to evade cellular immunity are predictable. *Nature communications*, 7:11660, 2016.

[10] N. H. Barton and J. B. Coe. On the application of statistical physics to evolutionary biology. *Journal of Theoretical Biology*, 259, 2009.

[11] Thomas Bataillon, Thomas H.G. Ezard, Michael Kopp, and Joanna Masel. Genetics of adaptation and fitness landscapes: From toy models to testable quantitative predictions. *Evolution*, 76, 2022.

[12] Michael F Berger, Eran Hodis, Timothy P Heffernan, Yonathan Lissanu Deribe, Michael S Lawrence, Alexei Protopopov, Elena Ivanova, Ian R Watson, Elizabeth Nickerson, Papia Ghosh, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*, 485(7399):502–506, 2012.

[13] Shimon Bershtein, Michal Segal, Roy Bekerman, Nobuhiko Tokuriki, and Dan S Tawfik. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929–932, 2006.

[14] Graham R Bignell, Chris D Greenman, Helen Davies, Adam P Butler, Sarah Edkins, Jenny M Andrews, Gemma Buck, Lina Chen, David Beare, Calli Latimer, et al. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–898, 2010.

[15] Collin M Blakely, Thomas BK Watkins, Wei Wu, Beatrice Gini, Jacob J Chabon, Caroline E McCoach, Nicholas McGranahan, Gareth A Wilson, Nicolai J Birkbak, Victor R Olivas, et al. Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nature Genetics*, 49(12):1693–1704, 2017.

[16] Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16, 05 2015.

[17] Jesse D Bloom, Lizhi Ian Gong, and David Baltimore. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, 328(5983):1272–1275, 2010.

[18] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of $2N_e s$ from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

[19] Benedetta Bolognesi, Andre J. Faure, Mireia Seuma, Jörn M. Schmiedel, Gian Gaetano Tartaglia, and Ben Lehner. The mutational landscape of a prion-like domain. *Nature Communications*, 10, 2019.

[20] Richard Bradshaw, Bhavesh M Patel, Edward W Tate, Robin J Leatherbarrow, and Ian R Gould. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein–protein interaction. *Protein Engineering, Design and Selection*, 24:197–207, 07 2010.

[21] Michael S Breen, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor A Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, 2012.

[22] Rachel B Brem, John D Storey, Jacqueline Whittle, and Leonid Kruglyak. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436(7051):701, 2005.

[23] Jessica L. Bridgford, Su Min Lee, Christine M.M. Lee, Paola Guglielmelli, Elisa Rumi, Daniela Pietra, Stephen Wilcox, Yash Chhabra, Alan F. Rubin, Mario Cazzola, Alessandro M. Vannucchi, Andrew J. Brooks, Matthew E. Call, and Melissa J. Call. Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood*, 135, 2020.

[24] Carlos D Bustamante, Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, Stephen Glanowski, David M Tanenbaum, Thomas J White, John J Sninsky, Ryan D Hernandez, et al. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157, 2005.

[25] Thomas C Butler, John P Barton, Mehran Kardar, and Arup K Chakraborty. Identification of drug resistance mutations in HIV from constraints on natural evolution. *Physical Review E*, 93(2):022412, 2016.

[26] Örjan Carlborg and Chris S Haley. Epistasis: Too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[27] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[28] Ashley JR Carter, Joachim Hermisson, and Thomas F Hansen. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theoretical Population Biology*, 68(3):179–196, 2005.

[29] Brian Charlesworth. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetics Research*, 63(03):213–227, 1994.

[30] Brian Charlesworth, MT Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

[31] 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.

[32] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: A key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[33] Paul D. Cotter, Lucy H. Deegan, Elaine M. Lawton, Lorraine A. Draper, Paula M. O'Connor, Colin Hill, and R. Paul Ross. Complete alanine scanning of the two-component lantibiotic lacticin 3147: generating a blueprint for rational drug design. *Molecular Microbiology*, 62:735–747, 09 2006.

[34] Vincent Dahirel, Karthik Shekhar, Florencia Pereyra, Toshiyuki Miura, Mikita Artyomov, Shiv Talsania, Todd M Allen, Marcus Altfeld, Mary Carrington, Darrell J Irvine, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences*, 108(28):11530–11535, 2011.

[35] J Arjan G M de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, 2014.

[36] J Arjan GM de Visser, Tim F Cooper, and Santiago F Elena. The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3617–3624, 2011.

[37] J Arjan GM de Visser and Santiago F Elena. The evolution of sex: Empirical insights into the roles of epistasis and drift. *Nature Reviews Genetics*, 8(2):139–149, 2007.

[38] 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.

[39] Adam S. Dingens, Priyamvada Acharya, Hugh K. Haddox, Reda Rawi, Kai Xu, Gwo Yu Chuang, Hui Wei, Baoshan Zhang, John R. Mascola, Bridget Carragher, Clinton S. Potter, Julie Overbaugh, Peter D. Kwong, and Jesse D. Bloom. Complete functional mapping of infection- and vaccine-elicited antibodies against the fusion peptide of HIV. *PLoS Pathogens*, 14, 2018.

[40] Adam S Dingens, Hugh K Haddox, Julie Overbaugh, and Jesse D Bloom. Comprehensive mapping of hiv-1 escape from a broadly neutralizing antibody. *Cell host & microbe*, 21(6):777–787, 2017.

[41] Sara Domínguez-García, Carlos García, Humberto Quesada, and Armando Caballero. Accelerated inbreeding depression suggests synergistic epistasis for deleterious mutations in Drosophila melanogaster. *Heredity*, 123(6):709–722, 2019.

[42] Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular Biology and Evolution*, 32, 2015.

[43] Michael B Doud, Juhye M Lee, and Jesse D Bloom. How single mutations affect viral escape from broad and narrow antibodies to h1 influenza hemagglutinin. *Nature communications*, 9(1):1386, 2018.

[44] Alexei J Drummond, Geoff K Nicholls, Allen G Rodrigo, and Wiremu Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

[45] Richard Durrett. *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media, 2008.

[46] Daniel Esposito, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler, and Alan F. Rubin. Mavedb: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1), 2019.

[47] Warren J Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*. Springer Science & Business Media, 2012.

[48] Maha R Farhat, B Jesse Shapiro, Karen J Kieser, Razvan Sultana, Karen R Jacobson, Thomas C Victor, Robin M Warren, Elizabeth M Streicher, Alistair Calver, Alex Sloutsky, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. *Nature Genetics*, 45(10):1183–1189, 2013.

[49] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522, 2014.

[50] Marcus W Feldman, Sarah P Otto, and Freddy B Christiansen. Population genetic perspectives on the evolution of recombination. *Annual Review of Genetics*, 30(1):261–295, 1996.

[51] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D Walker, and Arup K Chakraborty. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[52] Anna Ferrer-Admetlla, Christoph Leuenberger, Jeffrey D Jensen, and Daniel Wegmann. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. *Genetics*, 203(2):831–846, 2016.

[53] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, 33(1):268–280, 2015.

[54] Gregory M. Findlay, Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, 513, 2014.

[55] Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution*, 31, 2014.

[56] Matthieu Foll, Hyunjin Shim, and Jeffrey D Jensen. WFABC: a Wright–Fisher ABC–based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1):87–98, 2015.

[57] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–746, 2010.

[58] Douglas M. Fowler, Carlos L. Araya, Wayne Gerard, and Stanley Fields. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27, 2011.

[59] Douglas M Fowler, Carlos L Araya, Wayne Gerard, and Stanley Fields. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27(24):3430–3431, 2011.

[60] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[61] Susanne U Franssen, Viola Nolte, Ray Tobler, and Christian Schlötterer. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. *Molecular Biology and Evolution*, 32(2):495–509, 2015.

[62] Gaurav D Gaiha, Elizabeth J Rossin, Jonathan Urbach, Christian Landeros, David R Collins, Chioma Nwonu, Itai Muzhingi, Olivia M Waring, Alicja Piechocka-Trocha, Michael Waring, et al. Structural topology defines protective CD8+ T cell epitopes in the HIV proteome. *Science*, 364(6439):480–484, 2019.

[63] Molly Gasperini, Lea Starita, and Jay Shendure. The power of multiplexed functional analysis of genetic variants. *Nature protocols*, 11(10):1782–1787, 2016.

[64] Lisbeth Gauguin, Carlie Delaine, Clair L Alvino, Kerrie A McNeil, John L Wallace, Briony E Forbes, and Pierre De Meyts. Alanine scanning of a putative receptor binding surface of insulin-like growth factor-i. *Journal of Biological Chemistry*, 283:20821–20829, 07 2008.

[65] Sergey Gavrilets. *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press, 2004.

[66] Philip J Gerrish and Richard E Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144, 1998.

[67] Zachariah Gompert. Bayesian inference of selection in a heterogeneous environment from genetic time-series data. *Molecular Ecology*, 25(1):121–134, 2016.

[68] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, 2013.

[69] 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, 2017.

[70] Michael Gundry and Jan Vijg. Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 729, 2012.

[71] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to hiv's envelope protein on viral replication in cell culture. *PLoS pathogens*, 12(12):e1006114, 2016.

[72] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7:e34420, 2018.

[73] Thomas F Hansen. Why epistasis is important for selection and adaptation. *Evolution*, 67(12):3501–3511, 2013.

[74] Zhangyi He, Mark Beaumont, and Feng Yu. Effects of the ordering of natural selection and population regulation mechanisms on Wright-Fisher models. *G3: Genes, Genomes, Genetics*, 7(7):2095–2106, 2017.

[75] Ryan T Hietpas, Jeffrey D Jensen, and Daniel NA Bolon. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901, 2011.

[76] Geoffrey E. Hill. Genetic hitchhiking, mitonuclear coadaptation, and the origins of mt dna barcode gaps. *Ecology and Evolution*, 10(17):9048–9059, 2020.

[77] Nancy Hom, Lauren Gentles, Jesse D. Bloom, and Kelly K. Lee. Deep Mutational Scan of the Highly Conserved Influenza A Virus M1 Matrix Protein Reveals Substantial Intrinsic Mutational Tolerance. *Journal of Virology*, 93, 2019.

[78] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128, 2017.

[79] Mohammad Tofazzal Hossain Howlader, Yasuhiro Kagawa, Ai Miyakawa, Ayaka Yamamoto, Tetsuya Taniguchi, Tohru Hayakawa, and Hiroshi Sakai. Alanine scanning analyses of the three major loops in domain ii of bacillus thuringiensis mosquitocidal toxin cry4aa. *Applied and Environmental Microbiology*, 76:860–865, 11 2009.

[80] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert RH Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, et al. Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences*, 109(39):15553–15559, 2012.

[81] Diarmaid Hughes and Dan I Andersson. Evolutionary consequences of drug resistance: Shared principles across diverse targets and organisms. *Nature Reviews Genetics*, 16(8):459–471, 2015.

[82] Christopher JR Illingworth. Fitness inference from short-read data: Within-host evolution of a reassortant H5N1 influenza virus. *Molecular Biology and Evolution*, 32(11):3012–3026, 2015.

[83] Christopher JR Illingworth, Andrej Fischer, and Ville Mustonen. Identifying selection in the within-host evolution of influenza using viral sequence data. *PLoS Computational Biology*, 10(7):e1003755, 2014.

[84] Christopher JR Illingworth and Ville Mustonen. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*, 189(3):989–1000, 2011.

[85] 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.

[86] Christopher JR Illingworth, Leopold Parts, Stephan Schiffels, Gianni Liti, and Ville Mustonen. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular Biology and Evolution*, 29(4):1187–1197, 2011.

[87] Christopher JR Illingworth, Jayna Raghwani, David Serwadda, Nelson K Sewankambo, Merlin L Robb, Michael A Eller, Andrew R Redd, Thomas C Quinn, and Katrina A Lythgoe. A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection. *PLoS Pathogens*, 16(6):e1008171, 2020.

[88] Arya Iranmehr, Ali Akbari, Christian Schlötterer, and Vineet Bafna. CLEAR: Composition of likelihoods for evolve and resequence experiments. *Genetics*, 206(2):1011–1023, 2017.

[89] Marcus Jäger, Maria Dendle, and Jeffery W. Kelly. Sequence determinants of thermodynamic stability in a WW domain - An all--sheet protein. *Protein Science*, 18, 2009.

[90] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

[91] 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.

[92] Motoo Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[93] Joel G. Kingsolver and Douglas W. Schemske. Path analyses of selection. *Trends in Ecology Evolution*, 6:276–280, 09 1991.

[94] T. Kortemme, D. E. Kim, and D. Baker. Computational alanine scanning of protein-protein interfaces. *Science Signaling*, 2004:pl2–pl2, 02 2004.

[95] Roger D Kouyos, Olin K Silander, and Sebastian Bonhoeffer. Epistasis between deleterious mutations and the evolution of recombination. *Trends in Ecology & Evolution*, 22(6):308–315, 2007.

[96] Viktor Kožich, Jitka Sokolová, Veronika Klatovská, Jakub Krijt, Miroslav Janošík, Karel Jelínek, and Jan P. Kraus. Cystathionine -synthase mutations: Effect of mutation topology on folding and activity. *Human Mutation*, 31, 2010.

[97] Rainer Kress. *Numerical analysis*. Springer-Verlag, New York, 1998.

[98] Claus Kristensen, Thomas Kjeldsen, Finn C. Wiberg, Lauge Schäffer, Morten Hach, Svend Havelund, Joseph Bass, Donald F. Steiner, and Asser S. Andersen. Alanine scanning mutagenesis of insulin. *Journal of Biological Chemistry*, 272:12978–12983, 05 1997.

[99] Vladimir Kubyshkin and Nediljko Budisa. The alanine world model for the development of the amino acid repertoire in protein biosynthesis. *International Journal of Molecular Sciences*, 20:5507, 11 2019.

[100] Miguel Lacerda and Cathal Seoighe. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, 198:1237–1250, 2014.

[101] 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.

[102] Erik Laurini, Domenico Marson, Suzana Aulic, Maurizio Fermeglia, and Sabrina Pricl. Computational alanine scanning and structural analysis of the sars-cov-2 spike protein/angiotensin-converting enzyme 2 complex. *ACS Nano*, 14:11821–11830, 08 2020.

[103] Brian Lee, Muhammad Saqib Sohail, Elizabeth Finney, Syed Faraz Ahmed, Ahmed Abdul Quadeer, Matthew R. McKay, and John P. Barton. Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data, url: https://www.medrxiv.org/content/early/2022/01/01/2021. 12.31.21268591.full.pdf. *medRxiv*, unpublished data.

[104] Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35):E8276–E8285, 2018.

[105] Fabrice Lefèvre, Marie Hélène Rémy, and Jean Michel Masson. Alanine-stretch scanning mutagenesis: A simple and efficient method to probe protein structure and function. *Nucleic Acids Research*, 25, 1997.

[106] Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331, 2011.

[107] Chuan Li, Wenfeng Qian, Calum J Maclean, and Jianzhi Zhang. The fitness landscape of a trna gene. *Science*, 352(6287):837–840, 2016.

[108] Laurence Loewe and William G. Hill. The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365:1153–1167, 04 2010.

[109] Raymond HY Louie, Kevin J Kaczorowski, John P Barton, Arup K Chakraborty, and Matthew R McKay. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proceedings of the National Academy of Sciences*, page 201717765, 2018.

[110] Elena R Lozovsky, Rachel F Daniels, Gavin D Heffernan, David P Jacobus, and Daniel L Hartl. Relevance of higher-order epistasis in drug resistance. *Molecular Biology and Evolution*, 38(1):142–151, 2021.

[111] Helmut Lutkepohl. Handbook of matrices. *Computational Statistics and Data Analysis*, 2(25):243, 1997.

[112] Xiaotu Ma, Ying Shao, Liqing Tian, Diane A. Flasch, Heather L. Mulder, Michael N. Edmonson, Yu Liu, Xiang Chen, Scott Newman, Joy Nakitandwe, and et al. Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(1), 2019.

[113] Anna-Sapfo Malaspinas. Methods to characterize selective sweeps using time serial samples: an ancient dna perspective. *Molecular ecology*, 25(1):24–41, 2016.

[114] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

[115] Jaclyn K Mann, John P Barton, Andrew L Ferguson, Saleha Omarjee, Bruce D Walker, Arup Chakraborty, and Thumbi Ndung'u. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS computational biology*, 10(8):e1003776, 2014.

[116] Alexander Y. Maslov, Wilber Quispe-Tintaya, Tatyana Gorbacheva, Ryan R. White, and Jan Vijg. High-throughput sequencing in mutation detection: A new generation of genotoxicity tests? *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 776, 2015.

[117] Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013.

[118] Sebastian Matuszewski, Marcel E Hildebrandt, Ana-Hermina Ghenu, Jeffrey D Jensen, and Claudia Bank. A statistical guide to the design of deep mutational scanning experiments. *Genetics*, 204(1):77–87, 2016.

[119] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[120] Irina S. Moreira, Pedro A. Fernandes, and Maria J. Ramos. Computational alanine scanning mutagenesis—an improved methodological approach. *Journal of Computational Chemistry*, 28:644–654, 2006.

[121] Kim L Morrison and Gregory A Weiss. Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*, 5:302–307, 06 2001.

[122] Pablo R Murcia, Joseph Hughes, Patrizia Battista, Lucy Lloyd, Gregory J Baillie, Ricardo H Ramirez-Gonzalez, Doug Ormond, Karen Oliver, Debra Elton, Jennifer A Mumford, et al. Evolution of an eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathogens*, 8(5):e1002730, 2012.

[123] Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences*, 107(9):4248–4253, 2010.

[124] Richard A Neher and Thomas Leitner. Recombination rate and selection strength in hiv intra-patient evolution. *PLoS Computational Biology*, 6(1):e1000660, 2010.

[125] Richard A Neher and Boris I Shraiman. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16):6866–6871, 2009.

[126] Gabriele Pedruzzi, Ayuna Barlukova, and Igor M Rouzine. Evolutionary footprint of epistasis. *PLoS Computational Biology*, 14(9):e1006426, 2018.

[127] Velislava N. Petrova and Colin A. Russell. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16, 2018.

[128] Franziska Pfeiffer, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L. Schultze, and Günter Mayer. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 8(1), 2018.

[129] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[130] Martin O Pollard, Deepti Gurdasani, Alexander J Mentzer, Tarryn Porter, and Manjinder S Sandhu. Long reads: Their purpose and place. *Human Molecular Genetics*, 27(R2):R234–R241, 2018.

[131] Vadim Puller, Richard Neher, and Jan Albert. Estimating time of HIV-1 infection from next-generation sequence diversity. *PLoS Computational Biology*, 13(10):e1005775, 2017.

[132] Linqiong Qiu, Yuna Yan, Zhaoxi Sun, Jianing Song, and John Z.H. Zhang. Interaction entropy for computational alanine scanning in protein-protein binding. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8:e1342, 10 2017.

[133] Ahmed Abdul Quadeer, David Morales-Jimenez, and Matthew R McKay. Co-evolution networks of HIV/HCV are modular with direct association to structure and function. *PLoS Computational Biology*, 14(9):e1006409, 2018.

[134] T. Michael Redmond, Eugenia Poliakov, Shirley Yu, Jen Yue Tsai, Zhongjian Lu, and Susan Gentleman. Mutation of key residues of RPE65 abolishes its enzymatic role as isomerohydrolase in the visual cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2005.

[135] Nicholas Renzette, Daniel R Caffrey, Konstantin B Zeldovich, Ping Liu, Glen R Gallagher, Daniel Aiello, Alyssa J Porter, Evelyn A Kurt-Jones, Daniel N Bolon, Yu-Ping Poh, et al. Evolution of the influenza A virus genome during development of oseltamivir resistance in vitro. *Journal of Virology*, 88(1):272–281, 2014.

[136] Matthew S Rich, Celia Payen, Alan F Rubin, Giang T Ong, Monica R Sanchez, Nozomu Yachie, Maitreya J Dunham, and Stanley Fields. Comprehensive analysis of the SUL1 promoter of Saccharomyces cerevisiae. *Genetics*, 203(1):191–202, 2016.

[137] Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications.* Springer-Verlag, 2nd edition, 1989.

[138] Francesca Rizzato, Alice Coucke, Eleonora de Leonardis, John P. Barton, Jérôme Tubiana, Rémi Monasson, and Simona Cocco. Inference of compressed potts graphical models. *Physical Review E*, 101(1), 2020.

[139] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

[140] Jeremy I. Roop, Noah A. Cassidy, Adam S. Dingens, Jesse D. Bloom, and Julie Overbaugh. Identification of HIV-1 envelope mutations that enhance entry using macaque CD4 and CCR5. *Viruses*, 12, 2020.

[141] Alan F. Rubin, Hannah Gelman, Nathan Lucas, Sandra M. Bajjalieh, Anthony T. Papenfuss, Terence P. Speed, and Douglas M. Fowler. A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, 18, 08 2017.

[142] Alan F Rubin, Hannah Gelman, Nathan Lucas, Sandra M Bajjalieh, Anthony T Papenfuss, Terence P Speed, and Douglas M Fowler. A statistical framework for analyzing deep mutational scanning data. *Genome biology*, 18(1):1–15, 2017.

[143] Merijn LM Salverda, Eynat Dellus, Florien A Gorter, Alfons JM Debets, John Van Der Oost, Rolf F Hoekstra, Dan S Tawfik, and J Arjan GM de Visser. Initial mutations direct alternative pathways of protein evolution. *PLoS Genetics*, 7(3):e1001321, 2011.

[144] Joshua G Schraiber. A path integral formulation of the Wright–Fisher process with genic selection. *Theoretical Population Biology*, 92:30–35, 2014.

[145] Joshua G Schraiber, Steven N Evans, and Montgomery Slatkin. Bayesian inference of natural selection from allele frequency time series. *Genetics*, 203(1):493–511, 2016.

[146] Joshua G Schraiber, Robert C Griffiths, and Steven N Evans. Analysis and rejection sampling of wright–fisher diffusion bridges. *Theoretical population biology*, 89:64–74, 2013.

[147] Ernesto Segredo-Otero and Rafael Sanjuán. Genetic complementation fosters evolvability in complex fitness landscapes. *Scientific Reports*, 13, 2023.

[148] Guy Sella and Aaron E. Hirsh. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2005.

[149] Karen Silence, M Hartmann, Karl-Heinz Gührs, A Gase, Bernhard Schlott, D Collen, and H.R. Lijnen. Structure-function relationships in staphylokinase as revealed by "clustered charge to alanine" mutagenesis. *Journal of Biological Chemistry*, 270:27192–27198, 11 1995.

[150] Shane M. Simonsen, Lillian Sando, K. Johan Rosengren, Conan K. Wang, Michelle L. Colgrave, Norelle L. Daly, and David J. Craik. Alanine scanning mutagenesis of the prototypic cyclotide reveals a cluster of residues essential for bioactivity. *Journal of Biological Chemistry*, 283:9805–9813, 04 2008.

[151] John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(01):23–35, 1974.

[152] Paul D Sniegowski and Philip J Gerrish. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1544):1255–1263, 2010.

[153] Yq Shirleen Soh, Louise H. Moncla, Rachel Eguia, Trevor Bedford, and Jesse D. Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *eLife*, 8, 2019.

[154] Muhammad Saqib Sohail, Raymond H Louie, Zhenchen Hong, John P Barton, and Matthew R McKay. Inferring epistasis from genetic time-series data. *Molecular Biology and Evolution*, 39(10), 2022.

[155] Muhammad Saqib Sohail, Raymond HY Louie, Matthew R McKay, and John P Barton. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology*, 39(4):472–479, 2021.

[156] Lea M. Starita and Stanley Fields. Deep mutational scanning: A highly parallel method to measure the effects of mutation on protein function. *Cold Spring Harbor Protocols*, 2015, 2015.

[157] Lea M. Starita, Jonathan N. Pruneda, Russell S. Lo, Douglas M. Fowler, Helen J. Kim, Joseph B. Hiatt, Jay Shendure, Peter S. Brzovic, Stanley Fields, and Rachel E. Klevit. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 2013.

[158] Lea M. Starita, David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, 200, 2015.

[159] Tyler N. Starr and Joseph W. Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, 2016.

[160] Matthias Steinrücken, Anand Bhaskar, and Yun S Song. A novel spectral method for inferring general diploid selection from time series genetic data. *The Annals of Applied Statistics*, 8(4):2203–2222, 2014.

[161] Nicholas Stoler and Anton Nekrutenko. Sequencing error profiles of illumina sequencing instruments. *NAR Genomics and Bioinformatics*, 3(1), 2021.

[162] Natalja Strelkowa and Michael Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682, 2012.

[163] Qingling Tang and Aron W. Fenton. Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. *Human Mutation*, 38:1132–1143, 06 2017.

[164] Paula Tataru, Thomas Bataillon, and Asger Hobolth. Inference under a Wright-Fisher model using an accurate beta approximation. *Genetics*, 201(3):1133–1141, 2015.

[165] Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the Wright-Fisher model using allele frequency data. *Systematic Biology*, 66(1):e30–e46, 2017.

[166] Thomas Taus, Andreas Futschik, and Christian Schlötterer. Quantifying selection with pool-seq time series data. *Molecular Biology and Evolution*, 34(11):3023–3034, 2017.

[167] Jonathan Terhorst, Christian Schlötterer, and Yun S Song. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genetics*, 11(4):e1005069, 2015.

[168] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, 2014.

[169] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[170] Hande Topa, Ágnes Jónás, Robert Kofler, Carolin Kosiol, and Antti Honkela. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, 31(11):1762–1770, 2015.

[171] Armin Töpfer, Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel. Probabilistic inference of viral quasispecies subject to recombination. *Journal of Computational Biology*, 20(2):113–123, 2013.

[172] Harold P. De Vladar and Nicholas H. Barton. The contribution of statistical physics to evolutionary biology. *Trends in Ecology and Evolution*, 26, 2011.

[173] Michael J Wade. A gene's eye view of epistasis, selection and speciation. *Journal of Evolutionary Biology*, 15(3):337–346, 2002.

[174] Xiaoyue Wang, Audrey Q Fu, Megan E McNerney, and Kevin P White. Widespread genetic epistasis among cancer genes. *Nature Communications*, 5(1):1–10, 2014.

[175] Robert G. Webster and Elena A. Govorkova. Continuing challenges in influenza. *Annals of the New York Academy of Sciences*, 1323, 2014.

[176] Huijin Wei and Xianghua Li. Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Frontiers in Genetics*, 14, 2023.

[177] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[178] Jochen Weile and Frederick P. Roth. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics*, 137(9):665–678, 2018.

[179] Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, 2006.

[180] Daniel M Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B Heckendorn. The influence of higher-order epistasis on biological fitness landscape topography. *Journal of Statistical Physics*, 172(1):208–225, 2018.

[181] Daniel M Weinreich, Yinghong Lan, C Scott Wylie, and Robert B Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707, 2013.

[182] Daniel M Weinreich, Richard A Watson, and Lin Chao. Perspective: Sign epistasis and genetic costraint on evolutionary trajectories. *Evolution*, 59(6):1165–1174, 2005.

[183] G. A. Weiss, C. K. Watanabe, A. Zhong, A. Goddard, and S. S. Sidhu. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proceedings of the National Academy of Sciences*, 97:8950–8954, 07 2000.

[184] Robert J Woods, Jeffrey E Barrick, Tim F Cooper, Utpala Shrestha, Mark R Kauth, and Richard E Lenski. Second-order selection for evolvability in a large *Escherichia coli* population. *Science*, 331(6023):1433–1436, 2011.

[185] Sewall Wright. *Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies.* 1969.

[186] Katherine S Xue, Terry Stevens-Ayers, Angela P Campbell, Janet A Englund, Steven A Pergam, Michael Boeckh, and Jesse D Bloom. Parallel evolution of influenza across multiple spatiotemporal scales. *eLife*, 6:e26875, 2017.

[187] Lucy R Yates and Peter J Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806, 2012.

[188] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[189] Courtney M Yuen and David R Liu. Dissecting protein structure and function using directed evolution. *Nature Methods*, 4:995–997, 12 2007.

[190] Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. Population genomics of intrapatient HIV-1 evolution. *eLife*, 4:e11282, 2015.

[191] Fabio Zanini, Vadim Puller, Johanna Brodin, Jan Albert, and Richard A Neher. In vivo mutation rates and the landscape of fitness costs of HIV-1. *Virus Evolution*, 3(1):vex003, 2017.

[192] Tian-hao Zhang, Lei Dai, John P Barton, Yushen Du, Yuxiang Tan, Wenwen Pang, Arup K Chakraborty, James O Lloyd-Smith, and Ren Sun. Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genetics*, 16(10):e1009009, 2020.

[193] Tal Zinger, Maoz Gelbart, Danielle Miller, Pleuni S Pennings, and Adi Stern. Inferring population genetics parameters of evolving viruses using time-series data. *Virus Evolution*, 5(1):vez011, 2019.