NATURAL SELECTION INFERENCE AND SAMPLING FORMULAE

By

PAVEL KHROMOV

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

Written under the direction of

Alexandre V. Morozov

And approved by

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE DISSERTATION

Natural Selection Inference and Sampling Formulae

By PAVEL KHROMOV

Dissertation Director:

Alexandre V. Morozov

Theory of evolution provides a simple, flexible, and elegant framework to study and explain life around us. Population genetics is a mathematical wing of evolutionary theory. One of its key results connects steady state distribution of observable traits in a population with evolutionary parameters like mutation rate and fitness landscape in which the population is evolving. If number of trait types is very large, which is a relevant limit for molecular biology, then Ewens sampling formula provides a description of the steady state for the case with no natural selection acting on the population. We provide a generalization of this formula to arbitrary fitness landscape and use it to infer distribution of mutation rate and selection pressure along fruit fly chromosomes from sequenced data.

Acknowledgments

It is hard to accomplish something valuable. But when you are surrounded by great people this hardship turns into challenge from which you can learn, grow and even enjoy the ride. Steph Petrusz and Rob Isenhower, Matteo Turilli and Paloma Caravantes, Dima Grigoryev, Asya Zaytseva, Liza Guseva and Polina Kuznetsova, Volodya Lubyshev, Manos Maridakis, Jacek Cyranka, Triet Pham, Adrian Culver, Marina Kechkina, Max Miller, Mitya Karabash, Yana Bromberg, Jumana Dakka, Borya Radnaev, Jeremy Pronchik, Liza Muravieva, Masha Danilenko, and Bryan Leung. I would like to say thank you to all of you. Special thanks to my family Vika, Julia, Alexander, Larisa, and Galina. Ted Malliaris thanks for awesome collaboration, and finally Alex Morozov thanks for your guidance, patience and support.

Dedication

In memory of Lilia Khromova.

Loving mother and best friend.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Theory of evolution provides a simple, flexible, and elegant framework to study and explain life around us. Population genetics is a mathematical wing of evolutionary theory which lies at the center of our research. In this chapter we introduce basic notions and concepts that our results are built on.

1.1 Evolutionary forces

Population genetics studies changes in a population under three evolutionary forces: mutation, selection, and drift.

To illustrate these forces, imagine we study a population of cells. Cells compete for finite resources and divide to produce offspring. But sometime an offspring is not an exact copy of its parent. Errors do occur during this process, sometimes due to a pure chance or thanks to external influences like chemical agents or radiation. These are all example of mutations.

Broadly speaking, these errors can have three outcomes. If the changes happen in a part of DNA that is responsible for critical cell functions, then life of this new cell will be affected. This could be twofold, if the mutation breaks something in the cell's fine tuned machinery, this will inevitably lower its chances to compete for resources and divide.

We can express that by saying that fitness of the offspring is lower that of its parent's. Alternatively, as there is always room for improvement, so if the mutation happen to streamline the function of the cell, then the cell gains an edge over its peers and now it will likely to win more often when competing for resources and leave more progeny. This can be summarized by saying that the offspring fitness is higher than its parent's. Finally, third option is when the mutation happens in a region of DNA that is not responsible for any critical cell functions and we don't observe any immediate effect on cell's viability. Then we conclude that the offspring and the parent have the same fitness.

And lastly there are always random effects acting on the population of cells that remove cells from the population without paying any respect to cell's fitness. Just as piano that is about to fall on a cell does not care about cell's fitness. This is genetic drift.

1.2 Stochastic models of evolution

Pioneering attempts to mathematically model this phenomena were made by Wright [1] and Fisher [2]. Alternative but equivalent method was introduced by Moran in [3]. I will follow the latter below.

Let us go back to a population of cells. Instead of looking at the whole DNA, we consider its subsequence – a gene – that is responsible for a particular function of the cell. Then two cells either share the same subsequence or not, which means they have different variant of gene which are called alleles. To evolve this population, we make changes every discrete time step – a generation – while maintaining constant size N of the population. For the sake of establishing notations, if among these N cells we observe K allelic types, then there are n_i cells of type i and $\sum_{i=1}^{K} n_i = N$.

Let us start by considering neutral case, meaning all K alleles have same fitness. We pick an allele from the population to become a parent. As long as the alleles are equally fit, picking a parent boils down to uniform sampling, *i.e.* each allele has equal chance of $\frac{1}{N}$ to become a parent. Once we introduce fitness landscape, each allele is assigned its fitness f_i and the uniform sampling is replaced by weighted one. Now probability for i-th allele A_i to become a parent is $f_i n_i / \sum_{j=1}^{N} f_j n_j$. Once we have a parent allele A_i, there is a chance μ_{ij} that the offspring will be one of $K-1$ other alleles $A_{j \neq i}$. That leaves a chance of $1 - \sum_{j=1}^{K} \mu_{ij}$ for mutation to not occur leaving parent and offspring alleles identical. One can think of this setup by means of a graph with K vertices representing different alleles. Edges are assigned weights being mutation rates between alleles. These weights need not to be symmetric and could be zero if corresponding alleles cannot mutate into each other. If every allele can mutate to any other allele with the same rate $\mu_{ij} = \mu$ corresponding graph becomes complete. This complete graph or as we also refer to it fully-connected (FC) network case is the starting point for my research. Finally, we uniformly sample (fitness does not play role here) the population to pick an allele that we replace with the offspring.

Instead of cells we could have considered other organisms focusing on other kind of traits: say eye color for humans, beak shape for birds or wing shape for fruit flies, and study how a particular trait is represented in a given population.

1.3 Master equation

This dynamics can be described by master equation. Consider a population of N alleles of two types A and B. Let fitness of A be $f_A = 1 + s$ and $f_B = 1$. Number of A alleles $n_A = n$ is sufficient to describe the state of the system as $n_B = N - n$ at any given time

t. If there are n alleles A at time t then there are two ways to end up with $n+1$ alleles at $t+1$: First, pick one of A's to be a parent, not have a mutation when producing the offspring and replace one of B's with the offspring. The odds of that happening are

$$\frac{f_A n_A}{f_A n_A + f_B n_B}(1-\mu)\frac{n_B}{N}. \tag{1.1}$$

Second, is to pick B to be a parent, have a mutation when producing the offspring, and replace one of B's with the offspring. Probability of that is

$$\frac{f_B n_B}{f_A n_A + f_B n_B}\mu\frac{n_B}{N}. \tag{1.2}$$

Therefore transition probability $\mathbb{P}[n+1|n]$ is simply the sum of 1.1 and 1.2. Similarly, to end up with $n-1$ alleles of A at $t+1$ while having n at t we either pick one of A's to be a parent, mutate the offspring, and replace one of A's with the offspring or pick one of B's to be a parent, avoid mutation and replace one of A's with the offspring. Then

$$\mathbb{P}[n-1|n] = \frac{f_A n_A}{f_A n_A + f_B n_B}\mu\frac{n_A}{N} + \frac{f_B n_B}{f_A n_A + f_B n_B}(1-\mu)\frac{n_A}{N} \tag{1.3}$$

Finally, as long as we can't increase or decrease n_A by more than one

$$\mathbb{P}[n|n] = 1 - \mathbb{P}[n+1|n] - \mathbb{P}[n-1|n]. \tag{1.4}$$

If $p(n,t)$ is a probability to have n alleles of type A at time t then the master equation is

$$p(n, t+\Delta t) = \sum_{n'} \mathbb{P}[n|n']p(n'|t) \tag{1.5}$$

which we can rewrite in gain minus loss form

$$p(n, t+\Delta t) - p(n,t) = \sum_{n'\neq n} \mathbb{P}[n|n']p(n'|t) - \sum_{n'\neq n} \mathbb{P}[n'|n]p(n|t). \tag{1.6}$$

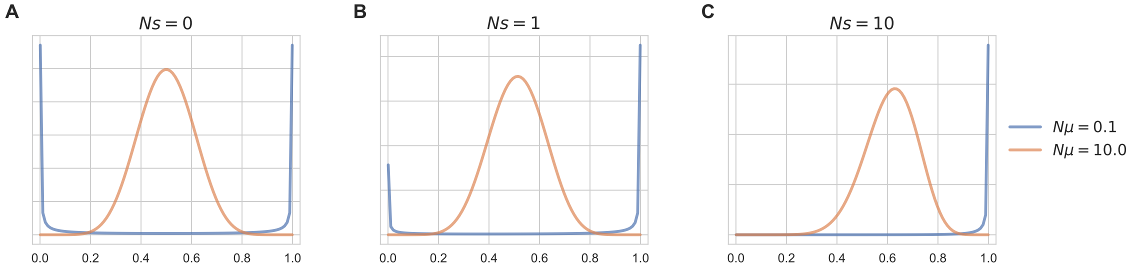

Figure 1.1: **Steady state allele frequency distributions.** Panels A through C represent increase in selection pressure. Blue line shows distribution for monomorphic population while orange for polymorphic.

1.4 Diffusion limit

Now we consider large populations while assuming $N\mu$ and Ns_i finite. It makes sense to introduce allele frequencies $x = n/N$. Expanding 1.6 in Δt and $\Delta x = 1/N$ and keeping only leading terms in $1/N$ and scaling $\Delta t = 2/N^2$ we arrive at Fokker-Plank equation

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}(Vp) - \frac{\partial}{\partial x}(Mp). \tag{1.7}$$

Here M is mean allele frequency change Δx

$$M = \mathbb{E}[\Delta x] = \Delta x \mathbb{P}[x|x+\Delta x] + (-\Delta x)\mathbb{P}[x|x-\Delta x] = N\mu\frac{1-2x}{2} + \frac{N}{2}x(1-x)\frac{\partial\langle f\rangle}{\partial x} \tag{1.8}$$

and V is its variance

$$V = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = x(1-x) \tag{1.9}$$

with $\langle f\rangle = \sum_{i=1}^{N} f_i x_i$ being mean population fitness.

Setting time derivative in (1.7) to zero we get a steady state allele frequency distribution [4, 5]

$$p \sim e^{Nsx}[x(1-x)]^{N\mu-1}. \tag{1.10}$$

Examples of this distribution are shown on Fig. 1.1. We see that in case of $\theta = 0.1$ the population tends to be in either of two states $x = 0$ or $x = 1$ which corresponds to the whole population being made up of just a single allele. This limit is called monomorphic. In the neutral case $Ns = 0$ (panel A) both states have the same probability but if we start introducing selection pressure $Ns = 1$ and $Ns = 10$ (panels B and C) we see the state corresponding to the highest fitness has progressively higher chance to be observed. Now if we consider higher mutation rate $N\mu = 10$ we see that the population consisting of a mix of different alleles. This limit is called polymorphic. In the neutral case (panel A) the mean of the distribution corresponds to the situation when both alleles are equally represented in the population. But when we increases selection (panels B and C) the mean shifts towards state with higher fitness hence alleles with higher fitness make up more then half of the population on average.

In case of K alleles we are dealing with a probability density $p(x_1, \ldots, x_K) = p(\mathbf{x})$ where all frequencies x_i are normalized $\sum_{i=1}^{K} x_i = 1$. If we pick x_K to be a dependent variable corresponding Fokker-Plank equation is [4, 5]

$$\frac{\partial p}{\partial t} = \frac{1}{2} \sum_{i,j=1}^{K-1} \frac{\partial^2}{\partial x_i \partial x_j} (V_{ij} p) - \sum_{i=1}^{K-1} \frac{\partial}{\partial x_i} (M_i p) \tag{1.11}$$

where mean

$$M_i = \mathbb{E}[\Delta x_i] = \frac{\epsilon - K\epsilon x_i}{2} + \frac{N}{2} x_i \left[\frac{\partial \langle f \rangle}{\partial x_i} - \sum_{j=1}^{K} x_j \frac{\partial \langle f \rangle}{\partial x_j} \right] \tag{1.12}$$

with $\epsilon = N\mu/(K-1)$ and covariance matrix

$$V_{ij} = \mathbb{E}[\Delta x_i \Delta x_j] - \mathbb{E}[\Delta x_i]\mathbb{E}[\Delta x_j] = \begin{cases} x_i(1 - x_i) & i = j, \\ x_i x_j & i \neq j \end{cases} . \tag{1.13}$$

Steady state solution to (1.11) is [4, 5]

$$p(\mathbf{x}) \sim e^{N\langle f\rangle} \prod_{i=1}^{K} x_i^{\epsilon-1} \tag{1.14}$$

up to a normalization constant. We can use this distribution to predict steady state frequencies given evolutionary parameters and the other way around, given steady state frequencies we could infer evolutionary parameters.

1.5 Infinite allele limit

It turns out though we cannot describe large K systems using (1.14) as it was shown this distribution becomes ill defined in infinite allele limit [6, 7]. Intuitively it could be seen by calculating the means of steady state frequencies go to zero as $\mathbb{E}[x_i] \sim 1/K$. This limit is not a mathematical abstraction, in fact it has direct connection to molecular biology. Consider a chunk of DNA of size 100 nucleotides. This could be a small gene for example (most genes span way more than 100 nucleotides but we will stick to a conservative estimate). As each nucleotide could be either of four amino acids A, C, T or G then there is at most $4^{100} \sim 10^{60}$ possible variants of this sequence. Hence one needs to find appropriate degrees of freedom to describe this limit and find the distribution of these degrees of freedom. This was done by Ewens [8] for neutral populations. We will spend subsequent chapters introducing this distribution, generalizing it to arbitrary fitness landscape, mastering fast calculation of this generalized distribution, and finally developing a method to infer evolutionary parameters with a help of this distribution and applying it to fruit fly sequencing data.

Chapter 2

Sampling Formula for Arbitrary Fitness Landscape

With the advent of high-throughput molecular biology techniques, it has recently become possible to carry out large-scale genotype-phenotype assays in molecular systems [9, 10, 11, 12, 13]. For example, Podgornaia and Laub have recently mapped all $20^4 = 1.6 \times 10^5$ possible combinations of four key residues in the *E. coli* protein kinase PhoQ, and assayed each mutant for the signaling function mediated by its binding partner PhoP [9]. This study revealed 1659 functional PhoQ variants, which can be thought of as forming the upper plane on the fitness landscape; all non-functional variants form the lower plane. The upper plane is divided into several clusters under single-point amino acid or nucleotide mutations – sequences within each cluster can mutate into each other through neutral substitutions only. The two-plane landscape is epistatic – the effect of a given mutation depends on the amino acids at the other three positions, in agreement with previous reports on the major role of epistasis in molecular evolution [14, 15, 16, 17].

The picture of a "coarse-grained" fitness landscape stratified into several distinct phenotypes is in agreement with other recent high-throughput experiments aimed at elucidating the relationship between sequence and function [15, 18, 19, 10, 11, 12]. Although these experiments typically yield continuous distributions of selection coefficients, the distributions tend to be bi-modal, with one peak corresponding to strongly deleterious and lethal mutations and another to weakly deleterious and neutral ones [20, 21, 22]. These observations suggest stratifying the fitness landscape into functional and non-functional phenotypes; intermediate fitness states such as those corresponding to weakly deleterious phenotypes can be added if necessary to refine the picture.

Overall, given the astronomically large number of alleles, the typical size of neutrally-connected clusters of sequences can be assumed to be much larger than the population size. Then evolutionary dynamics on a multiple-plane landscape will be characterized by mutation-selection-drift balance [1, 23, 24, 5, 25, 26, 27, 28] in the infinite-allele limit. At steady state, population statistics, such as the mean and the variance of the number of distinct alleles or the probability of observing a given pattern of allelic diversity in a sample of sequences, do not change anymore, even though the population continues to explore new alleles through mutation [28]. In the absence of selection, the steady-state allele sampling probability was derived by Ewens [8]. The Ewens sampling formula can be used to understand allelic diversity in neutral populations and to test for deviations from the neutral expectation; [29] its essential limitation is that, essentially, each allele is allowed to mutate into every other allele [28]. The Ewens formula arises naturally in many sampling problems in biological and physical sciences [30, 31, 32]. However, in order to understand molecular evolution in the presence of selection and make quantitative predictions of selection coefficients, it is necessary to extend it to more general fitness distributions.

Previous work in this area has focused mostly on the symmetric overdominance model, first analyzed in this context by Watterson [5, 33]. This is a diploid model in which all heterozygotes have the same selective advantage over all homozygotes, such that the mean population fitness depends on the square of allele frequencies. Since the sampling formula for this model is challenging to evaluate and therefore has never been used in practical calculations, subsequent work in the field focused on various approximations to the exact result, which require additional assumptions such as weak selection [5] or large sample sizes [34]. In particular, Joyce and collaborators have discussed asymptotic properties of the sampling distributions under a model of selection with multiple fitness states [35, 36], as well as the symmetric overdominance model [37]. More recently, Watterson's model of selection was generalized by Handa [38] and Huillet [39], who considered mean population fitness involving allele frequencies raised to the arbitrary power $q \geq 1$. They obtained sampling probabilities expressed in terms of multi-dimensional integrals which would be difficult to employ in practical calculations. In any event, only the $q = 1$ (neutral evolution) and $q = 2$ (symmetric overdominance) cases appear to have biological meaning.

Furthermore, Ethier and Kurtz have studied allelic diversity in a general model of selection in which fitness of each new allele is a symmetric function of the allelic states of its two parents, focusing on the proofs of existence and uniqueness of a steady state in the infinite-allele limit. [40, 41] Desai et al. have investigated sampling probabilities in a model (previously introduced by Charlesworth et al. [42] and Hudson and Kaplan [43]) based on a sequence of neutral and negatively selected sites [44]. This model has no interactions between sites, and therefore can be treated using the Poisson Random Field approach [45]. Since molecular evolution is characterized by prominent epistasis and correlated fitness values between parents and their offspring, the approach of Desai et al. cannot be applied to genomic data without careful numerical analysis of all the approximations involved.

Finally, several prior publications have focused on steady-state population statistics other than sampling probabilities. In particular, Li used the steady-state approach to obtain the frequency spectrum for a general landscape, and derived expressions for the mean number of alleles in a sample, as well as the mean and the variance of heterozygosity [25, 26, 27]. Ewens and Li derived frequency spectra for landscapes with two and three distinct fitness states and used them to compute the mean number of distinct alleles and the mean heterozygosity [46]. Griffiths derived a general integral expression for the frequency spectrum in a genic selection model [47].

Here we demonstrate an extension of the Ewens sampling formula to arbitrary fitness landscapes with genic selection. First, we follow previous work [1, 23, 24, 5, 25, 26, 27, 28] in assuming that the population adopts a steady state characterized by mutation-selection-drift balance. The steady state depends on the mean population fitness, which involves a linear combination of gene frequencies. Next, we derive a general sampling formula valid for any mutation rate μ, population size N, sample size $n \ll N$, and the number of alleles K with arbitrary fitness. We find that the most general sampling formula is difficult to employ in numerical calculations with large finite values of K, but small values of K and the infinite K limit are more manageable. Here we focus on the infinite-allele ($K \to \infty$) approximation with several phenotypic states, inspired by recent high-throughput molecular evolution experiments [15, 18, 19, 9, 10, 11, 12, 13]. We have developed a numerical technique based on the efficient calculation of Bell polynomials, which is distinct from previous efforts to compute sampling probabilities [48, 49]. Our approach enables us to study selection signatures and deviations from neutrality on landscapes with arbitrary fitness distributions.

We contrast our predictions with the effective population size approximation [42, 44].

We also compare our results with explicit simulations, using the Moran population genetics model [3] with single-point mutations as a benchmark against which the accuracy of the "full connectivity" assumption is checked. Finally, we investigate the limitations of the infinite-allele assumption. Our results are applicable to understanding the nature of allelic diversity under selection, mutation and drift. Moreover, our sampling formulas can form a basis of a quantitative, numerically feasible test for detecting the presence of selection and estimating its strength in evolving populations. Population-level allele diversity data are made increasingly available through high-throughput sequencing techniques, making our approach a practical and timely tool for studying the role of selection in evolution – a topic of much current interest and debate [22, 50, 51, 52, 53, 54, 55].

2.1 Sampling probability with selection

We consider a haploid population of fixed size N (our results also hold for diploid populations, as long as fitness values are assigned to individual genes rather than organisms). Each organism in the population is represented by a single allele in the state i, with fitness f_i; there are K distinct allelic states. Mutations occur with a probability μ per generation, changing the original allele into one of the $K-1$ remaining alleles. Thus the probability of offspring A_j produced by parent $A_{i \neq j}$ is $\mu/(K-1)$ (note that our approach can be easily generalized to the case of final-state-dependent mutation rates: $\mu_{ij} = \mu_j$, $\forall i$ in $A_i \rightarrow A_j$). We can view this system as an "allelic network" with the topology of a complete graph, with K vertices representing allelic states and edges representing mutational moves. Stochastic evolution of the population can then be described using Moran [3, 56] or Wright-Fisher [1, 56] models of population dynamics.

The steady-state distribution of allelic frequencies for these models is given by [1, 23, 24, 25, 5]

$$p(\mathbf{x}) = \frac{1}{Z} e^{N\langle f\rangle} \prod_{i=1}^{K} x_i^{\epsilon-1}, \tag{2.1}$$

where $\mathbf{x} = (x_1, \ldots, x_K)$ is a vector of allelic frequencies, $\epsilon = \theta/(K-1)$ with $\theta = N\mu$ for Moran and $\theta = 2N\mu$ for Write-Fisher models correspondingly, $\langle f\rangle = \sum_{i=1}^{K} f_i x_i$ is mean population fitness, and Z is a normalization constant.

In many situations relevant to molecular evolution, the number of alleles K is much larger than the population size N. In this case, the steady state in terms of allele frequencies is unlikely to be reached on relevant evolutionary time scales. Mathematically, the $K \to \infty$ limit of Eq. 2.1 becomes ill-defined [6, 7]. Nonetheless, the steady state is well-defined in terms of allelic *counts* rather than frequencies of specific alleles [28]. In other words, the allelic diversity of the population (e.g. as characterized by the mean and the variance of the distribution of the number of distinct allelic types) is tractable and will no longer change in steady state, although new alleles will continue to be explored through mutation.

Since only a subset of the entire population is typically available for analysis, we shall focus on the probabilities of allelic counts in samples of size $n \ll N$. To introduce the concept of allelic counts, let us for a moment consider a finite number of allelic types, e.g. $K = 5$, and call the corresponding alleles A, B, C, D, E. Suppose that we take a sample of $n = 4$ alleles from the population and we first observe allele A, then C, then A again, and finally D. We can record this sequence of alleles as an ordered list (A, C, A, D). However, typically we are not interested in the order in which alleles appear in the sample, and therefore record the result as an unordered list $\{A, A, C, D\}$, which shows that allele A has appeared twice and alleles C and D have appeared once each. Here we have used the notation $\{a, b, \ldots, z\}$ for unordered lists ($\{a, b, \ldots, z\} = \{b, a, \ldots, z\}$), and $(a, b, \ldots, z)$ for ordered lists ($(a, b, \ldots, z) \neq (b, a, \ldots, z)$).

Alternatively, we can record non-zero allelic counts, which yields $n_A = 2, n_C = 1, n_D = 1$. Finally, we can dispense with the allele labels altogether, identifying each allele in the sample as either new or already seen. In this case, we are left with an unordered list of counts $\{2, 1, 1\}$, meaning that we have observed 4 alleles of 3 different types, with one type represented by two alleles and the other two types by one each. In general, we will refer to $\mathbf{n} = \{n_1, \ldots, n_k\}$ as the sample configuration or the allelic counts. An equivalent representation would be to use a histogram which records how many groups of j identical alleles occur in the sample, with j ranging from 1 to n. In our example, there is one group of two identical alleles and two groups of one allele each, so that (A, C, A, D) is recorded as the allelic histogram $(a_1 = 2, a_2 = 1, a_3 = 0, a_4 = 0)$. All results in the paper are presented in terms of the counts $\{n_1, \ldots, n_k\}$ rather than the histogram $(a_1, \ldots, a_n)$.

It turns out that the allelic counts are appropriate variables in the infinite allele limit. The celebrated Ewens sampling formula [28, 8] expresses the probability of observing a particular sample configuration $\mathbf{n}$ in the absence of selection:

$$\mathbb{P}[\mathbf{n}] = N_P \frac{1}{k!} \frac{n!}{\prod_{i=1}^{k} n_i} \frac{\theta^k}{\theta^{(n)}}. \tag{2.2}$$

where N_P is the total number of distinct permutations of the allelic counts, and $\theta^{(n)} = \theta(\theta + 1) \ldots (\theta + n - 1)$ is the rising factorial.

Following an approach developed by Watterson [5], we generalize the Ewens sampling formula to the case of multiple fitness states. We define $\boldsymbol{\gamma}$, a vector whose components, γ_m, are fractions of all alleles with fitness f_m. Allowing m to range from 1 to M ($\sum_{m=1}^{M} \gamma_m = 1$) results in a landscape with $M \ll K$ distinct fitness states. Unless $\gamma_m \sim 1/K$, there is an infinite number of alleles with the same fitness, so that the landscape looks like M fitness planes interconnected through mutations. For this reason we shall often refer to phenotypic states as fitness planes and to the fitness landscape as the multiple-plane landscape.

2.1.1 Allele frequency distribution

Eq. 2.1 can be rewritten as follows:

$$p(\mathbf{x}) = \frac{1}{B(\boldsymbol{\epsilon})\mathcal{F}(\boldsymbol{\epsilon}; |\boldsymbol{\epsilon}|; \boldsymbol{\beta})} \prod_{i=1}^{K} x_i^{\epsilon-1} e^{\beta_i x_i}, \tag{2.3}$$

where $\boldsymbol{\epsilon} = (\epsilon, \ldots, \epsilon)$ is a K-dimensional vector of rescaled mutation rates, $|\boldsymbol{\epsilon}| = K\epsilon \simeq \theta$ is the L_1-norm of $\boldsymbol{\epsilon}$,

$$B(\mathbf{a}) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{K} a_i)} \tag{2.4}$$

is the generalized beta function, and

$$\mathcal{F}(\mathbf{a}; b; \mathbf{z}) = \sum_{j_1=0}^{\infty} \cdots \sum_{j_K=0}^{\infty} \frac{a_1^{(j_1)} \ldots a_K^{(j_K)}}{b^{(j_1+\ldots+j_K)}} \frac{z_1^{j_1}}{j_1!} \cdots \frac{z_K^{j_K}}{j_K!} = \sum_{j=0}^{\infty} \frac{B_j(\alpha_1, \ldots, \alpha_j)}{j! b^{(j)}} \tag{2.5}$$

is a generalization of the confluent hypergeometric function ${}_1F_1(a; b; z)$ to vector arguments. Here, $a^{(j)} = \Gamma(a+j)/\Gamma(a)$ is the rising factorial, B_j is the j-th complete Bell polynomial, and $\alpha_j = (j-1)! \sum_{i=1}^{n} a_i z_i^j$. To obtain Eq. 2.3, we have used the following result for integrating over the $(K-1)$-dimensional simplex Σ_{K-1}:

$$\int_{\Sigma_{K-1}} \prod_{i=1}^{K} x_i^{\nu_i - 1} dx_i = \frac{\prod_{i=1}^{K} \Gamma(\nu_i)}{\Gamma(\sum_{i=1}^{K} \nu_i)}. \tag{2.6}$$

A $(K-1)$-dimensional simplex Σ_{K-1} is a subspace of $\mathbb{R}^K : (x_1, \ldots, x_K) \in [0,1]^K$ which satisfies $\sum_{i=1}^{K} x_i = 1$. We have expanded the exponent in Eq. 2.1 in a Taylor series and applied Eq. 2.6 to each term in the resulting expansion.

2.1.2 Strongly monomorphic limit

In this limit the mutation rate tends to zero while the population size is kept fixed, $\epsilon \to 0$ [56, 57, 58, 59]. Consider the Fourier transform of the steady-state distribution in Eq. 2.3:

$$\tilde{p}(\mathbf{k}) = \int_{\Sigma_{K-1}} d\mathbf{x}\, e^{i\mathbf{k}\cdot\mathbf{x}} p(\mathbf{x}), \tag{2.7}$$

where the integral is over the $(K-1)$-dimensional simplex. Using Eq. 2.5, we can write the Fourier transform as a ratio of two generalized hypergeometric functions:

$$\tilde{p}(\mathbf{k}) = \frac{\mathcal{F}(\boldsymbol{\epsilon}; |\boldsymbol{\epsilon}|; \boldsymbol{\beta} + i\mathbf{k})}{\mathcal{F}(\boldsymbol{\epsilon}; |\boldsymbol{\epsilon}|; \boldsymbol{\beta})}. \tag{2.8}$$

Taking the $\epsilon \to 0$ limit yields

$$\tilde{p}_{\text{mono}}(\mathbf{k}) = \frac{\sum_{m=1}^{K} e^{\beta_m + ik_m}}{\sum_{m=1}^{K} e^{\beta_m}}. \tag{2.9}$$

Thus the steady-state distribution in the monomorphic limit is given by:

$$p_{\text{mono}}(\mathbf{x}) = \int \frac{d\mathbf{x}}{\text{Vol}(\Sigma_{K-1})} e^{-i\mathbf{k}\cdot\mathbf{x}} \tilde{p}_{\text{mono}}(\mathbf{k}) = \frac{\sum_{m=1}^{K} e^{\beta_m} \delta(\mathbf{x} - \mathbf{1}_m)}{\sum_{m=1}^{K} e^{\beta_m}}, \tag{2.10}$$

where $\text{Vol}(\Sigma_{K-1}) = \sqrt{K}/(K-1)!$ is the volume of the $(K-1)$-dimensional simplex and $(\mathbf{1}_m)_i = \delta_{mi}$. The population resides in one of the K monomorphic states available to it, with the probability of being in a particular state exponentially weighted by its fitness [60, 61, 62].

2.1.3 Probability of a sample of alleles

In this section we derive the sampling probability when the number of alleles K is finite. Let us find the probability $\mathbb{P}[\mathbf{n}]$ of observing counts $\mathbf{n} = \{n_1, \ldots, n_k\}$, assuming that the population has reached steady state in terms of its allelic diversity. Before considering general case, we illustrate our approach using an example with only $K = 3$ allelic types: $\mathcal{A} = (A, B, C)$. We wish to calculate the probability of observing counts $\{2, 1\}$ in a sample of size $n = 3$, which is assumed to be much less than the population size N. There are 18 samples that contribute to this counts:

$$\begin{array}{ccc} AAB & ABA & BAA \\ AAC & ACA & CAA \\ BBC & BCB & CBB \end{array}$$

$$\begin{array}{ccc} ABB & BAB & BBA \\ ACC & CAC & CCA \\ BCC & CBC & CCB \end{array}$$

The probability of choosing A first, then A again and finally B is

$$\begin{aligned} \mathbb{P}[(A, A, B)] &= \int x_A^2 x_B^1 \; p(x_A, x_B, x_C) \; dx_A dx_B dx_C \\ &= \int x_A^2 x_B^1 \; p(x_A, x_B) \; dx_A dx_B, \end{aligned} \tag{2.11}$$

where $p(x_A, x_B, x_C)$ is given by Eq. 2.3. Consequently, the probability of observing two A's and one B in *any* order is given by [5]

$$\mathbb{P}[\{A, A, B\}] = \binom{3}{2\;1} \int x_A^2 x_B^1 \; p(x_A, x_B) \; dx_A dx_B, \tag{2.12}$$

where $\binom{3}{2\;1}$ is the multinomial coefficient. Introducing a set $S_2\mathcal{A} = \{(A, B), (A, C), (B, C)\}$, which permutes allelic identities in an ordered manner (i.e., the overall allele ordering from A to B to C is preserved in each pair of alleles), we can take into account the first 9 configurations in the table above:

$$\mathbb{P}[\{A,A,B\}] + \mathbb{P}[\{A,A,C\}] + \mathbb{P}[\{B,B,C\}] = \binom{3}{2\ 1} \sum_{\sigma \in S_2\mathcal{A}} \int x_{\sigma_1}^2 x_{\sigma_2}^1 \; p(x_{\sigma_1}, x_{\sigma_2}) \; dx_{\sigma_1} dx_{\sigma_2}. \tag{2.13}$$

In order to include 9 remaining configurations in the table, we need to switch the order of the alleles: $\{(A,B),(A,C),(B,C)\} \to \{(B,A),(C,A),(C,B)\}$. But switching the alleles in each pair amounts to replacing $x_{\sigma_1}^2 x_{\sigma_2}^1$ with $x_{\sigma_2}^2 x_{\sigma_1}^1 = x_{\sigma_1}^1 x_{\sigma_2}^2$ in Eq. 2.13. Thus we can summarize the entire table by introducing a set $P(n_1, \ldots, n_k)$ of all distinct permutations of the counts $\{n_1, \ldots, n_k\}$, which determine the powers to which the allelic frequencies are raised in Eq. 2.13. In our example $P(2,1) = \{(2,1),(1,2)\}$. Therefore,

$$\begin{aligned} \mathbb{P}[\{2,1\}] &= \binom{3}{2\ 1} \sum_{\nu \in P(2,1)} \sum_{\sigma \in S_2\mathcal{A}} \int x_{\sigma_1}^{\nu_1} x_{\sigma_2}^{\nu_2} \; p(x_{\sigma_1}, x_{\sigma_2}) \; dx_{\sigma_1} dx_{\sigma_2} && (2.14) \\ &= \binom{3}{2\ 1} \sum_{\nu \in P(2,1)} \sum_{\sigma \in S_2\mathcal{A}} \mathbb{E}\left[\prod_{i=1}^{2} x_{\sigma_i}^{\nu_i}\right]. && (2.15) \end{aligned}$$

The above example can be easily generalized to describe the probability $\mathbb{P}[\{n_1, \ldots, n_k\}]$ of observing arbitrary counts. To do so, we enumerate all K alleles, forming a unique ordered list $\mathcal{A} = (1, \ldots, K)$. Second, we choose a subset $\sigma = (\sigma_1, \ldots, \sigma_k)$ of size k from $\mathcal{A}$ without replacement, so that the allelic order is preserved: $\sigma_1 < \ldots < \sigma_k$ (note that no subsets are allowed to contain repeating elements of $\mathcal{A}$). Then $S_k\mathcal{A}$ can be naturally defined as a set which contains all ordered subsets of $\mathcal{A}$ of size k. Finally, as before $P(\mathbf{n})$ is a set of all distinct permutations of allelic counts. Following these steps we have

$$\mathbb{P}[\mathbf{n}] = \binom{n}{n_1 \ldots n_k} \sum_{\nu \in P(\mathbf{n})} \sum_{\sigma \in S_k\mathcal{A}} \mathbb{E}\left[\prod_{i=1}^{k} x_{\sigma_i}^{\nu_i}\right], \tag{2.16}$$

where the expectation is calculated with respect to the steady-state allele distribution, Eq. 2.3.

We can use sampling probability (Eq. 2.16) to compute the distribution of the number of different allelic types k:

$$\mathbb{P}[k] = \sum_{\substack{n_1 \geq \ldots \geq n_k \\ n_1 + \ldots + n_k = n}} \mathbb{P}[\mathbf{n}], \tag{2.17}$$

where the summation runs over all ordered partitions of n into k positive integers.

2.1.4 Generalized sampling formula

As Eq. 2.16 demonstrates, evaluation of sample probabilities requires calculation of moments of allele frequency distributions. This could be done by taking derivatives of the normalization constant $Z = B(\boldsymbol{\epsilon})\mathcal{F}(\boldsymbol{\epsilon}; |\boldsymbol{\epsilon}|; \boldsymbol{\beta})$ in Eq. 2.3 with respect to the corresponding components of $\boldsymbol{\beta}$:

$$\mathbb{E}\left[\prod_{i=1}^{k} x_i^{\nu_i}\right] = \frac{1}{Z}\prod_{i=1}^{k}\left(\frac{\partial}{\partial \beta_i}\right)^{\nu_i} Z. \tag{2.18}$$

Then Eq. 2.16 takes the form

$$\mathbb{P}[\mathbf{n}] = \binom{n}{n_1 \ldots n_k} \frac{\prod_{i=1}^{k} \epsilon^{(n_i)}}{(K\epsilon)^{(n)}} \sum_{\nu \in P(\mathbf{n})} \sum_{\sigma \in S_k \mathcal{A}} \frac{\mathcal{F}(\boldsymbol{\epsilon} + \boldsymbol{\nu}_\sigma; K\epsilon + n; \boldsymbol{\beta})}{\mathcal{F}(\boldsymbol{\epsilon}; K\epsilon; \boldsymbol{\beta})}, \tag{2.19}$$

where $\boldsymbol{\nu}_\sigma$ is a K-dimensional vector whose σ_i-th components are ν_i with $i = 1, \ldots, k$ and all the other components are zero. Here, we have used the fact that differentiating Eq. 2.5 with respect to $\mathbf{z}$ yields a simple result similar to that known for the regular confluent hypergeometric function:

$$\prod_{i=1}^{k}\left(\frac{\partial}{\partial z_i}\right)^{n_i} \mathcal{F}(\mathbf{a}; b; \mathbf{z}) = \frac{\prod_{i=1}^{k} (a_i)^{(n_i)}}{b^{(n)}} \mathcal{F}\left(\mathbf{a} + \sum_{i=1}^{k} n_i \mathbf{1}_i; b + n; \mathbf{z}\right),$$

where $n = \sum_{i=1}^{k} n_i$ and $(\mathbf{1}_i)_j = \delta_{ij}$. As discussed above, the sum over σ extends over all distinct subsets of k alleles sampled from K uniquely ordered alleles and subject to the $\sigma_1 < \ldots < \sigma_k$ constraint. Therefore $\boldsymbol{\nu}_\sigma$ has $K - k$ zero and k non-zero components which are distributed according to σ. The sum over ν extends over all distinct permutations of allelic counts which sum up to n. Eq. 2.19 is valid for an arbitrary fitness landscape and an arbitrary number of alleles K.

2.1.5 Neutral limit of the sampling formula

When all alleles have the same fitness, the general sampling formula given by Eq. 2.19 should reduce to the Ewens formula for neutral evolutionary dynamics [28, 8]. Indeed, with all β_i set to zero, the generalized hypergeometric function $\mathcal{F}(\mathbf{a}; b; \mathbf{0})$ (Eq. 2.5) becomes 1. Then for the finite number of alleles K

$$\mathbb{P}[\mathbf{n}] = N_P \frac{n!}{(K\epsilon)^{(n)}} \binom{K}{k} \prod_{i=1}^{k} \frac{\epsilon^{(n_i)}}{n_i!}, \tag{2.20}$$

where $N_P = |P(\mathbf{n})|$ is the total number of distinct permutations of allelic counts. In the limit of an infinite number of alleles $K \to \infty$, Eq. 2.20 reduces to Eq. 2.2. Changing variables to allelic histogram counts yields $\prod_{i=1}^{k} n_i = \prod_{j=1}^{n} j^{a_j}$ and $N_P = k!/\prod_{j=1}^{n} a_j!$, resulting in

$$\mathbb{P}[(a_1, \ldots, a_n)] = \frac{n!}{\prod_{j=1}^{n} a_j! j^{a_j}} \frac{\theta^k}{\theta^{(n)}}. \tag{2.21}$$

Eq. 2.21 is a standard form of the Ewens sampling formula [28, 8].

2.1.6 Sampling formula for a population with two fitness states

As a straightforward generalization of the neutral case, consider a system with I alleles of fitness f_2 and $K - I$ alleles with fitness $f_1 > f_2$. Thus the fitness landscape consists of two interconnected “planes”. We can assume without loss of generality that alleles 1 through I belong to the lower plane and alleles $I + 1$ through K belong to the higher plane. Then

$\gamma = I/K$ defines a fraction of nodes on the lower plane and the fitness vector is

$$\boldsymbol{\beta} = (\underbrace{\beta, ..., \beta}_{I}, \underbrace{0, ..., 0}_{K-I}), \tag{2.22}$$

with I non-zero entries followed by $K - I$ zeros, and $\beta = -Ns$. If the first i counts come from the lower plane and the other $k - i$ counts come from the upper plane, we have

$$\boldsymbol{\nu}^Y = (\underbrace{\overbrace{\nu_1, ..., \nu_i}^{i}, 0, ..., 0}_{I}, \underbrace{\overbrace{\nu_{i+1}, ..., \nu_k}^{k-i}, 0, ..., 0}_{K-I}), \tag{2.23}$$

plus all alternative assignments of the first i counts within the first I entries of $\boldsymbol{\nu}^Y$, and the remaining $k - i$ counts within the last $K - I$ entries of $\boldsymbol{\nu}^Y$, such that the original order of the non-zero count entries is not changed. In this case, the generalized hypergeometric function reduces to the confluent hypergeometric function:

$$\mathcal{F}(\boldsymbol{\epsilon} + \boldsymbol{\nu}^Y; |\boldsymbol{\epsilon}| + n; \boldsymbol{\beta}) = {}_1F_1(\gamma\theta + \sum_{m=1}^{i} \nu_m; \theta + n; \beta). \tag{2.24}$$

Then for finite K the sampling probability is given by:

$$\mathbb{P}[\mathbf{n}] = \binom{n}{n_1 \dots n_k} \frac{\prod_{i=1}^{k} \epsilon^{(n_i)}}{(K\epsilon)^{(n)}} \binom{K}{k} \sum_{\nu \in P(\mathbf{n})} \sum_{i=0}^{k} \frac{{}_1F_1\left(\gamma\theta + \sum_{m=1}^{i} \nu_m; \theta + n; \beta\right)}{{}_1F_1\left(\gamma\theta; \theta; \beta\right)} \frac{\binom{I}{i}\binom{K-I}{k-i}}{\binom{K}{k}}. \tag{2.25}$$

Here, the $\binom{I}{i}$ and $\binom{K-I}{k-i}$ binomial factors are due to assigning non-zero counts to alternative positions within $\boldsymbol{\nu}^Y$, as described above. Taking the infinite allele ($K \to \infty$) limit with γ fixed, we arrive at

$$\mathbb{P}[\mathbf{n}] = \frac{n!}{k!} \frac{1}{\prod_{i=1}^{k} n_i} \frac{\theta^k}{\theta^{(n)}} \sum_{\nu \in P(\mathbf{n})} \sum_{i=0}^{k} \frac{{}_1F_1\left(\gamma\theta + \sum_{m=1}^{i} \nu_m; \theta + n; \beta\right)}{{}_1F_1\left(\gamma\theta; \theta; \beta\right)} \binom{k}{i} \gamma^i (1-\gamma)^{k-i}. \tag{2.26}$$

Thus hypergeometric sampling of Eq. 2.25 reduces to binomial sampling in the infinite-allele limit.

2.1.7 Sampling formula for a population with multiple fitness states

Let us now generalize the result of the previous section to the case of multiple fitness states: each allele can be assigned a distinct fitness value f_m, $m = 1, \ldots, M$. In other words, the fitness landscape consists of multiple planes, with $I_m = \gamma_m K$ nodes of fitness f_m on the m-th plane, so that $\sum_{m=1}^{M} \gamma_m = 1$. Then the sampling probability for finite K is given by

$$\mathbb{P}[\mathbf{n}] = \binom{n}{n_1 \ldots n_k} \frac{\prod_{i=1}^{k} \epsilon^{(n_i)}}{(K\epsilon)^{(n)}} \binom{K}{k} \sum_{\nu \in P(\mathbf{n})} \sum_{Y \in \mathcal{Y}(\mathbf{I},\mathbf{n})} \frac{\mathcal{F}(\boldsymbol{\gamma}\theta + \boldsymbol{\nu}^{Y}; \theta + n; \boldsymbol{\beta})}{\mathcal{F}(\boldsymbol{\gamma}\theta; \theta; \boldsymbol{\beta})} \frac{\binom{I_1}{i_1} \cdots \binom{I_M}{i_M}}{\binom{K}{k}}, \tag{2.27}$$

and its infinite allele limit is given by

$$\mathbb{P}[\mathbf{n}] = \frac{n!}{k!} \frac{1}{\prod_{i=1}^{k} n_i} \frac{\theta^k}{\theta^{(n)}} \sum_{\nu \in P(\mathbf{n})} \sum_{Y \in \mathcal{Y}(\mathbf{n})} \frac{\mathcal{F}(\boldsymbol{\gamma}\theta + \boldsymbol{\nu}^{Y}; \theta + n; \boldsymbol{\beta})}{\mathcal{F}(\boldsymbol{\gamma}\theta; \theta; \boldsymbol{\beta})} \binom{k}{i_1 \ldots i_M} \gamma_1^{i_1} \cdots \gamma_M^{i_M}. \tag{2.28}$$

The sums in Eqs. 2.27 and 2.28 take into account all possible ways of sampling n alleles from M planes (Fig 2.1). To explain these sums, let us imagine distributing n books over M shelves. The books come in k indivisible volume sets, and the i-th set has ν_i identical books in it. We would like to find all book-to-shelf arrangements, keeping in mind that shelves have finite capacities: only I_m books can be placed on the m-th shelf. One way to describe any book-to-shelf arrangement is to use an M-dimensional vector $\boldsymbol{\nu}^Y$ which records how many books are placed on each shelf. For example, if $M > k$, a vector $\boldsymbol{\nu}^Y = (\nu_1, \ldots, \nu_k, 0, \ldots, 0)$ with $M - k$ zeros following k non-zero entries describes placing volume sets on shelves in a particular order: the first volume set goes on the first shelf, the second volume on the second shelf and so on (assuming that the shelves are large enough to accommodate the volume sets), until no more books are left, so that the remaining $M - k$ shelves remain empty. Permutations of this arrangement, expressed as permutations of $\boldsymbol{\nu}^Y$

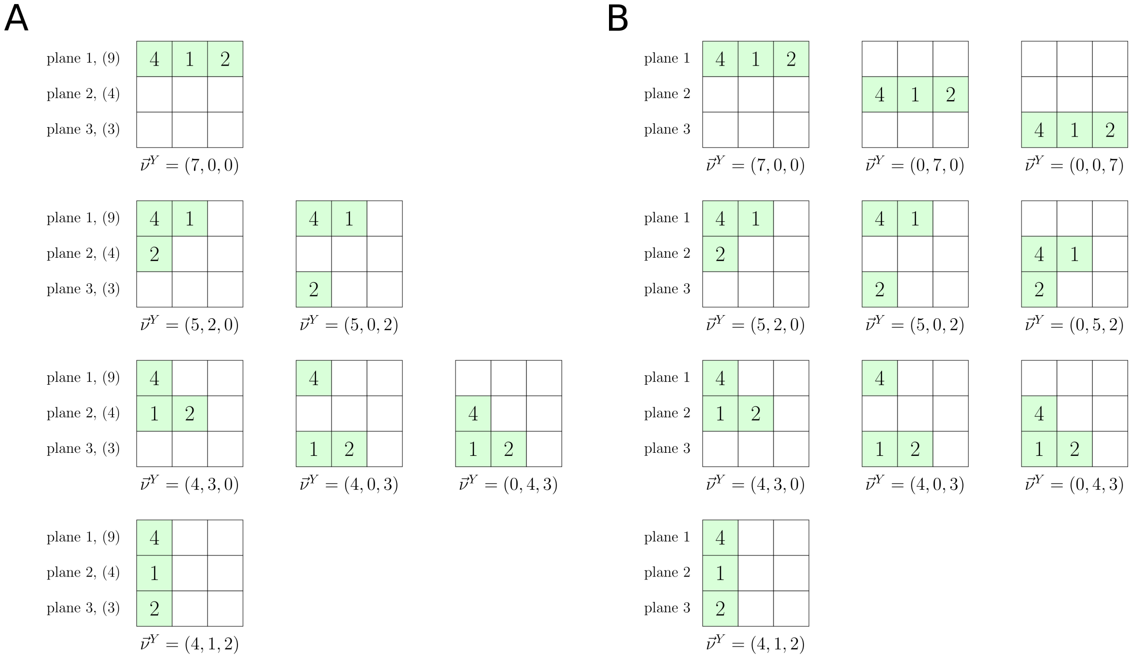

Figure 2.1: **Summations in the sampling formula for a population with multiple fitness states.** Illustration of summations over $\mathcal{Y}(\mathbf{I}, \mathbf{n})$ and $\mathcal{Y}(\mathbf{n})$ in Eqs. 2.27 and 2.28 respectively, for a list of allelic counts $\mathbf{n} = \{4, 1, 2\}$. (A) The finite plane case. Finite plane capacities are shown in parentheses. (B) The infinite plane case.

vector elements, are also allowed (again, assuming that all the shelves are large enough). We can also put more than one volume set on a single shelf, leading to arrangements such as $(\nu_1 + \nu_2, \nu_3, \ldots, \nu_k, 0, \ldots, 0)$ with $M - k + 1$ zero and $k - 1$ non-zero entries. As before, this arrangement is allowed only if the number of books on each shelf does not exceed shelf capacities. Note that the question of capacity does not arise in the infinite allele limit, since the shelves become effectively infinitely long.

In order to systematically list all the arrangements for volume sets $(\nu_1, \ldots, \nu_k)$, we follow a simple rule: if the k-th set of ν_k books is placed on the m-th shelf, the $(k+1)$-th set of ν_{k+1} books goes either on the same shelf or on the m'-th shelf with $m' > m$.

Taking elements of $(\nu_1, \ldots, \nu_k)$ one by one and changing the initial shelf (onto which the 1st volume set is placed) and the number of volume sets on each shelf, we can generate a set of all permutations of $\boldsymbol{\nu}^Y$ elements. We shall call this set $\mathcal{Y}(\mathbf{I}, \mathbf{n})$ since it depends on both the shelf capacities $\mathbf{I} = (I_1, \ldots, I_M)$ and the volume sets $\mathbf{n}$. In the limit of infinite shelf capacity the dependence on shelf sizes disappears, and the set of all permutations will be called $\mathcal{Y}(\mathbf{n})$. To include all possible arrangements, we need to perform the book-placing procedure for each distinct permutation of $\mathbf{n}$.

Now, if we replace shelves with fitness planes and volume sets with allelic counts, we obtain an algorithm for generating all allowed placements of allelic counts on fitness planes. The non-negative indices $i_1, \ldots, i_M$ in Eqs. 2.27 and 2.28 represent the number of volume sets (allelic counts) on each shelf (fitness plane). The distribution of alleles among fitness planes of finite capacity is illustrated in Fig 2.1A for $M = 3$ and a vector of allelic counts $\boldsymbol{\nu} = (4, 1, 2)$; the infinite-plane case is shown in Fig 2.1B.

Next, let us consider the monomorphic limit of Eq. 2.28. It can be shown that

$$\mathcal{F}(\theta\boldsymbol{\gamma}; \theta; \boldsymbol{\beta}) \xrightarrow[\theta \to 0]{} \sum_{m=1}^{M} \gamma_m e^{\beta_m}, \tag{2.29}$$

leading to

$$\begin{aligned} \mathbb{P}[\{n\}] &= 1 + O(\theta), \\ \mathbb{P}[\{n_1, \ldots, n_k\}] &= O(\theta^{k-1}). \end{aligned} \tag{2.30}$$

Therefore, as expected, the $\mathbb{P}[\{n\}]$ $(k = 1)$ term dominates in the monomorphic limit.

By construction, Eq. 2.28 reduces to the neutral limit, Eq. 2.2, when all fitness values are the same. In addition, the neutral limit is reproduced in the strongly polymorphic limit

$$\mathcal{F}(\boldsymbol{\gamma}\theta + \boldsymbol{\nu}_Y; \theta + n; \boldsymbol{\beta}) \xrightarrow[\theta \to \infty]{} \mathcal{F}(\boldsymbol{\gamma}\theta; \theta; \boldsymbol{\beta}), \tag{2.31}$$

and Eq. 2.28 reduces to the neutral result. This is expected since selection effects become vanishingly small in this regime.

2.1.8 Efficient evaluation of sampling probabilities

To evaluate sampling probabilities, we need to compute $\mathcal{F}(\mathbf{a}, b, \mathbf{z})$ (Eq. 2.5) efficiently. The calculation of $\mathcal{F}(\mathbf{a}, b, \mathbf{z})$ is performed by filling a square matrix with the partial Bell polynomials $B_{n,k}$, from which complete Bell polynomials can be calculated from the rows as $B_n = \sum_{k=1}^{n} B_{n,k}$. We use the following convolution identity: $(\mathbf{x} \diamondsuit \mathbf{y})_n = \sum_{j=1}^{n-1} \binom{n}{j} x_j y_{n-j}$. Note that the identity is commutative, i.e. $(\mathbf{x} \diamondsuit \mathbf{y})_n = (\mathbf{y} \diamondsuit \mathbf{x})_n$, and that the summation limits are such that the convolution of two vectors with nonzero elements will always have a zero as its first element. Let $\mathbf{x}^{k\diamondsuit}$ denote the vector that results when $\mathbf{x}$ is convolved with itself k times. The convolution matrix C is lower triangular and has the vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ as its leftmost column, $\mathbf{x}^{2\diamondsuit}$ as the second leftmost, etc. Partial Bell polynomials can then be calculated as:

$$B_{n,k}(x_1, \ldots, x_{n-k+1}) = \frac{(\mathbf{x}^{k\diamondsuit})_n}{k!} = \frac{C_{n,k}}{k!}. \tag{2.32}$$

The matrix elements $C_{n,k}$ can be calculated starting from the top of the matrix, left-to-right within each row. The sum in Eq. 2.5 runs over complete Bell polynomials in ascending order, so that convergence can be checked after the completion of each row. We specify a relative precision, e.g. $\tilde{\epsilon} = 10^{-12}$, and terminate the computation of $\mathcal{F}$ once the contribution of the current term j is small enough compared to the partial sum from 0 to $j-1$: $|\mathcal{F}_j / \mathcal{F}_{\text{partial}}| < \tilde{\epsilon}$.

Our main result is the following expression for the sampling probability (details of the derivation are available in Materials and Methods):

$$\begin{aligned}\mathbb{P}[\mathbf{n}] =& \frac{n!}{k!}\frac{1}{\prod_{i=1}^{k} n_i}\frac{\theta^k}{\theta^{(n)}}\times \\ & \sum_{\nu\in P(\mathbf{n})}\sum_{Y\in\mathcal{Y}(\mathbf{n})}\frac{\mathcal{F}(\boldsymbol{\gamma}\theta+\boldsymbol{\nu}^Y;\theta+n;\boldsymbol{\beta})}{\mathcal{F}(\boldsymbol{\gamma}\theta;\theta;\boldsymbol{\beta})}\binom{k}{i_1\dots i_M}\gamma_1^{i_1}\dots\gamma_M^{i_M}.\end{aligned} \tag{2.33}$$

Here, $\mathcal{F}(\mathbf{a};b;\mathbf{z})$ is a generalization of the confluent hypergeometric function ${}_1F_1(a;b;z)$ to vector arguments. The double sum in Eq. 2.33 takes into account all possible ways of assigning observed allelic counts $\mathbf{n}$ to M fitness planes; $\boldsymbol{\nu}^Y$ is an auxiliary vector which encodes these assignments Fig 2.1. Each assignment contributes differently to the final expression due to the non-trivial fitness landscape. The fitness values are stored in the vector $\boldsymbol{\beta}$, whose components are fitness differences $\beta_m = N(f_m - f_1)$ scaled by the population size N. For example, in the case of two fitness states $\beta_1 = 0$ and $\beta_2 = N(f_2 - f_1) = Ns$, where s is the selection coefficient. Finally, $i_1 \dots i_M$ indicate the number of distinct allelic types sampled from the corresponding fitness plane ($\sum_{m=1}^{M} i_m = k$).

The first line in Eq. 2.33 is simply the Ewens formula (Eq. 2.2) without N_P, which is the value returned by the double sum on the second line when all fitness values are equal. The version of the sampling formula with selection (Eq. 2.33) suitable for a finite number of alleles K is provided in Materials and Methods. In the main text we shall focus on the infinite allele limit. Despite the seemingly complicated structure of Eq. 2.33, it can be used in efficient numerical calculations. The following sections are devoted to exploring the properties of this formula and discussing its applicability and accuracy if some of the model assumptions are relaxed.

2.2 The effective population size approximation

According to the effective population size (EPS) approximation [42, 44] in the monomorphic limit population dynamics is effectively neutral with a rescaled population size N^*. Indeed, in this limit Eq. 2.33 reduces to

$$\mathbb{P}[\mathbf{n}] \xrightarrow[\theta \to 0]{} \frac{N_P}{k!} \frac{n!}{\prod_{i=1}^{k} n_i} \theta^{k-1} (1-\gamma)^{k-1} \tag{2.34}$$

in the two-plane case. The $\theta \to 0$ limit corresponds to the $s \gg \mu$ regime with s being finite; Eq. 2.34 is the same as the neutral sampling formula (Eq. 2.2) in the monomorphic limit if the population size is rescaled: $N \to N^* = (1-\gamma)N$. This result can be generalized to the landscape with multiple fitness planes, in which case $N^* = \gamma_m N$, where γ_m is a fraction of nodes with the highest fitness.

However, the EPS approximation breaks down in the polymorphic regime. Indeed, if we take the $\theta \to \infty$ limit while keeping s/μ finite, it can be shown for the two-plane landscape that

$$\frac{\mathbb{P}[\mathbf{n}]}{\mathbb{P}[\mathbf{n}, s=0]} \xrightarrow[\theta \to \infty]{} \sum_{m=0}^{\infty} c_m \left(\frac{s}{\mu}\right)^m \equiv \lambda \tag{2.35}$$

where $\mathbb{P}[\mathbf{n}, s=0]$ is given by Eq. 2.2, and the coefficients c_m depend solely on the allelic counts $n_1, \ldots, n_k$. Since the right-hand side of Eq. 2.35 does not depend on the population size, it can be used to define $N^* = \lambda^{1/(k-n)} N$. However, this definition will be sample-specific, as λ depends on the allelic counts via c_m's. Thus there is no universal rescaling of the population size in the strongly polymorphic regime, and therefore evolutionary dynamics is non-neutral.

2.3 Detection of selection signatures

As discussed above, in general we expect allele diversity to deviate from neutrality, making it possible to detect selection signatures using a set of sequences sampled from the population. To investigate non-neutral population dynamics, we compute probabilities for all integer partitions $\mathbf{n} = \{n_1, \ldots, n_k\}$ of n alleles sampled from the population evolving under selection (Eq. 2.33), and compare them with steady-state partition probabilities obtained under neutral evolution (Eq. 2.2) and the monomorphic EPS approximation (Eq. 2.34).

We use the Kullback-Leibler (KL) distance to quantify the difference between two probability distributions [63]: $\mathrm{KL}(p||q) = \sum_i p_i \log(p_i/q_i)$, where i is the partition label. For the two-plane system, we first compare partition probabilities under selection, $p_i = \mathbb{P}[\mathbf{n}, \theta, \beta]$, with the corresponding neutral probabilities, $q_i = \mathbb{P}[\mathbf{n}, \theta, \beta = 0]$. In Fig 2.2A, we plot the KL divergence as a function of the mutation rate and the selection strength for the two-plane fitness landscape. We observe that evolutionary dynamics is essentially neutral if selection is weak ($s \leq \mu$); in addition, the range of selection coefficients for which neutrality holds increases in the monomorphic regime ($N\mu \leq 1$). On the other hand, population statistics is clearly non-neutral when the population is polymorphic and when the separation between the two fitness planes is large. Next, we compute the KL divergence $\mathrm{KL}(p||q^*)$ between the EPS probability distribution, $q_i^* = \mathbb{P}[\mathbf{n}, \theta^*, \beta = 0]$, where $\theta^* = (1-\gamma)\theta$, and p_i (Fig 2.2B). We see that the EPS approximation fails in the polymorphic, weak-selection regime. Overall, the neutral and EPS approximations are approximately complementary: for example, in the strong-selection ($s \gg \mu$) polymorphic regime, when evolutionary dynamics becomes non-neutral, it is well approximated by the EPS model.

In Fig 2.2C we show KL divergences between partition probability distributions on two-

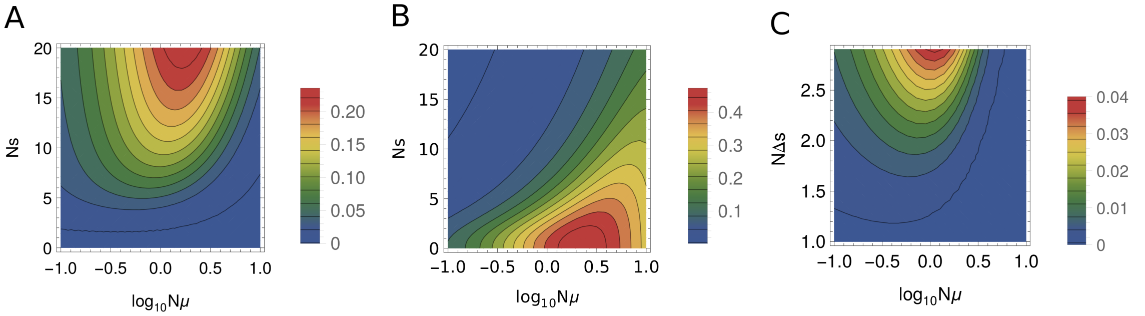

Figure 2.2: **KL divergences of partition probabilities.** Probabilities of all possible partitions of $n = 3$ alleles ($\{3\}$, $\{2, 1\}$ $\{1, 1, 1\}$) were sampled from a population of size $N = 10^3$. (A) and (B) KL divergences for the two-plane fitness landscape as a function of the mutation rate $N\mu$ and the selection coefficient Ns scaled by the population size, for partition probabilities with and without selection (A), and partition probabilities with selection compared with the EPS approximation (Eq. 2.34) (B). (C) KL divergences for the sampling probabilities of all possible partitions on a three-plane vs. two-plane landscape. Alleles in the three planes have fitnesses 1, $1 + s - \Delta s$ and $1 + s - \Delta s$ respectively, with $Ns = 6$ for both two and three-plane landscapes.

and three-plane fitness landscapes. We observe that the partition probabilities are essentially two-plane (i.e., there are no selection signatures indicating presence of intermediate-fitness alleles) if the population is monomorphic ($N\mu \leq 1$), or if the distance between the two upper planes is smaller than the mutation rate ($\Delta s \leq \mu$). However, there is a considerable parameter region in which deviations between two and three-plane sampling probabilities appear to be significant (with KL divergences between the two distributions of 0.01 or more), making it possible to detect three distinct fitness states in the sampling data.

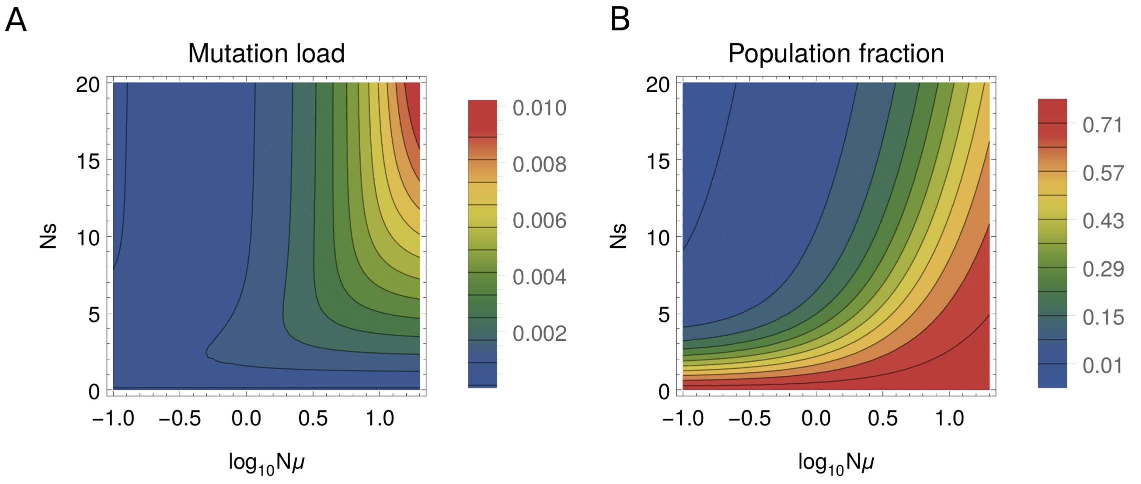

Figure 2.3: **Mutation load and population fraction for the two-plane fitness landscape.** (A) Mutation load (Eq. 2.36) and (B) population fraction in the lower plane, as a function of the mutation rate ($N\mu$) and the selection strength (Ns) rescaled by the population size.

2.4 Mutation load

By definition, the mutation load is given by [56, 59] $L = (f_{\max} - \langle f \rangle)/f_{\max}$, where $f_{\max}$ is the maximum fitness and $\langle f \rangle = \sum_{i=1}^{K} x_i f_i$ is the mean population fitness. To estimate the mutation load at steady state, we compute the expected value of the mean population fitness over multiple realizations of the stochastic process, $\mathbb{E}[\langle f \rangle]$.

For the two-plane system, this computation leads to

$$L = \frac{s\gamma}{1+s} \frac{{}_1F_1(\gamma\theta + 1; \theta + 1; -Ns)}{{}_1F_1(\gamma\theta; \theta; -Ns)}. \tag{2.36}$$

Another indicative quantity is the average fraction of the population with low fitness, $\mathbb{E}[x_{\text{low}}]$. For the two-plane system it is given by $\mathbb{E}[x_{\text{low}}] = L(1+s)/s$.

Values of mutation load for the two-plane fitness landscape are shown in Fig 2.3A over a range of selection strengths and mutation rates. As expected, we observe that the largest deviations from the maximum fitness occur in the strong-mutation, strong-selection regime, where a fraction of the population is constantly displaced to the lower plane by mutation, incurring a fitness cost. Correspondingly, at a given value of selection strength the mutation load increases with the mutation rate. In the monomorphic regime the mutation load is vanishingly low because the entire population condenses to a single allelic state and moves randomly on the upper plane. The fraction of the population on the lower fitness plane is shown in Fig 2.3B. The fraction is high when the separation between the two planes is low and, at a fixed separation, it increases with the mutation rate.

2.5 Fitness landscape models and numerical simulations

To check our main result (Eq. 2.33), we have compared it to the outcomes of numerical simulations of two models. In the first model, each allele is allowed to mutate into any of the other $K - 1$ alleles with equal probability. We call this model fully-connected (FC); derivations of the Ewens sampling formula and our generalization of it (Eq. 2.33) were carried out for the FC model. The second model is more realistic: an organism is represented by a sequence of integers $S = (a_1, \ldots, a_L)$ of length L and alphabet size A, meaning that $0 \leq a_i \leq A - 1$. A mutation replaces an integer at a randomly chosen site with one of the remaining $A - 1$ integers; all the replacements have equal probabilities. We call this model a single-point mutation (SPM) model; it is a more realistic description of protein or nucleotide sequence evolution.

To assign a fitness value to each allele, we focus on the landscapes in which alleles can have either low or high fitness values (the two-plane model), or low, intermediate, and high fitness values (the three-plane model). The fractions of alleles found in each plane are given by $\boldsymbol{\gamma}$: $\boldsymbol{\gamma} = (\gamma, 1 - \gamma)$ for the two-plane model and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, 1 - \gamma_1 - \gamma_2)$ for the three-plane model. In the FC model, the mutational neighborhood of each allele is the same, so that any desired allele fractions $\boldsymbol{\gamma}$ can be implemented. However, in the SPM model the fractions of neutral, beneficial and deleterious moves in each plane will depend on $\boldsymbol{\gamma}$ and the assignment of states to planes. We wished to produce non-trivial distributions of neutral moves on the fitness planes, with mutational neighborhoods of some alleles being completely neutral in each plane. Another condition was that the number of alleles in each plane should decrease with its fitness, to reflect the fact that beneficial mutations are rare.

To fulfill these requirements, we chose to assign fitness values in the SPM model in the following way. We use the sequence length $L = 10$ and the alphabet size $A = 4$. For each sequence $S = (a_1, \ldots, a_L)$ we compute a score $z = a_1 + \ldots + a_L$. We compare these scores with a set of cutoffs $(c_1, \ldots, c_{M-1})$ for the M-plane landscape. For the two-plane landscape, the fitness is 1 if $z \leq c_1$, and $1 + s$ otherwise. We use the cutoff $c_1 = 17$, which yields $\boldsymbol{\gamma} = (0.758, 0.242)$. For the three-plane landscape, if $z \leq c_1$ the fitness is 1, if $c_1 < z \leq c_2$ the fitness is $1 + s - \Delta s$, and if $z > c_2$ the fitness is $1 + s + \Delta s$. We choose the cutoffs $c_1 = 17$ and $c_2 = 21$, which lead to $\boldsymbol{\gamma} = (0.758, 0.210, 0.032)$. In order to compare FC and SPM simulations directly, we use the same values of $\boldsymbol{\gamma}$ in the corresponding FC models.

Our numerical simulations have been carried out using the Moran model of population genetics [28, 3]. Specifically, we have evolved a population of $N = 10^3$ haploid organisms, each of which could be in one of K allelic states. At each step a parent is chosen by randomly sampling the population with weights proportional to the fitness of each individual. An offspring is then produced as an exact copy of the parent. Next, the offspring undergoes

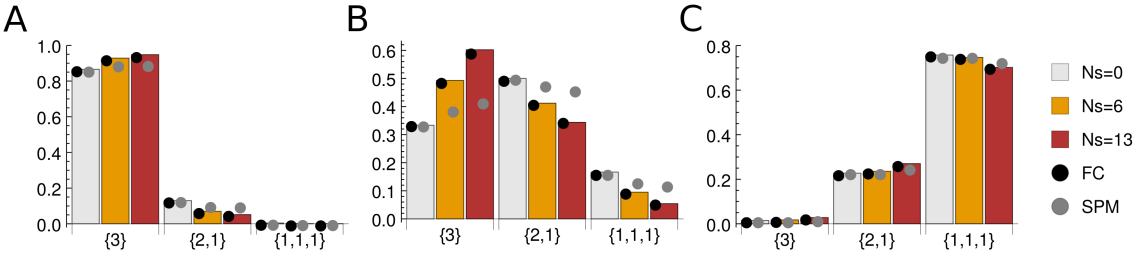

Figure 2.4: **Partition probabilities for the two-plane fitness landscape.** Shown are sampling probabilities of all partitions with $n = 3$: $\{3\}$, $\{2, 1\}$, $\{1, 1, 1\}$. Bars: theoretical predictions in the infinite allele limit. Black circles: numerical simulations on the FC sequence network. Grey circles: numerical simulations on the SPM sequence network. In all simulations, alphabet size $A = 4$, sequence length $L = 10$, and population size $N = 10^3$ were used. Partition probabilities were estimated from 10^6 samples as described in the main text. (A) Monomorphic population, $N\mu = 0.1$. (B) Weakly polymorphic population, $N\mu = 1.0$. (C) Strongly polymorphic population, $N\mu = 10.0$. The corresponding KL divergences are listed in Table 2.1. Note that the error bars of the partition probabilities are too small to be shown, due to extensive sampling in our numerical simulations.

mutation with the probability μ. Finally, the population is uniformly sampled to choose an organism that will be replaced by the offspring, keeping the overall population size constant. Probabilities of sampling n individuals from the population were calculated as averages over 10^6 samples gathered from 10^3 independent runs. For each run, a randomly generated initial population was evolved to steady state, after which n individuals were sampled from the population with replacement 10^3 times, waiting $\sim 1/\mu$ generations between subsequent samples.

Note that in the neutral case the exact mapping between θ and μ is given by $\theta = N\mu/(1-\mu)$ for the Moran model. [28] However, it is unclear if this mapping can be extended to the non-neutral cases considered here. In any event, for the population size and the values of θ investigated below, $\mu = \theta/(N + \theta) \simeq \theta/N$. Therefore, we use the diffusion theory result $\theta = N\mu$ in comparing theoretical predictions with numerical simulations.

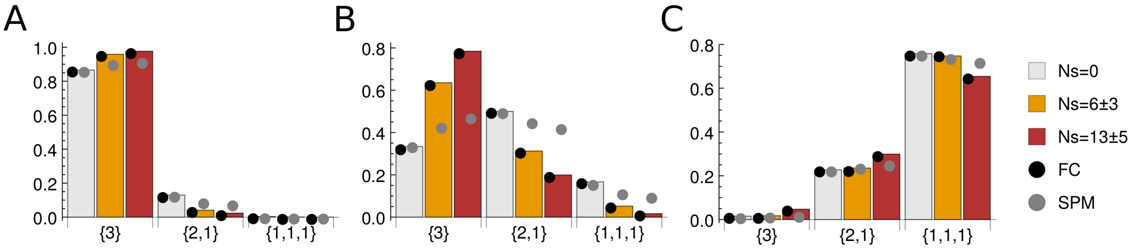

Figure 2.5: **Partition probabilities for the three-plane fitness landscape.** All notation and symbols are as in Fig 2.4. The corresponding KL divergences are listed in Table 2.1.

2.6 Partition probabilities on fully-connected vs. single-point-mutant networks

Here we investigate the extent to which sampling probabilities change in the SPM sequence evolution model described above, compared to the FC fitness landscape. We are especially interested in the limits of the predictive power of our theoretical framework, which necessarily involves the FC assumption. In Fig 2.4 and Table 2.1 we compare theoretical predictions with numerical simulations on the FC and SPM networks in the two-plane system for the sample of $n = 3$ alleles. Overall, as expected, we observe an excellent agreement between theory and simulations on FC networks. Furthermore, we see that the agreement between SPM simulations and our theoretical results is reasonable: in nearly all cases, the predicted ranking of the sample partitions, as well as the ranking of any given sample partition with respect to the selection strength, Ns, are preserved. The largest discrepancies occur in the weakly polymorphic ($N\mu = 1$), non-neutral regime ($Ns = 6, 13$).

The situation is qualitatively similar when a three-plane fitness landscape is considered (Fig 2.5, Table 2.1). We again observe an excellent agreement between theory and FC simulations and, overall, a reasonable agreement between theory and SPM simulations, with

		Single-plane landscape	**Two-plane landscape**	
		$Ns = 0$	$Ns = 6$	$Ns = 13$
$N\mu = 0.1$	**FC**	1×10^{-5}	2×10^{-5}	3×10^{-5}
	SPM	1×10^{-5}	9×10^{-3}	2×10^{-2}
	Ratio	1.000	0.452	0.425
$N\mu = 1.0$	**FC**	2×10^{-5}	8×10^{-5}	1×10^{-4}
	SPM	1×10^{-4}	2×10^{-2}	9×10^{-2}
	Ratio	1.000	0.363	0.508
$N\mu = 10.0$	**FC**	1×10^{-6}	6×10^{-5}	2×10^{-4}
	SPM	1×10^{-4}	4×10^{-5}	3×10^{-3}
	Ratio	1.000	0.331	0.345

		Three-plane landscape	
		$Ns = 6 \pm 3$	$Ns = 13 \pm 5$
$N\mu = 0.1$	**FC**	4×10^{-5}	2×10^{-6}
	SPM	2×10^{-2}	3×10^{-2}
	Ratio	0.370	0.380
$N\mu = 1.0$	**FC**	1×10^{-6}	6×10^{-6}
	SPM	8×10^{-2}	2×10^{-1}
	Ratio	0.378	0.434
$N\mu = 10.0$	**FC**	2×10^{-5}	4×10^{-5}
	SPM	2×10^{-4}	2×10^{-2}
	Ratio	0.595	0.488

Table 2.1: **KL divergences between theoretical predictions and numerical simulations for single-plane, two-plane (Fig 2.4), and three-plane (Fig 2.5) fitness landscapes, with the sample size $n = 3$.** Note: **FC** = KL(p = numerical FC $||$ q = theory), **SPM** = KL(p = numerical SPM $||$ q = theory), **Ratio** = KL(p = theory $||$ q = numerical SPM)/KL(p = theory $||$ q = numerical neutral SPM).

the largest discrepancies again occurring in the weakly polymorphic, non-neutral regime. These observations remain true when samples with $n = 4$ and 5 alleles are considered (Tables 2.2,2.3).

Finally, we have checked whether our theoretical predictions, which rely on the full-connectivity assumption, are closer to the non-neutral rather than neutral SPM steady-state dynamics in numerical simulations: if this is the case, we should be able to predict

		Single-plane landscape	**Two-plane landscape**	
		$Ns = 0$	$Ns = 6$	$Ns = 13$
$N\mu = 0.1$	**FC**	1×10^{-5}	6×10^{-6}	6×10^{-6}
	SPM	1×10^{-5}	9×10^{-3}	2×10^{-2}
	Ratio	1.000	0.394	0.527
$N\mu = 1.0$	**FC**	9×10^{-5}	3×10^{-5}	8×10^{-5}
	SPM	9×10^{-4}	3×10^{-2}	1×10^{-1}
	Ratio	1.000	0.527	0.542
$N\mu = 10.0$	**FC**	2×10^{-5}	6×10^{-6}	7×10^{-5}
	SPM	2×10^{-4}	1×10^{-4}	3×10^{-3}
	Ratio	1.000	0.418	0.199

		Three-plane landscape	
		$Ns = 6 \pm 3$	$Ns = 13 \pm 5$
$N\mu = 0.1$	**FC**	1×10^{-6}	5×10^{-6}
	SPM	2×10^{-2}	4×10^{-2}
	Ratio	0.397	0.432
$N\mu = 1.0$	**FC**	2×10^{-4}	2×10^{-4}
	SPM	1×10^{-1}	3×10^{-1}
	Ratio	0.442	0.486
$N\mu = 10.0$	**FC**	7×10^{-6}	1×10^{-5}
	SPM	2×10^{-4}	2×10^{-2}
	Ratio	0.677	0.406

Table 2.2: **KL divergences between theoretical predictions and numerical simulations for single-plane, two-plane (Fig 2.4), and three-plane (Fig 2.5) fitness landscapes, with the sample size $n = 4$.** Note: **FC** = KL(p = numerical FC || q = theory), **SPM** = KL(p = numerical SPM || q = theory), **Ratio** = KL(p = theory || q = numerical SPM)/KL(p = theory || q = numerical neutral SPM).

selection signatures in populations evolving under single-point mutations using our methodology. We have computed the ratio of KL distances defined in the Table 2.1 caption; this ratio is less than 1 if the theoretical predictions with selection are closer to the corresponding SPM simulation than to the neutral SPM simulation, and greater than 1 otherwise. We observe that the ratio is less than 1 in all cases with selection and for all sample sizes (Tables 2.1–2.3), indicating that the error introduced by the FC assumption is less than

		Single-plane landscape	**Two-plane landscape**	
		$Ns = 0$	$Ns = 6$	$Ns = 13$
$N\mu = 0.1$	**FC**	1×10^{-5}	2×10^{-5}	3×10^{-5}
	SPM	3×10^{-5}	1×10^{-2}	2×10^{-2}
	Ratio	1.000	0.441	0.385
$N\mu = 1.0$	**FC**	9×10^{-5}	1×10^{-4}	3×10^{-4}
	SPM	5×10^{-4}	4×10^{-2}	1×10^{-1}
	Ratio	1.000	0.428	0.485
$N\mu = 10.0$	**FC**	1×10^{-5}	1×10^{-5}	5×10^{-4}
	SPM	1×10^{-3}	5×10^{-4}	8×10^{-3}
	Ratio	1.000	0.461	0.548

		Three-plane landscape	
		$Ns = 6 \pm 3$	$Ns = 13 \pm 5$
$N\mu = 0.1$	**FC**	5×10^{-6}	3×10^{-6}
	SPM	3×10^{-2}	4×10^{-2}
	Ratio	0.379	0.429
$N\mu = 1.0$	**FC**	7×10^{-4}	4×10^{-5}
	SPM	1×10^{-1}	3×10^{-1}
	Ratio	0.426	0.514
$N\mu = 10.0$	**FC**	1×10^{-4}	1×10^{-3}
	SPM	4×10^{-4}	4×10^{-2}
	Ratio	0.546	0.516

Table 2.3: **KL divergences between theoretical predictions and numerical simulations for single-plane, two-plane (Fig 2.4), and three-plane (Fig 2.5) fitness landscapes, with the sample size $n = 5$.** Note: **FC** = KL(p = numerical FC || q = theory), **SPM** = KL(p = numerical SPM || q = theory), **Ratio** = KL(p = theory || q = numerical SPM)/KL(p = theory || q = numerical neutral SPM).

the distance between selective and neutral systems (note that the ratio is 1 by definition in the single-plane neutral case).

2.7 Infinite-allele assumption

Although our approach is valid for an arbitrary number of alleles K, statistics of allele diversity in a population under selection become substantially easier to deal with in the

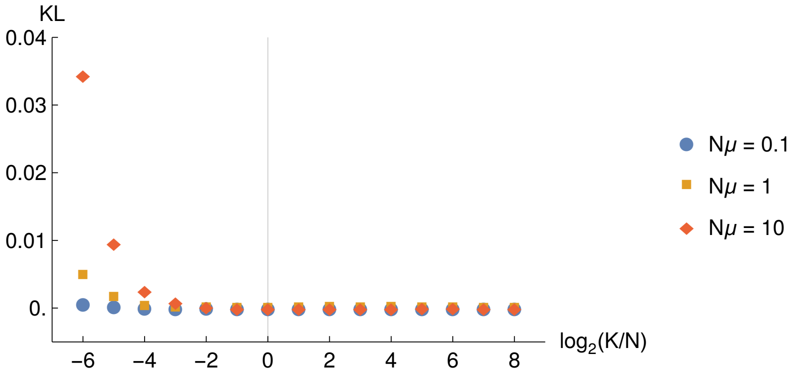

Figure 2.6: **Test of the infinite-allele assumption.** Shown are KL divergences between computational and theoretical partition probabilities on the FC two-plane fitness landscape ($Ns = 6$, $\boldsymbol{\gamma} = (0.758, 0.242)$), as a function of the log ratio between the total number of alleles K and the population size N. The sample size is $n = 3$; partition probabilities were estimated from 10^6 samples. Population size is $N = 10^3$, and the total number of alleles is $K = 10^3 \times 2^i$, $i \in \{-6 \ldots 8\}$. For smaller networks, the number of sequences in the upper and lower planes had to be rounded to the nearest integer. Diamonds: polymorphic population ($N\mu = 10.0$), squares: weakly polymorphic population ($N\mu = 1.0$), circles: monomorphic population ($N\mu = 0.1$). The solid vertical line corresponds to $K = N$.

infinite-allele limit. As discussed in the Introduction, this limit is justified since our focus here is on evolution of protein, RNA and DNA sequences, where the number of alleles grows exponentially with sequence length. Nonetheless, we have systematically investigated the extent of deviations between our infinite-allele theoretical results and simulations as the number of alleles K decreases and becomes comparable to the population size N. Fig 2.6 shows the KL divergence between partition probabilities derived theoretically for the two-plane landscape in the infinite-allele limit (Eq. 2.33) and obtained numerically on finite-size FC networks. We consider three regimes: monomorphic ($N\mu = 0.1$), weakly polymorphic ($N\mu = 1.0$), and strongly polymorphic ($N\mu = 10.0$). In the latter two cases, noticeable deviations between theory and simulations begin to appear below the $K \sim N$ regime; the agreement improves as the population becomes more monomorphic. We conclude that our theory is applicable over a wide range of mutation rates, as long as the network size is comparable to, or greater than, the population size.

2.8 Discussion and Conclusion

One of the most challenging problems in evolutionary biology is to understand evolutionary dynamics of molecular loci, such as protein or RNA-coding sequences, or gene regulatory regions. The number of nucleotides at these loci, L, is large enough so that the total number of possible sequences, $K = A^L$, is astronomical, far exceeding the population size N. Under these conditions the evolution of a molecular locus, assumed to be decoupled by recombination from the rest of the genome, reaches a "de-labelled" steady state. The allelic diversity in the steady-state population is determined by the balance of forces of selection and drift on one hand, and mutation on the other. The former act to reduce allelic diversity, while the latter acts to increase it. As a result, population statistics such as the mean number of distinct alleles, or the probability of seeing a certain allelic configuration in a sample, do not change with time, even though new genotypes continue to be explored on the effectively infinite allelic network.

The steady-state allelic diversity in an infinite-allele neutral system was explored by Ewens [28, 8]. The main result of that study, the Ewens sampling formula, is widely used in population genetics. However, selection is bound to play a key role in molecular evolution, and recent high-throughput studies connecting protein sequences with phenotypes [15, 18, 19, 9, 10, 11, 12] reveal a more complex picture of molecular evolution: generally, a functional protein is disrupted by a fraction of mutations (e.g., through substitution of a hydrophobic residue for a hydrophilic one in the protein core). Other mutations do not significantly change protein stability, binding affinity, or binding specificity, and are therefore effectively neutral. Occasionally, a mutation is found which increases the fitness of an already functional, adapted protein, but these mutations are very infrequent. Overall, recent experimental studies indicate that "coarse-grained" fitness landscapes comprised of multiple interconnected planes (i.e., several distinct fitness states) are a reasonable representation. The simplest landscape of this kind has just two fitness states, with functional sequences on the upper plane and non-functional sequences on the lower plane [9]. Multiple-plane fitness landscapes constructed in this way are characterized by extensive epistasis under the single-point mutational move set, which is likely to be pervasive in molecular evolution [14, 15, 16, 17].

Since molecular evolution may be described by steady-state dynamics on multiple-plane fitness landscapes, it is of great interest to generalize the Ewens sampling formula to arbitrary fitness distributions, and to the case of several distinct fitness states in particular. Tractable expressions for sampling probabilities would enable inference of selection coefficients, relative numbers of alleles in each fitness state, and mutation rates, using DNA, RNA, or protein sequences sampled from the population as input to the inference procedure. Here we report an extension of the Ewens sampling formula to arbitrary fitness distributions, focusing on the multiple-plane case which yields substantial simplifications in the infinite-allele limit. Unlike techniques based on the Poisson random field framework [45], such as the sampling probability formulas developed by Desai et al. [44], our approach does not rely on assuming independent evolution at each site along the sequence. However, an essential drawback of the Ewens sampling formula and our generalization of it is the "full-connectivity" assumption (i.e., that each allele can mutate into every other allele). Furthermore, the sampling formula becomes intractable for large sample sizes, since the number of terms to sum over in Eq. 2.33 becomes too large.

Therefore, in order to study the limits of applicability of our theory, we have carried out extensive comparisons with numerical simulations on multiple-plane fitness landscapes. First, we checked the full-connectivity assumption inherent in the Ewens approach by comparing the sampling probabilities of our theory with those obtained by simulation of steady-state populations evolving on single-point-mutant networks. We find that the agreement, although dependent on the details of the fitness landscape model, the values of selection coefficients, and mutation rates (and least reliable in the weakly polymorphic regime), remains strong enough overall to encourage application of our theoretical results to sequence data. We also find that the error introduced by the full-connectivity assumption, as measured by the KL distance, is less than the distance between sampling probabilities in neutral and non-neutral systems. Note that our SPM model of the fitness landscape was constructed specifically to create a non-trivial distribution of neutral, deleterious and beneficial single-point mutations for the alleles, in some sense making it as distant from the fully connected network as possible. Thus we expect the errors inherent in our theoretical framework to be smaller (or at least not much worse) in applications to natural systems. Second, we have checked the infinite-allele assumption by systematically reducing the number of alleles until it became lower than the population size. We find that, for a wide range of mutation rates, deviations between theory and simulations become significant only when the number of alleles approaches the population size from above. Thus our assumption of the infinite network size is justified for sufficiently long loci, such as those encoding transcribed or regulatory regions.

Robust inference of selection coefficients from a sample of sequences collected from an evolving population requires statistics of allelic diversity to deviate substantially from the neutral expectation. If selection cannot be ruled out *a priori*, the use of our generalized Ewens sampling formula, which is valid throughout the entire parameter space, is necessary for inferring selection signatures and mutation rates from data. Moreover, allelic diversity generated by steady-state evolutionary dynamics on a three-plane fitness landscape is sufficiently distinct from its two-plane counterpart in the strong-selection, weakly polymorphic regime, opening up a possibility of inferring multiple selection coefficients from a sample of sequences. Another hallmark of non-neutral population dynamics is de-localization of the population to multiple fitness planes. With a two-plane landscape, we expect the fraction of the population on the lower plane to increase with the mutation rate and decrease with the distance between the two planes. Our investigation of the mutation load confirms these predictions.

In summary, we have generalized the Ewens sampling formula to populations evolving under selection. Although in principle our results are valid for arbitrary fitness distributions, focusing on the infinite allele limit and landscapes characterized by several distinct fitness states yields substantial simplifications, making our approach computationally tractable and thus applicable to inferring selection signatures from high-throughput sequence data. Such multiple-state "coarse-grained" fitness distributions appear to be a reasonable starting point supported by recent large-scale genotype-phenotype maps in molecular systems [15, 18, 19, 9, 10, 11, 12]. Unlike previous approaches, we do not assume that each site along the sequence evolves independently – an assumption that has recently been challenged in molecular evolution studies [14, 15, 16, 17]. However, we do make the infinite allele assumption, and, as in the Ewens original formula [8], assume that each allele can mutate into any other allele. Therefore, we check our theory against numerical simulations in model systems where these assumptions are relaxed, and find that our predictions remain accurate enough to enable inference of evolutionary parameters from sequencing data.

Chapter 3

Inference and Sampling Formulae

We have developed a computational approach to simultaneous genome-wide inference of key population genetics parameters: selection strengths, mutation rates rescaled by the effective population size and the fraction of viable genotypes, solely from an alignment of genomic sequences sampled from the same population. Our approach is based on a generalization of the Ewens sampling formula, used to compute steady-state probabilities of allelic counts in a neutrally evolving population, to populations subjected to selective constraints. Patterns of polymorphisms observed in alignments of genomic sequences are used as input to Approximate Bayesian Computation, which employs the generalized Ewens sampling formula to infer the distributions of population genetics parameters. After carrying out extensive validation of our approach on synthetic data, we have applied it to the evolution of the *Drosophila melanogaster* genome, where an alignment of 197 genomic sequences is available for a single ancestral-range population from Zambia, Africa. We have divided the *Drosophila* genome into 100-bp windows and assumed that sequences in each window can exist in either low- or high-fitness state. Thus, the steady-state population in our model is subject to a constant influx of deleterious mutations, which shape the observed frequencies of allelic counts in each window. Our approach, which focuses on deleterious mutations and accounts for intra-window linkage and epistasis, provides an alternative description of background selection. We find that most of the *Drosophila* genome evolves under selective constraints imposed by deleterious mutations. These constraints are not confined to known functional regions of the genome such as coding sequences and may reflect global biological processes such as the necessity to maintain chromatin structure. Furthermore, we find that inference of mutation rates in the presence of selection leads to mutation rate estimates that are several-fold higher than neutral estimates widely used in the literature. Our computational pipeline can be used in any organism for which a sample of genomic sequences from the same population is available.

3.1 Introduction

Explaining the origin of genetic variation observed in natural populations is a long-standing problem in evolutionary biology. While earlier views of genome evolution were based on neutral theory, which emphasizes the role of random genetic drift in the observed patterns of intra- and interspecies variation [64], recent studies have questioned the applicability of the neutral theory or its nearly neutral extension [65] to genome evolution.

One of the strongest cases for the pervasive role of natural selection in metazoans has been made in *Drosophila melanogaster* [61], where population genetics modeling is enabled and guided by the availability of hundreds of sequenced genomes, including 197 from a single population in Zambia, Africa [66, 67], and by the functional genomics databases such as FlyBase [68]. In particular, genomic data from *D. melanogaster* and related species was used to argue that a large fraction of the *D. melanogaster* genome, including non-coding regions, is under widespread purifying and positive selection [69, 70]. Genetic linkage, long recognized as a key evolutionary force [71, 72], is also likely to play an important role in fly evolution, either through selective sweeps: hitchhiking of sites adjacent to a beneficial (adaptive) site that rises rapidly to fixation [73, 74, 75] or through background selection: continuous generation and removal of strongly deleterious mutations by natural selection, in the presence of recombination [42, 76, 77].

Although the evidence that selective forces and genetic linkage play key roles in *Drosophila* evolution is compelling, the exact nature and the relative contributions of the underlying evolutionary processes are less clear [78, 79]. In particular, selective sweeps and background selection produce qualitatively similar outcomes that may be difficult to differentiate given available genomic data [22]. A recent study has argued that background selection alone can account for a large fraction of the observed patterns of nucleotide diversity in *Drosophila melanogaster* [80]. The study predicted nucleotide diversity at neutral sites in the presence of selection against deleterious mutations at genetically linked sites [76, 81]. The model required several inputs: deleterious mutation rates, a parameterized distribution of deleterious selection coefficients, and the recombination frequency between the focal neutral site and the site under selection.

Here we provide an alternative approach to modeling background selection. We model the fitness landscape explicitly by assigning each allele either to a low- or a high-fitness state. The fitness difference between the two fitness states, the overall mutation rate and the fraction of alleles in the high-fitness state are inferred from rather than input into the model. To carry out the inference, we assume that the population has reached steady state and use a generalization of the neutral Ewens sampling formula [8, 28] to fitness landscapes with multiple distinct states [82]. As with the neutral Ewens sampling formula, basic input data consists of counts of distinct alleles in samples of aligned genomic sequences. In the absence of selection, the population maintains a steady state characterized by mutation-drift balance. As the strength of selection (i.e., the fitness difference between the two fitness states) is increased, alleles become more and more concentrated on the upper fitness plane, with each high-fitness allele subject to both neutral and deleterious mutations. Selection against deleterious mutations affects the observed frequency spectrum and the distribution of allelic counts, allowing us to carry out the inference process. We parse the *D. melanogaster* genome into non-overlapping 100-bp windows. Each window contains up to 197 aligned sequences from the Zambian fly population, providing the allelic counts that serve as input to the computational inference pipeline. Due to the complexity of the generalized Ewens sampling formula, we have opted for Approximate Bayesian Computation (ABC), a Bayesian inference method that can be used to estimate posterior distributions of model parameters [83, 84, 85].

The key strength of the generalized Ewens sampling approach is that it goes beyond standard statistical tests of natural selection [86], yielding explicit estimates of key evolutionary parameters such as mutation rate and selection strength. At the same time, in contrast to the Poisson Random Field approach, [45, 87, 88, 89] our methodology does not assume that each nucleotide evolves independently, and therefore is capable of treating site linkage and epistasis within each genomic window. However, as with every population genetics model, several simplifying assumptions have to be made. First, the Ewens sampling approach implies that any allele can mutate into any other allele with a single mutation rate. We investigate this issue both in this work and in Ref. [82] and conclude that systematic errors caused by this assumption are likely to be modest. Second, unlike the approach to background selection originally described in Refs. [76, 81], we do not treat recombination explicitly. This issue is mitigated by employing short 100-bp genomic windows (compared e.g. to 1-100 kbp windows in Ref. [80]), which however make the analysis more computationally challenging. Third, we assume that the genomic sequences in each window are in a de-labeled steady state [28]: although the sequences keep mutating into novel alleles, the de-labeled population statistics such as the frequency spectrum and the number of distinct alleles are stationary on average. If the steady-state assumption is correct, our inference should not be affected by past expansions and contractions in the population size, such as the significant bottleneck inferred for the Zambian *D. melanogaster* population that we use in this work [90]. Moreover, even if the population is still expanding after the bottleneck, our approach may still be applicable but will yield evolutionary parameters based on a reduced effective population size which is in turn affected by implicit bottleneck parameters.

Using the approach described above, we find that most of the fly genome evolves under selective pressure from deleterious mutations, and that for the alleles in the high-fitness state, most mutations are deleterious. The former observation is in line with the emerging consensus view on the selective constraints imposed on the evolution of the *D. melanogaster* genome [61], and with a recent observation of the major role of purifying selection in establishing the observed patterns of nucleotide diversity across the fly genome [80]. Our approach thus establishes a baseline against which other selective signatures such as selective sweeps or balancing selection are to be discerned. Our methodology presents an alternative to statistical tests for selection, including background selection under recombination, and it is reassuring that it reaches broadly similar conclusions as the previous studies. Moreover, it provides an alternative to the Poisson Random Field framework as a population genetics method capable of estimating selection strength directly from single-population DNA sequence data. Finally, our approach yields a revised estimate of θ, mutation rates rescaled by the effective population size, suggesting that widely used neutral estimates need to be reconsidered, as neutral estimates of θ are systematically biased towards lower values.

3.2 Overview of Approximate Bayesian Computation (ABC)

ABC is an efficient inference method that can be used to estimate posterior distributions of model parameters in cases where the likelihood function is either unavailable as a closed-form expression or computationally costly to evaluate [83, 84, 85]. Let us suppose that we have observed a statistic $\mathbf{x_0}$ and we would like to learn the distribution of model parameters $\boldsymbol{\alpha}$ that would be consistent with the observed statistic $\mathbf{x_0}$. We also have a probabilistic model M that can generate the statistic of interest from the underlying parameters: $\boldsymbol{\alpha} \to \mathbf{x}$. Theoretically, it means that there exists a likelihood function $p(\mathbf{x}|\boldsymbol{\alpha})$ but in practice either the closed form of this distribution is unknown or the computation of the likelihood is prohibitively expensive.

In such cases, Approximate Bayesian Computation (ABC) can be employed to infer the distribution of model parameters. [83, 84, 85] ABC is based on the following rejection algorithm: First, pick a prior $p(\boldsymbol{\alpha})$ and sample from it to get an empirical distribution of model parameters $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)$. Using the model M, generate the corresponding sample for the statistic of interest: $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ (i.e., generate $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ by sampling once from each $p(\mathbf{x}|\boldsymbol{\alpha}_i)$, $i = 1 \ldots m$). Second, choose an appropriate measure of distance d in the space of statistics. Then the rejection algorithm amounts to keeping only the parameters

$\boldsymbol{\alpha}_i$ that generated the statistics $\mathbf{x}_i$ which satisfy $d(\mathbf{x}_i, \mathbf{x_0}) < \epsilon$ for a certain tolerance ϵ. The tolerance hyperparameter ϵ can be chosen by requiring that we keep only a certain percentage or a certain number of m datapoints that best match the observation [83]. The final distributions of the model parameters retained after the rejection step constitute empirical approximations to posterior probabilities.

3.3 *D. melanogaster* population genomics data

In this work, we employ sequenced and aligned haploid embryo genomes from a single ancestral range population of *D. melanogaster* from Zambia, Africa, which were made available in phase 3 of the Drosophila Population Genomics Project (DPGP3). [66, 67] Specifically, our data consists of $n = 197$ aligned sequences for chromosomes 2L, 2R, 3L, and 3R, and $n = 195$ aligned sequences for chromosome X. To summarize the observed allelic diversity in this dataset, we have divided each chromosome into 100 bp non-overlapping windows. The 100-bp window size was chosen to be able to observe non-trivial patterns of allelic diversity while minimizing the effects of recombination. We remove all windows in which more than 20% of the sequences in the alignment have at least one undetermined nucleotide (labeled as 'N'). If the fraction of such sequences is $< 20\%$, we keep the window but remove the affected sequences. We also remove all monomorphic windows as their frequency spectrum is uninformative for our analysis. This preprocessing step results in retaining 68% of all windows in chromosome 2L, 73% in 2R, 70% in 3L, 75% in 3R, and 63% in X.

For each remaining window, we compute a vector of allelic counts $\mathbf{n} = \{n_1, \ldots, n_k\}$, where k is the total number of groups of identical sequences in the window and n_j is the number of identical sequences in group $j = (1 \dots k)$. Since $\sum_{j=1}^{n} n_j = n$, $\mathbf{n}$ can be viewed as an integer partition of n. As an example of the calculation of allelic counts, consider a genomic window in which out of $n = 197$ aligned sequences, 187 occur once, 3 sequences occur twice, and 1 sequence occurs 4 times. The corresponding vector of allelic counts is given by $\mathbf{n} = \{4, 2, 2, 2, 1, \dots, 1\}$, where 1 is repeated 187 times, so that $k = 191$. Note that the same information can be conveyed with the frequency spectrum widely used in population genetics [8, 46, 28]: $\mathbf{a} = (a_1, \dots, a_n)$, where a_j is the number of groups of identical elements of size j. The a_j counts are subject to normalization $\sum_{j=1}^{n} j a_j = n$. In the above example, $a_1 = 187$, $a_2 = 3$, $a_4 = 1$, and the rest of a_j's are equal to 0. We shall use the $\mathbf{n}$ representation in the rest of the paper.

3.4 Application of ABC in population genetics

The main objective of this study is to infer genome-wide distributions of key evolutionary parameters such as selection strengths and mutation rates directly from sequence alignments of *D. melanogaster* genomes sampled from a single population. To this end, we use the generalized Ewens sampling formula which describes allelic diversity of a steady-state population evolving on an arbitrary fitness landscape. [82] This is an infinite-allele model where any allele can mutate into any other allele with a single mutation rate. Unlike approaches that treat deleterious background selection with recombination, [76, 81, 80] our method does not take recombination into account explicitly; however, in contrast to the Poisson Random Field approach, [45, 87, 88, 89] our methodology does not assume site independence and therefore is capable of treating linkage within each genomic window.

For a general fitness landscape with M distinct fitness states, the probability of the vector of allelic counts $\mathbf{n}$ is given by [82]

$$\mathbb{P}[\mathbf{n}|\theta,\boldsymbol{\beta},\boldsymbol{\gamma}] = N_P \frac{n!}{k!} \frac{1}{\prod_{i=1}^{k} n_i} \frac{\theta^k}{\theta^{(n)}} \sum_{\nu \in P(\mathbf{n})} \sum_{Y \in \mathcal{Y}(\mathbf{n})} \frac{1}{N_P} \frac{\mathcal{F}(\boldsymbol{\gamma}\theta + \boldsymbol{\nu}^Y; \theta + n; \boldsymbol{\beta})}{\mathcal{F}(\boldsymbol{\gamma}\theta; \theta; \boldsymbol{\beta})} \binom{k}{i_1 \ldots i_M} \gamma_1^{i_1} \ldots \gamma_M^{i_m}. \tag{3.1}$$

Here $\theta = aN\mu$ is the rescaled mutation rate, where N is the effective population size, μ is the mutation rate per locus, and the prefactor $a = 2$ for the Wright-Fisher model [1] and 1 for the Moran model [3]. Fitness landscapes are encoded via a vector of selection strengths $\boldsymbol{s}$ rescaled by the population size, $\boldsymbol{\beta} = N\boldsymbol{s}$, and a vector $\boldsymbol{\gamma}$ which determines landscape geometry by specifying the fraction of alleles (genotypes) in each fitness state. For example, with only two fitness states ($M = 2$), we assign fitness 1 to the fraction $1 - \gamma$ of all alleles and fitness $1 + s$ to the remaining fraction γ, resulting in $\boldsymbol{\beta} = (0, Ns)$ and $\boldsymbol{\gamma} = (1 - \gamma, \gamma)$. In this case, $1 - \gamma$ can be interpreted as a fraction of deleterious mutations for an allele in a high-fitness state. Finally, N_P is the total number of distinct permutations of the allelic counts, and $\mathcal{F}(\mathbf{a}; b; \mathbf{z})$ is the generalized confluent hypergeometric function [82]. The double sum takes into account all the ways in which an allelic partition $\mathbf{n}$ can be distributed among M fitness states.

In the absence of selection, Eq. (3.1) reduces to the neutral Ewens sampling formula [8] written in terms of the allelic counts n_i:

$$\mathbb{P}[\mathbf{n}|\theta] = N_P \frac{n!}{k!} \frac{1}{\prod_{i=1}^{k} n_i} \frac{\theta^k}{\theta^{(n)}}, \tag{3.2}$$

where $\theta^{(n)} = \theta(\theta + 1) \ldots (\theta + n - 1)$ is the rising factorial.

Additional details, along with representative examples, on how Eq. (3.1) is derived and evaluated, including how the double summation is performed and the numerical treatment of generalized confluent hypergeometric functions, can be found in Ref. [82]. Here it suffices to treat Eq. (3.1) as a "black box" function which is used in the ABC inference pipeline. Our overall objective is to learn genome-wide distributions of model parameters $\boldsymbol{\alpha} = (\theta, \boldsymbol{\beta}, \boldsymbol{\gamma})$ that appear in the generalized Ewens formula (Eq. (3.1)). In this case, a natural choice of the summary statistic $\mathbf{x}$ would be allelic counts $\mathbf{n}$ observed in each 100-bp genomic window. However, with $n = 197$ the total number of possible partitions is $> 3 \times 10^{12}$, too large to perform the double summation in Eq. (3.1). Therefore, in each window we subsample aligned sequences $B = 10^4$ times with replacement, creating sequence samples of size $n' = 5$, for which the probabilities of the corresponding allelic counts $\mathbf{n'}$ are amenable to evaluation using Eq. (3.1) or Eq. (3.2). In this way, we arrive at the empirical distribution $\mathbb{P}[\mathbf{n'}]$, where $\mathbf{n'} = \{n'_1, \ldots, n'_k\}$ encodes all possible partitions for the smaller integer $n' = \sum_{j=1}^{k} n'_j$. Note that our observed summary statistic $\mathbf{x_0}$ is now $\mathbb{P}[\mathbf{n'}]$, an empirical histogram of the frequencies of allelic counts generated by subsampling alignments of $n' = 5$ sequences in each window, as described above. Since we do not have an explicit likelihood formula for this statistic, likelihood-based methods cannot be applied, whereas the ABC framework is still capable of yielding posterior distributions of model parameters.

Next, we choose model parameter priors for ABC sampling. In the case of the fitness landscape with $M = 2$ fitness states, the model parameters are $\boldsymbol{\alpha} = (\theta, Ns, \gamma)$, where γ is the fraction of nodes with higher fitness. We impose the following priors: $\log_{10}\theta \sim \mathsf{Normal}(\mu = 0, \sigma = 1)$, $Ns \sim \mathsf{HalfNormal}(\sigma = 20)$ (note that if X is distributed according to a normal distribution with zero mean, $|X|$ is distributed according to a half-normal distribution), and $-\log_{10}\gamma \sim \mathsf{HalfNormal}(\sigma = 6)$. These distributions reflect our prior expectations of the relevant parameter ranges. We create $m = 10^6$ simulated empirical histograms $\mathbf{x}_i = \mathbb{P}[\mathbf{n'}|\boldsymbol{\alpha}_i]$ $(i = 1 \ldots m)$ by sampling parameters $\boldsymbol{\alpha}$ from their priors and calculating $\mathbb{P}[\mathbf{n}'|\boldsymbol{\alpha}]$ via Eq. (3.1). We rank this dataset against the $\mathbf{x_0} = \mathbb{P}[\mathbf{n}']$ empirical histogram of partition frequencies observed in each window, on the basis of the chi-squared statistic used as a measure of the distance between the two histograms: $d^2(\mathbf{x}_i, \mathbf{x_0}) = \sum_p (x_{i,p} - x_{0,p})^2 / x_{i,p}$, where p labels allelic count partitions and $x_{i,p}$, $x_{0,p}$ are the predicted and observed frequencies of the allelic partition p. We set ABC tolerance ϵ by ranking all 10^6 simulated histograms against the histogram observed in a given window and keeping 10^2 histograms with the smallest d^2 score. We typically use median values to summarize the resulting posterior distributions of the model parameters, since they are less sensitive to outliers.

3.5 Recombination simulations

Population dynamics under mutation, selection and recombination is modeled using the Moran process [3]. The population consists of $N = 10^3$ sequences with alphabet $A = 4$ and length $L = 10$ sites. Each evolutionary simulation starts with a randomly generated population. Subsequent generations are obtained using the following rules:

- Two distinct parental sequences $s_i = (p_1, \ldots, p_L)$ and $s_j = (q_1, \ldots, q_L)$, $i \neq j$, are selected from the population $s_1, \ldots, s_N$ with weights proportional to their fitness.

- A random number r is uniformly sampled in the $(0, 1)$ range. If r is less than the recombination rate ρ, an offspring sequence $(p_1, \ldots, p_{b-1}, q_b, \ldots, q_L)$ is formed using the parent sequences s_i and s_j, where the breakpoint b is uniformly sampled from $2 \leq b \leq L - 1$. If $r > \rho$, recombination does not occur and the offspring is a copy of the first parent s_i.

- The offspring undergoes a single-point mutation with probability μ at a randomly chosen site $1 \leq k \leq L$.
- The population is uniformly sampled to remove a single sequence. The offspring sequence is then added to the population, keeping the total population size constant.

The above steps are repeated until the steady state is reached. Once in steady state, the population is sampled 10^4 times, skipping $1/\mu$ generations between consecutive samples.

3.6 Validation of the ABC inference pipeline using synthetic data

In order to validate our ABC approach to the genome-wide inference of evolutionary parameters, we first test our pipeline on simulated data. We use Eq. (3.1) with $M = 2$ fitness states to generate sampling probabilities for $n' = 5$ on a grid of rescaled selection coefficients, Ns, and the fraction of alleles with fitness $1 + s$, γ, for three values of the rescaled mutation rate θ. We use these probabilities directly as input to the ABC inference pipeline, which returns predictions for these 3 parameters. Note that this procedure is equivalent to producing an infinitely large sample of sets of 5 aligned sequences, so that we are not testing how sampling noise affects the accuracy of predictions (our sample size in each *D. melanogaster* genomic window, $B = 10^4$, is sufficiently large to minimize the sampling noise effects). Since we know the exact values of all 3 evolutionary parameters $\boldsymbol{\alpha} = (\theta, Ns, \gamma)$, we can systematically evaluate the accuracy of ABC inference across biologically relevant parameter ranges (Fig. 3.1). The same set of $m = 10^6$ simulated empirical histograms $\mathbb{P}[\mathbf{n}'|\boldsymbol{\alpha}_i]$ $(i = 1 \ldots m)$ is used here as in subsequent inference on fly genomic data (see Materials and Methods for details).

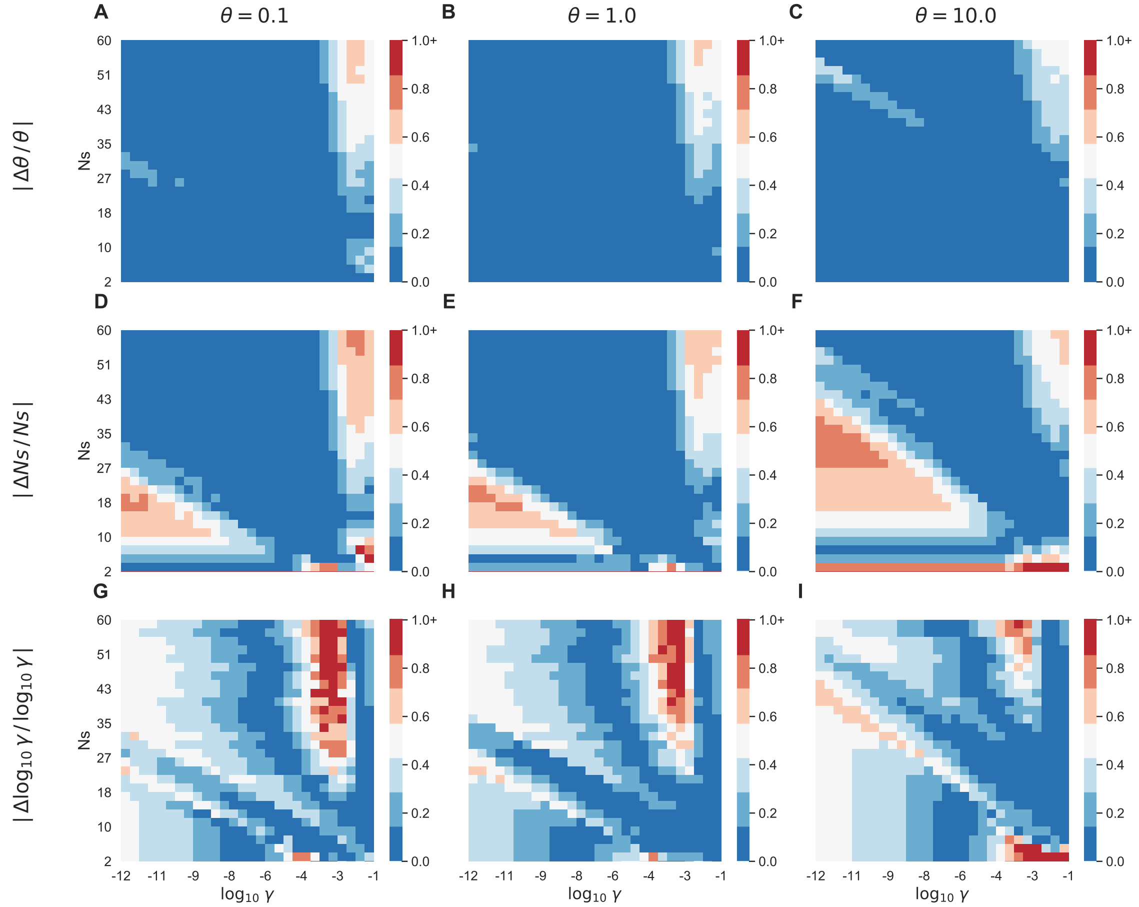

Figure 3.1: **Accuracy of ABC inference with selection on synthetic data**. Shown are relative errors $\Delta x/x = (x_{\text{inferred}} - x_{\text{true}})/x_{\text{true}}$ (where x_{inferred} is taken to be equal to the median value of the corresponding posterior distribution) for the rescaled mutation rate θ (A, D, G), the rescaled selection coefficient Ns (B, E, H), and the log-fraction of alleles with fitness $1 + s$, $\log_{10}\gamma$ (C, F, I). All relative errors are calculated for the same ranges of Ns and γ and for $\theta = 0.1, 1.0, 10.0$ corresponding to the monomorphic, intermediate, and polymorphic regimes.

Overall, the accuracy of our inference procedure is encouraging in this idealized setting, where the same model is used to generate the data and infer model parameters from it. In particular, we can reliably estimate mutation rates θ almost everywhere within the parameter ranges shown in Fig. 3.1A,D,G: the average relative error for θ inference is

≈ 0.09 over all 3 values of θ, while the maximum error is ≈ 0.66. For Ns and γ, there are parameter regions where the predictions yield significant errors, e.g. in the triangular-shaped area in Fig. 3.1B,E,H. It appears that the models are degenerate in these regions, such that multiple sets of parameters can fit input sampling probabilities in the $n' = 5$ histograms. As a result, the average relative error for Ns inference is ≈ 0.80, while the maximum error is ≈ 14.8. With $\log_{10} \gamma$ inference, the most prominent area where the ABC inference procedure yields significant errors is the vertical stripe in the large γ, large Ns quadrant (Fig. 3.1C,F,I). The area of the stripe shrinks as the population becomes more polymorphic. As with the Ns inference, the sampling probabilities within the stripe can be fit using multiple sets of parameters, so that the original parameter set is difficult to recover. The average and the maximum relative errors for $\log_{10} \gamma$ inference are ≈ 1.1 and ≈ 72.2, correspondingly.

We conclude from these numerical experiments that we should be able to infer the values of mutation rates from genomic data with a reasonably high degree of accuracy. Moreover, our predictions of selection coefficients, although subject to larger errors, should be accurate enough to serve as a test for the presence of natural selection signatures in genomic data. Finally, the predictions of the fraction of alleles with the higher fitness, γ, is the least reliable but also the least informative, since it depends on the assumption of two distinct fitness states. The structure of realistic fitness landscapes is likely to be considerably more complicated. Overall, the prediction accuracy is determined by which regions of the parameter space correspond to *D. melanogaster* genomic data.

To investigate the role of selection in predicting mutation rates, we have also carried out ABC inference using the neutral Ewens formula, Eq. (3.2), instead of its generalized version

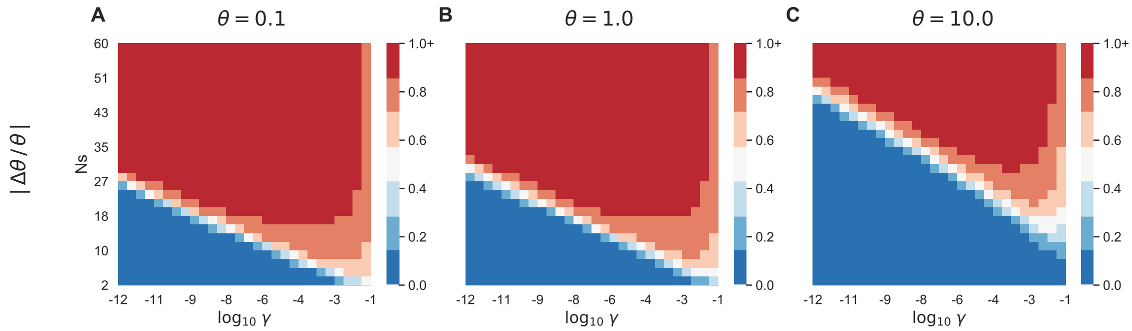

Figure 3.2: **Accuracy of neutral ABC inference on synthetic data**. Shown are relative errors $\Delta x/x = (x_{\text{inferred}} - x_{\text{true}})/x_{\text{true}}$ (where x_{inferred} is taken to be equal to the median value of the corresponding posterior distribution) for the rescaled mutation rate θ (A, B, C). All relative errors are calculated for the same ranges of Ns and γ as in Fig. 3.1 and for $\theta = 0.1, 1.0, 10.0$ corresponding to the monomorphic, intermediate, and polymorphic regimes.

with selection, Eq. (3.1), to generate $m = 10^6$ histograms $\mathbb{P}[\mathbf{n}'|\theta_i]$ using the values of θ_i ($i = 1 \ldots m$) previously employed for ABC inference with selection (Fig. 3.2). Since the input sampling probabilities are the same as in the previous test and therefore were generated under selection, we expect neutral inference to perform less well than the full treatment with selection. Indeed, considerable errors in θ predictions are produced under the neutral assumption in all three mutation regimes (Fig. 3.2A-C). The errors are small only in the regions where selection is weak and the fraction of high-fitness alleles is small: both of these factors make the mutational forces more dominant and the system is effectively neutral. We conclude that when population evolves under selection which is strong enough not to be dominated by mutational effects, neglecting selective forces (i.e., assuming neutrality) may easily lead to considerable errors in the inferred mutation rates.

3.6.1 Frequencies of allelic counts and ABC inference in the presence of recombination

Evolution of genomic sequences in *D. melanogaster* is subject to homologous recombination [91, 92, 93, 94, 95, 96]. Since our ABC inference pipeline does not include recombination explicitly (instead, we rely on the small window size, $L = 100$ bp, to minimize its effects in genome-wide analysis), we thought to investigate the influence of recombination on allelic count frequencies using a simple model system. To this effect, we have simulated a population of sequences with $L = 10$ sites subject to single-point mutation, recombination, and genetic drift (see Materials and Methods for details). Thus, each site in our population is equivalent to 10 bp in genomic windows. Since θ per bp is less than 0.01 per bp in *D. melanogaster* according to standard infinite-sites-based estimators [96] and therefore less than 1.0 per 100-bp genomic locus, we thought to investigate $n' = 5$ allelic count probabilities for $\theta = N\mu = (0.1, 1.0, 10.0)$ in our simulations.

Note that since the rate of spontaneous point-mutation events is $\approx 5 - 6 \times 10^{-9}$ per nucleotide per generation according to mutation-accumulation experiments [97, 98], the effective population size in the Zambian population under investigation is expected to be around 10^6 individuals [90], three orders of magnitude larger than the $N = 10^3$ population size we were able to implement in our simulations. Finally, fine-scale predictions of recombination rates using two *D. melanogaster* populations, one from North America and the other from Africa, yield ρ/μ estimates in the $\approx 2 - 4$ range [96]. Correspondingly, we have investigated $\rho/\mu = (0, 1, 2)$ cases for each of the three θ values mentioned above and for three values of selection strength: $Ns = (0, 6, 13)$ (Fig. 3.3). For each set of parameter values, we have compared theoretical predictions for $n' = 5$ allelic frequencies (Eq. (3.1)) with

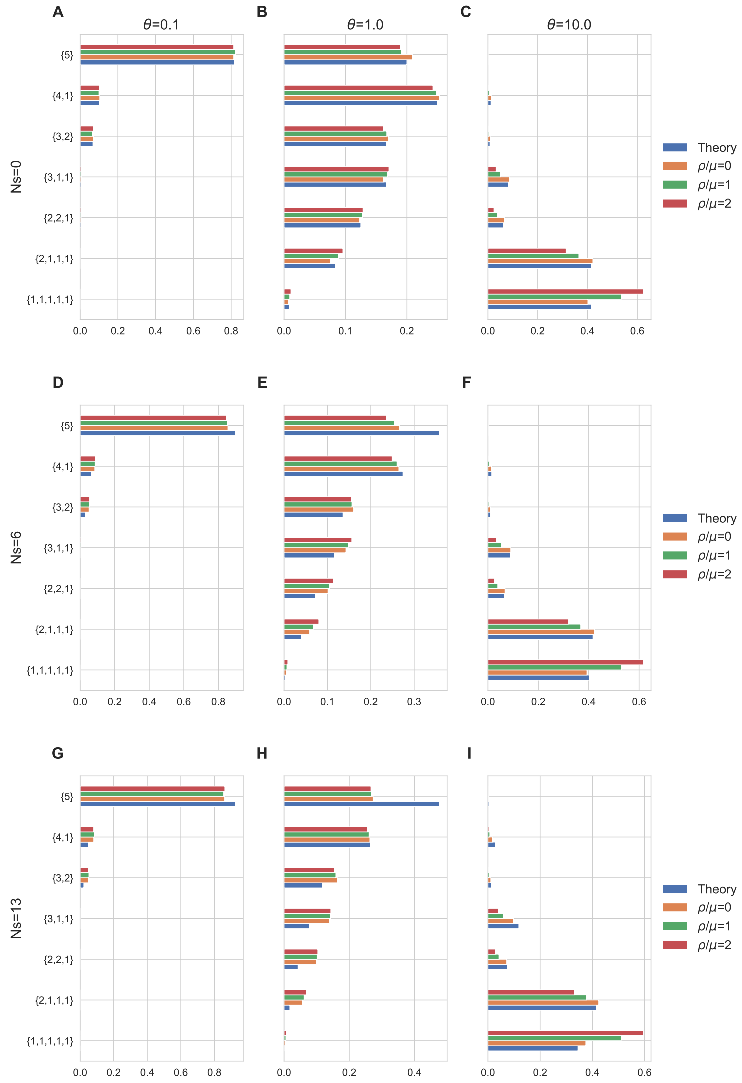

Figure 3.3: **Effect of recombination on sampling frequencies.** Columns of panels correspond to different selection strengths Ns and rows of panels to different values of the rescaled mutation rate θ. For cases with selection ($Ns = 6$ and 13), $\gamma = 0.242$ was used. In each panel, shown are sampling frequencies for the sample size $n' = 5$. Blue bars correspond to theoretical predictions (Eq. (3.1)), and orange, green and red bars correspond to numerical simulations averaged over 10^4 independent steady-state samples (see Materials and Methods for details). Simulations with $\rho/\mu = 0$: orange, $\rho/\mu = 1$: green, $\rho/\mu = 2$: red (ρ is the recombination rate and μ is the mutation rate per $L = 10$ locus).

steady-state numerical simulations. The $\rho/\mu = 0$ case with no recombination highlights the difference between the Ewens sampling approach, which assumes that each locus (i.e., nucleotide sequence in a 100-bp genomic window) can mutate into every other locus, and the numerical simulations in which only single-point mutations are allowed (see Ref. [82] for further analysis of this assumption).

Overall, we find reasonable agreement between predicted and observed allelic count frequencies. In the monomorphic limit ($\theta = 0.1$), the agreement is very good for all values of Ns and ρ/μ (Fig. 3.3A-C), indicating that the Ewens sampling approach would benefit from parsing the genome into even shorter, $10 - 20$ bp genomic windows that are characterized by θ values of similar magnitude. The agreement is significantly worse in the $\theta = 1.0$ regime when selection is present (Fig. 3.3D-F), indicating failure of the assumptions inherent in the Ewens sampling formula rather than the effects of recombination, which appear to be secondary in this case. The situation is reversed in the $\theta = 10.0$ regime (Fig. 3.3G-I), with recombination effects becoming more prominent. At the same time, deviations between generalized Ewens formula predictions and the $\rho/\mu = 0$ simulation are minimal in this regime, even in the presence of selection.

To investigate the effect of the observed discrepancies in allelic count frequencies on evolutionary parameter inference, we have used the $n' = 5$ allelic count frequencies from Fig. 3.3 as input to our ABC inference pipeline (Table 3.1). We find that θ is predicted very accurately for all parameter combinations, with the largest discrepancies observed when $\theta = 10.0$ and $\rho/\mu = 1$ or 2. For the rescaled selection coefficient Ns, the algorithm tends to predict non-zero values for a small fraction of alleles (e.g., $Ns = 4.01$ and $\gamma = 2.67 \times 10^{-9}$ in the case of neutral evolution with $\theta = 0.1$ and $\rho/\mu = 2$; Table 3.1, subsection A). In the ideal situation where, as in Fig. 3.1, Eq. (3.1) is used to both produce the $n' = 5$ allelic count frequencies and carry out ABC inference from them, the algorithm is able to make reasonably accurate predictions of both the selection strength and the fraction of viable alleles (cf. Theory columns in each subsection of Table 3.1).

$\boldsymbol{\theta = 0.1}$, $\boldsymbol{Ns = 0}$, $\boldsymbol{\gamma = 0}$

A	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	0.10	0.10	0.10	0.10
$(Ns)_m$	4.59	4.21	4.62	4.01
γ_m	1.24×10^{-7}	1.55×10^{-9}	2.00×10^{-6}	2.67×10^{-9}

$\boldsymbol{\theta = 0.1}$, $\boldsymbol{Ns = 6}$, $\boldsymbol{\gamma = 0.24}$

B	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	0.12	0.10	0.10	0.10
$(Ns)_m$	11.92	9.08	5.08	6.56
γ_m	0.23	0.58	0.02	0.53

$\boldsymbol{\theta = 0.1}$, $\boldsymbol{Ns = 13}$, $\boldsymbol{\gamma = 0.24}$

C	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	0.11	0.09	0.09	0.10
$(Ns)_m$	13.85	12.95	5.62	16.18
γ_m	0.20	0.59	0.16	0.61

$\boldsymbol{\theta = 1.0}$, $\boldsymbol{Ns = 0}$, $\boldsymbol{\gamma = 0}$

D	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	1.00	0.96	1.03	1.10
$(Ns)_m$	4.66	5.25	5.01	9.30
γ_m	4.59×10^{-8}	2.79×10^{-9}	4.13×10^{-9}	2.30×10^{-5}

$\boldsymbol{\theta = 1.0}$, $\boldsymbol{Ns = 6}$, $\boldsymbol{\gamma = 0.24}$

E	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	1.07	1.01	1.00	1.04
$(Ns)_m$	7.19	8.55	3.81	9.46
γ_m	0.27	0.58	0.09	8.80×10^{-5}

$\boldsymbol{\theta = 10.0}$, $\boldsymbol{Ns = 0}$, $\boldsymbol{\gamma = 0}$				
$\boldsymbol{\theta = 1.0}$, $\boldsymbol{Ns = 13}$, $\boldsymbol{\gamma = 0.24}$				
F	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	1.12	0.89	0.92	0.97
$(Ns)_m$	16.29	5.66	4.44	7.91
γ_m	0.22	0.54	0.03	4.48×10^{-4}
G	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	10.00	9.59	14.76	20.30
$(Ns)_m$	7.11	5.76	10.72	14.40
γ_m	1.00×10^{-6}	1.88×10^{-7}	0.03	0.06
$\boldsymbol{\theta = 10.0}$, $\boldsymbol{Ns = 6}$, $\boldsymbol{\gamma = 0.24}$				
H	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	9.86	9.31	14.56	19.61
$(Ns)_m$	10.03	7.73	13.25	14.21
γ_m	0.04	5.00×10^{-6}	0.01	0.02
$\boldsymbol{\theta = 10.0}$, $\boldsymbol{Ns = 13}$, $\boldsymbol{\gamma = 0.24}$				
I	**Theory**	$\boldsymbol{\rho/\mu = 0}$	$\boldsymbol{\rho/\mu = 1}$	$\boldsymbol{\rho/\mu = 2}$
θ_m	9.89	8.91	13.83	18.48
$(Ns)_m$	12.44	8.65	11.18	16.82
γ_m	0.24	0.01	0.07	0.03

Table 3.1: **ABC prediction accuracy in the presence of recombination.** Shown are median values of the model parameters: θ_m, $(Ns)_m$, and γ_m predicted using the ABC inference pipeline. Sampling frequencies ($n' = 5$) for each set of parameters in Fig. 3.3 served as input to the algorithm. To facilitate comparisons, each subsection of the Table is marked by the corresponding Fig. 3.3 panel label.

Predictions are also accurate in the $\theta = 0.1$ regime when allelic count frequencies from simulations with recombination are used as input to the ABC inference pipeline. However, the situation becomes more complicated when mutation rates increase: for example, in the $\theta = 1.0$ regime predicted Ns values alone are not a reliable indicator of selection strengths unless they are supplemented by the values of γ, which are much higher in the $Ns = 6$ and 13 cases compared to neutral evolution (Table 3.1, subsections D,E,F). Even

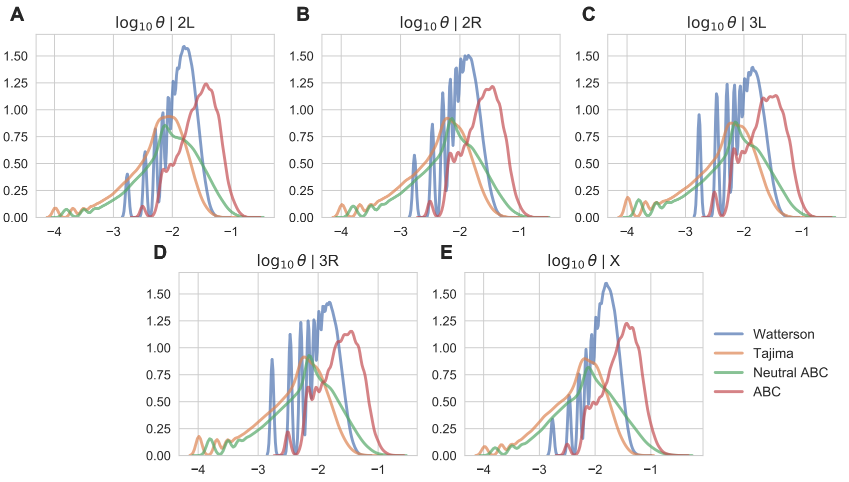

Figure 3.4: **Chromosome-wide distributions of mutation rates**. Shown are the histograms of Watterson (θ_W, blue), Tajima (θ_T, orange), ABC neutral (θ_n^{ABC}, green) and ABC with selection (θ_n^{ABC}, red) estimators of the logarithm of the mutation rate per bp rescaled by the population size, $\log_{10}\theta$, for chromosomes 2L (A), 2R (B), 3L (C), 3R (D), and X (E). The two ABC estimators are represented by their median values in each genomic window.

considered together, the Ns and γ predictions become unreliable in the $\theta = 10.0$ regime (Table 3.1, subsections G,H,I), indicating that we cannot distinguish selection strength differences of $\mathcal{O}(1)$ in this case. In summary, our analysis indicates that it is preferable to combine evidence from both Ns and γ predictions, and that it is prudent to limit the observations to qualitative conclusions such as presence or absence of selection, especially in windows with $\theta \gtrsim 10.0$.

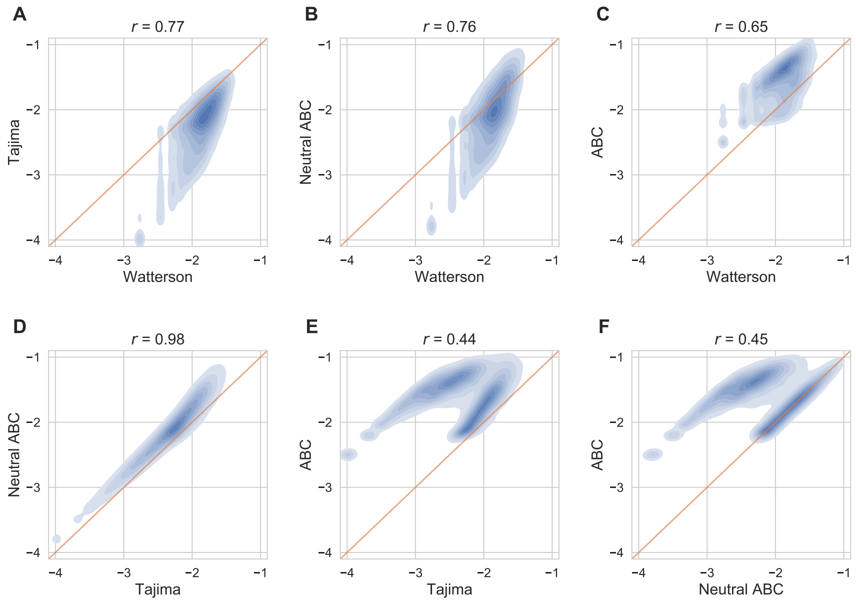

Figure 3.5: **Comparison of mutation rate estimators.** Shown are genome-wide density correlation plots (based on a random sample of 10^4 windows) between Watterson and Tajima estimators (A), Watterson and neutral ABC estimators (B), Watterson and ABC with selection estimators (C), Tajima and neutral ABC estimators (D), Tajima and ABC with selection estimators (E), neutral ABC and ABC with selection (F). r is the linear correlation coefficient, computed for all windows. Both ABC estimators are represented by their median values in each genomic window. All estimators are shown on the $\log_{10}$ scale. Thin red lines in all panels have unit slopes.

3.7 Evolutionary parameter inference in *D. melanogaster*

3.7.1 Inference of mutation rates

We have carried out genome-wide inference of rescaled mutation rates using several alternative methods. First, we use standard mutation rate estimators due to Watterson [99], which is based on the number of segregating sites observed in each window, and due to Tajima [100], which is based on the average number of nucleotide differences between all pairs of sequences. The normalized difference of the two estimators defines Tajima's D

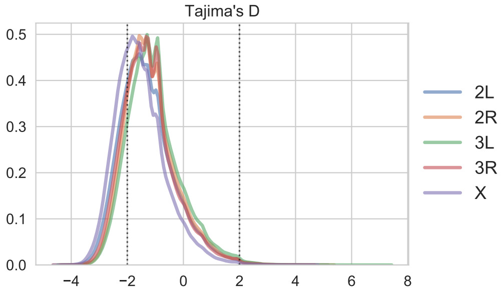

Figure 3.6: **Chromosome-wide distributions of Tajima's D statistic.** Vertical dotted lines indicate $D = -2$ and $D = 2$; all windows with $D < -2$ or $D > 2$ are assumed to be under significant selection. The fractions of windows with $D < -2$ are 0.24, 0.17, 0.15, 0.20 and 0.32 for chromosomes 2L, 2R, 3L, 3R and X, respectively. The corresponding fractions of windows with $D > 2$ are 0.004, 0.004, 0.006, 0.003, 0.002.

statistic widely used in tests for selection [86, 101]. For each valid 100 bp window with n aligned unambiguous sequences (see Materials and Methods for details), Watterson estimator is given by $\theta_W = S_n/H_{n-1}$, where S_n is the number of segregating sites and $H_n = \sum_{i=1}^{n} 1/i$ is the harmonic number. Tajima estimator, θ_T, is simply the average number of polymorphisms (nucleotide differences) in all pairwise alignments generated by n aligned 100-bp sequences in each window. Tajima and Watterson estimators are compared with two ABC-based inferences of θ, one employing the neutral formula (Eq. (3.2); θ_n^{ABC}) and the other one the formula with selection (Eq. (3.1); θ_s^{ABC}) (Fig. 3.4, Fig. 3.5).

First of all, we observe that, surprisingly, Tajima's estimator of the mutation rate, θ_T, is strongly correlated with the neutral ABC estimate, θ_n^{ABC}. In fact, the genome-wide linear correlation between these two approaches is $r = 0.98$ (Fig. 3.5D). Although the two approaches are mathematically different (using pairwise polymorphisms in the former and allelic counts in the latter), they yield very similar mutation rate estimates. In contrast, the correlation coefficient between θ_W and θ_T is much smaller ($r = 0.77$;

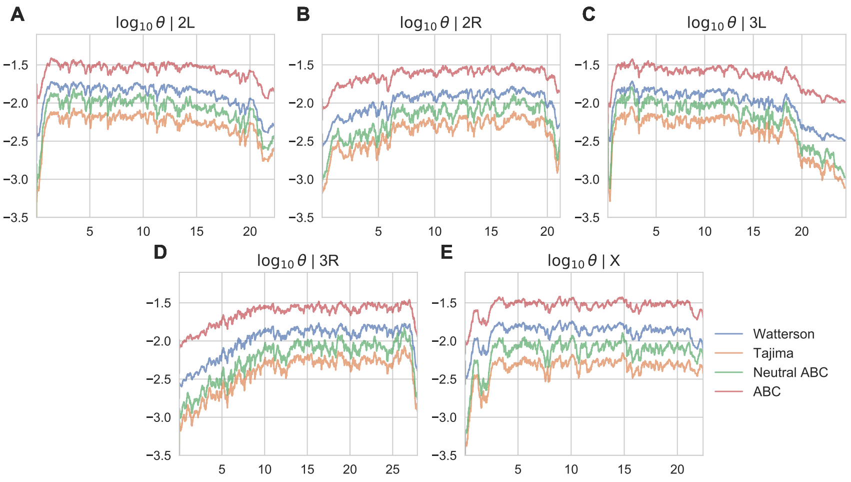

Figure 3.7: **Inferred mutation rates along *D. melanogaster* chromosomes**. Shown are Watterson (θ_W, blue), Tajima (θ_T, orange), ABC neutral (θ_n^{ABC}, green) and ABC with selection (θ_n^{ABC}, red) estimators of the logarithm of the mutation rate per bp rescaled by the population size, $\log_{10}\theta$, for each 100-bp window, for chromosomes 2L (A), 2R (B), 3L (C), 3R (D), and X (E), plotted vs. the genomic coordinate along each chromosome (in Mbp). The two ABC estimators are represented by their median values in each genomic window. All plotted values were smoothed with an exponentially weighted moving average with the center of mass of 1,000 windows, such that the exponential parameter $\alpha \simeq 10^{-3}$.

Fig. 3.5A) and $\theta_T - \theta_W < 0$ in most cases, indicating that Tajima's D statistic, which is $\sim \theta_T - \theta_W$, is affected by selection against genotypes carrying deleterious mutant alleles (for the purposes of interpreting this test, we assume that the Zambian *D. melanogaster* population under consideration is approximately stable in size). Indeed, the distribution of Tajima's D statistic for each chromosome is skewed towards negative values and its magnitude strongly suggests selective effects in a significant fraction of windows (Fig. 3.6). This can be explained by the fact that the number of segregating sites on which Watterson's estimator is based ignores the frequency of mutations and therefore is expected to be more strongly affected by the existence of rare deleterious mutants than the average number of pairwise nucleotide differences [100].

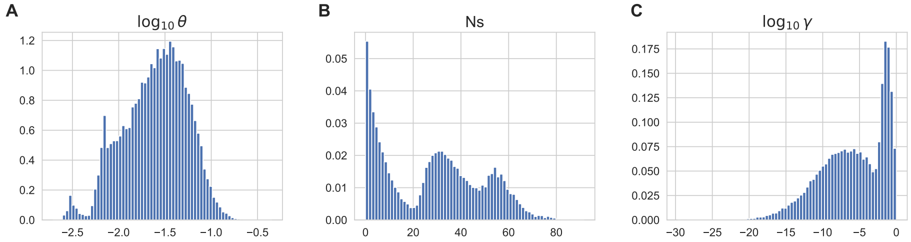

Figure 3.8: **Genome-wide posterior distributions of mutation rates, selection coefficients, and fraction of viable genotypes.** Each genome-wide posterior was created by combining individual posterior probabilities for each 100 bp window, which amounts to marginalizing all posteriors over the window index. (A): $\log_{10}\theta$ distribution (θ is the rescaled mutation rate per bp), (B): Ns distribution, (C): $\log_{10}\gamma$ distribution.

Both θ_W and θ_s^{ABC} tend to predict consistently higher values of mutation rates than the neutral ABC estimate (Fig. 3.4). This is clearly seen when inferred mutation rates are plotted along each chromosome (Fig. 3.7): ABC inference with selection predicts the highest mutation rates, followed by Watterson's estimate. However, θ_W is only moderately correlated with the ABC estimate under selection, θ_s^{ABC} ($r = 0.65$, Fig. 3.5C), indicating that both estimates are affected by selective forces in somewhat different ways. This is not surprising since, as mentioned above (and unlike ABC inference with selection), Watterson's estimator ignores the frequency of mutations. Remarkably, ABC inference with selection produces distributions of mutation rates that are nearly identical from chromosome to chromosome, indicating that the inference process is dominated by global polymorphism patterns rather than chromosome-specific features (Fig. 3.9A). To ensure that we do not lose information by focusing on the median values of rescaled mutation rates in each genomic window and to estimate the uncertainty of our predictions, we have constructed the genome-wide posterior probability by combining data from all windows (Fig. 3.8A). We find that the genome-wide distribution is essentially unimodal, with the shape similar to

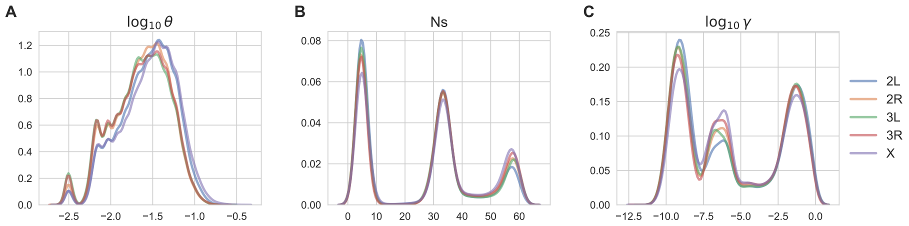

Figure 3.9: **ABC inference of mutation rates, selection coefficients, and fraction of viable genotypes.** Chromosome-wide distributions of $\log_{10}\theta$ per bp (A), Ns (B) and $\log_{10}\gamma$ (C) based on the generalized Ewens formula with selection (Eq. (3.1)). All predicted quantities are represented by their median values in each window. The curves for each chromosome represent a smoothed trace of a normalized histogram.

those seen in Fig. 3.9A and with predicted $\log_{10}\theta$ values predominantly concentrated in the $[-2.3, -0.8]$ range.

3.7.2 Inference of selection strengths and the fraction of viable genotypes

We observe three distinct peaks of selection strengths in each chromosome: weak selection (peak 1; $Ns < 15$), intermediate selection (peak 2; $15 \leq Ns \leq 45$), and strong selection (peak 3; $Ns > 45$) (Fig. 3.9B). As with the rescaled mutation rates, the peak structure is similar in all chromosomes. Thus ABC inference predicts that most of the fly genome evolves under selective constraints, in accordance with previous studies (see Ref. [61] for a comprehensive review). Interestingly, three distinct peaks are also observed in the genome-wide posterior probability of selection strengths (Fig. 3.8B); the probability that $Ns > 1$ is 94.2% according to this distribution. The fraction of genotypes in the high-fitness state (which we shall refer to as viable genotypes), γ, or, alternatively, the fraction of neutral mutations for a viable allele, also exhibits a characteristic peak structure which is fairly similar for all chromosomes (Fig. 3.9C). However, this structure is not observed in the

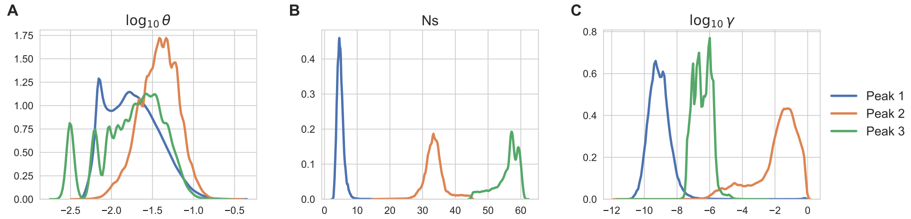

Figure 3.10: **ABC inference of mutation rates, selection coefficients, and fraction of viable genotypes partitioned by selection strength.** Genome-wide distributions of ABC inference results (median parameter values in each genomic window) for $\log_{10}\theta$ (θ is the rescaled mutation rate per bp) (A), Ns (B) and $\log_{10}\gamma$ (C) partitioned by the range of Ns: $Ns < 15$ (blue), $15 \leq Ns \leq 45$ (orange) and $Ns > 45$ (green). Each blue, orange and green curve is a smoothed trace of a normalized histogram.

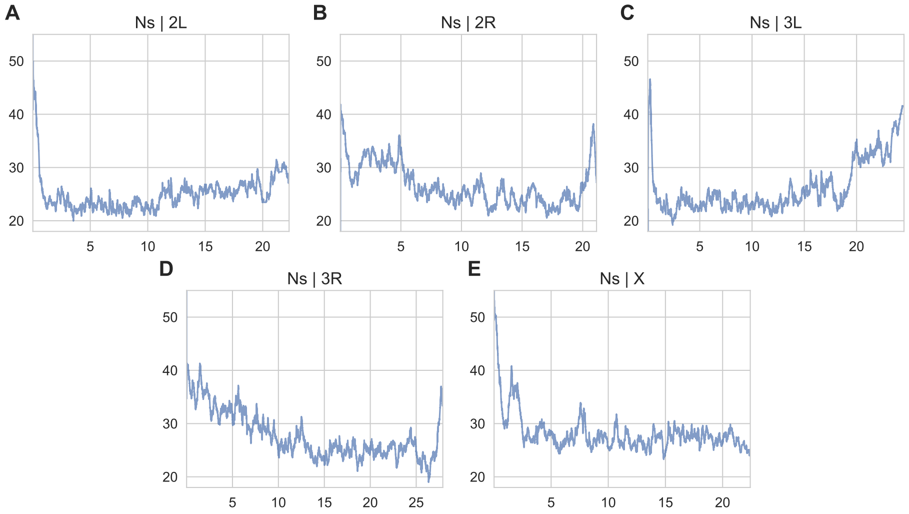

Figure 3.11: **Inferred selection coefficients along *D. melanogaster* chromosomes**. Shown are median values of selection coefficients (rescaled by the population size) predicted using ABC with selection, for each 100-bp window, for chromosomes 2L (A), 2R (B), 3L (C), 3R (D), and X (E), plotted vs. the genomic coordinate along each chromosome (in Mbp). All plotted values were smoothed with an exponentially weighted moving average with the center of mass of 1,000 windows, such that the exponential parameter $\alpha \simeq 10^{-3}$.

genome-wide posterior distribution, which is bimodal with a narrow peak in the [-3.0,0.0] range and a much broader peak in the [-20.0,-3.0) range (Fig. 3.8C).

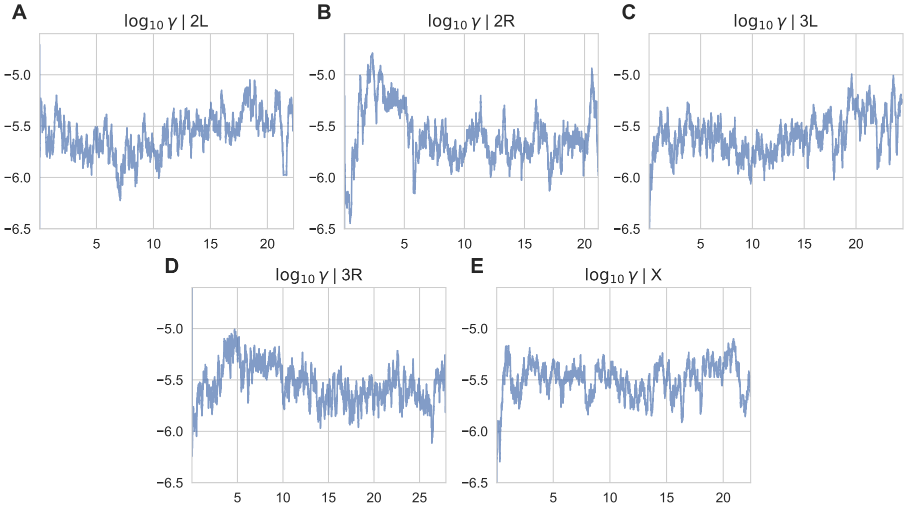

Figure 3.12: **Inferred fractions of high-fitness alleles along *D. melanogaster* chromosomes**. Shown are median values of the log fraction of viable genotypes, $\log_{10}\gamma$, predicted using ABC with selection, for each 100-bp window, for chromosomes 2L (A), 2R (B), 3L (C), 3R (D), and X (E), plotted vs. the genomic coordinate along each chromosome (in Mbp). All plotted values were smoothed with an exponentially weighted moving average with the center of mass of 1,000 windows, such that the exponential parameter $\alpha \simeq 10^{-3}$.

To investigate whether sequences in different fitness peaks correspond to distinct distributions of mutation rates and fractions of viable genotypes, we have divided all windows into 3 classes according to selection strength (Fig. 3.10). We observe that mutation rates do not correlate strongly with Ns peak identity, although sequences with intermediate selection strengths do tend to have somewhat higher mutation rates (Fig. 3.10A). In contrast, fractions of viable genotypes are partitioned by selection strength, with the sequences under strong selection characterized by intermediate values of $\log_{10}\gamma$ (Fig. 3.10C). In the light of our previous discussion of prediction accuracy on synthetic data with and without recombination (Fig. 3.3, Table 3.1), the intermediate-selection peak may be the most reliable since it is accompanied by sizable values of $\log_{10}\gamma$. With peaks 1 and 3, we cannot rule out

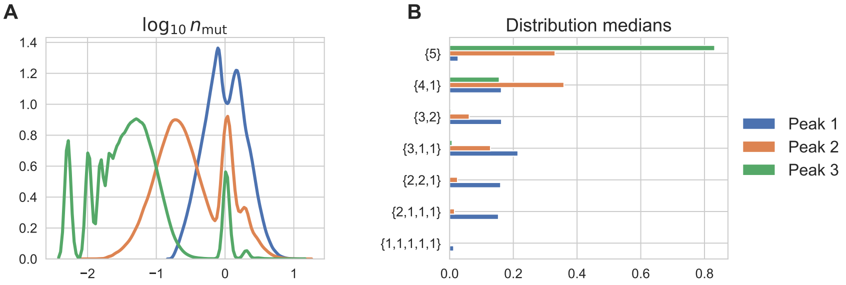

Figure 3.13: **Polymorphisms in sequence alignments and selection strength.** Histograms of the average number of mutations per sequence, n_{mut}, for 100-bp genomic windows associated with different selection peaks (Fig. 3.9B, Fig. 3.10B) (A). Median values of the distribution of $n' = 5$ allelic counts for all windows in the three selection peaks from Fig. 3.9B (B). In both panels, data from all chromosomes is combined. In panel B, each 100-bp window has a set of 10^2 $\mathbb{P}[\mathbf{n}'|\boldsymbol{\alpha}_i]$ distributions corresponding to 10^2 sets of model parameters $\boldsymbol{\alpha}_i$ with the smallest d^2 score (see Materials and Methods for details). These sets of histograms are combined into a single dataset for all windows that belong to a given selection peak and median values of the frequency distribution for each allelic configuration in the $n' = 5$ partition are reported.

the possibility that in some genomic windows, similar to predictions in Table 3.1, neutral evolution is in fact modeled by non-zero selection coefficients accompanied by low values of γ. In addition, selection strengths in peak 1 may be insufficient to reliably rule out the no-selection scenario. Finally, we note that plots of Ns and $\log_{10}\gamma$ vs. chromosome coordinates show no easily identifiable trends, except for the higher values of Ns accompanied by somewhat lower values of $\log_{10}\gamma$ in both sub-telomeric regions (Figs. 3.11,3.12).

Next, we have investigated the nature of genomic sequences that belong to the three selection peaks. We find that, as might be expected, the strength of selection is inversely correlated with the number of mutations observed in corresponding genomic sequences. Indeed, sequences evolving under the strongest selection (peak 3) are significantly less polymorphic than sequences predicted to be under weak selection (peak 1), with sequences in peak 2 occupying an intermediate position (Fig. 3.13A). Besides the number of mutations

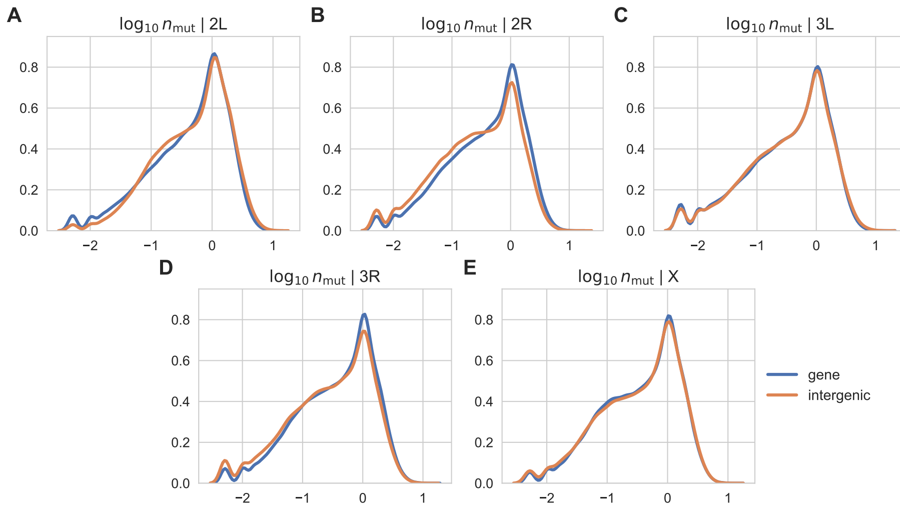

Figure 3.14: **Chromosome-wide distributions of the number of mutations partitioned into functional regions.** For each chromosome, shown is the average number of mutations per sequence in each alignment, n_{mut}, for genes (blue) and intergenic regions (orange). Only windows that fully overlap the functional region of interest (gene or intergenic) are included. Panels A through E show the n_{mut} distributions for chromosomes 2L, 2R, 3L, 3R and X, respectively.

per sequence, we have considered the distribution of allelic partitions in $n' = 5$ sequence alignments used in our inference procedure (Materials and Methods) (Fig. 3.13B). We observe that $n' = 5$ allelic counts that correspond to sequences under the strongest selective constraint (peak 3) are generated by sequences that are either all identical in the alignment ($\{5\}$) or with a single different sequence ($\{4, 1\}$). In contrast, sequence alignments in peak 1 are predominantly polymorphic, and sequence alignments in peak 2 occupy an intermediate position. Since mutation rates are predicted to be polymorphic for most windows (i.e., $\log_{10}\theta + 2 > 0$, where θ is the rescaled mutation rate per bp) regardless of their peak identity (Fig. 3.10A), the number of mutations in the alignment is, according to the ABC inference pipeline, indicative of selective constraints rather than the monomorphic limit.

	2L		**2R**		**3L**	
	gene	intergenic	gene	intergenic	gene	intergenic
peak 1	1.06 (0.65, 1.71)	1.15 (0.69, 1.87)	0.98 (0.60, 1.64)	0.86 (0.54, 1.48)	0.97 (0.60, 1.63)	0.97 (0.60, 1.66)
peak 2	0.30 (0.16, 0.97)	0.34 (0.17, 1.04)	0.27 (0.14, 0.92)	0.22 (0.12, 0.62)	0.25 (0.13, 0.88)	0.26 (0.13, 0.89)
peak 3	0.04 (0.02, 0.08)	0.06 (0.03, 0.11)	0.04 (0.02, 0.08)	0.04 (0.02, 0.07)	0.03 (0.02, 0.06)	0.03 (0.02, 0.07)

	3R		**X**	
	gene	intergenic	gene	intergenic
peak 1	0.94 (0.60, 1.58)	0.87 (0.55, 1.50)	0.99 (0.64, 1.64)	0.99 (0.64, 1.64)
peak 2	0.28 (0.15, 0.93)	0.24 (0.13, 0.79)	0.31 (0.16, 0.96)	0.31 (0.16, 0.96)
peak 3	0.05 (0.02, 0.09)	0.04 (0.02, 0.07)	0.05 (0.03, 0.09)	0.05 (0.03, 0.09)

Table 3.2: **Summary statistics for the average number of mutations per sequence, partitioned by functional region and selection strength.** Shown are the median values, followed by the first quartile and the third quartile in parentheses, of the distribution of n_{mut}, the average number of mutations per sequence in each 100 bp window, for all windows in a given chromosome, sorted by the selection peaks in Fig. 3.9B and by the functional region (gene or intergenic).

It would be natural to expect that sequences under stronger selective constraints are predominantly associated with functional genomic regions, such as coding sequences and promoters. However, we do not find any correlation between the average number of mutations per sequence in each 100-bp window and its location within either a genic or an intergenic region (Fig. 3.14; we employ FlyBase annotation v. 6.29 to map functional regions [68]). In fact, the distributions of the average number of mutations in genes and intergenic regions are remarkably similar for each chromosome (qualitatively similar results are obtained when considering exons and introns separately; data not shown). Moreover, distributions of the average number of mutations sorted by selection peaks in Fig. 3.10B are strongly overlapping in each chromosome (Table 3.2). We conclude that sequences under weak, intermediate and strong selection are distributed throughout the fly genome in a way that is independent of their standard functional annotation.

Finally, since our ABC inference pipeline produces sizable errors in some areas of the (θ, Ns, γ) parameter space (cf. Fig. 3.1), it is possible that our results are affected by inaccuracies in ABC computational predictions. However, a comparison of the ranges of predicted parameters in Fig. 3.10 with the prediction errors on synthetic data in Fig. 3.1 shows that most of our predictions are not concentrated in the problematic regions of the parameter space. For example, windows in peak 3 have $0.3 \lesssim \theta \lesssim 10$, $45 \lesssim Ns \lesssim 60$, and $-8 \lesssim \log_{10} \gamma \lesssim -5$. A comparison with the error plots in Fig. 3.1 shows that we can expect excellent accuracy for θ and Ns inference and reasonable accuracy for $\log_{10} \gamma$ inference within these ranges. The same is true, by and large, of the other two peaks. We conclude that the ABC inference procedure applied to *D. melanogaster* genomic data has sufficient internal consistency for at least qualitative conclusions regarding the magnitude of mutation and selection forces. This observation does not however preclude the possibility that our results are affected by the phenomena that are not explicitly included into the ABC model formulated above, such as recombination [91, 92, 93, 94, 95, 96], demographic effects [102, 90], and the assumption that any sequence in the 100-bp window can mutate into every other sequence [8].

3.8 Discussion and Conclusion

In this work, we have developed a novel computational approach to simultaneous genome-wide inference of mutation rates, selection strengths and the average fractions of beneficial, deleterious and neutral mutations per allele. The approach is based on applying Approximate Bayesian Computation [83, 84, 85] to the Ewens sampling formula which we have previously generalized to evolution under selection [82]. The generalized Ewens sampling formula provides an explicit closed-form solution for the probability of each partition of n alleles (for example, aligned sequences in a genomic window) into allelic counts. However, it is cumbersome to implement, requiring (i) a partition of n aligned sequences into all possible allelic partitions and (ii) for each allelic partition, a sum over all the ways in which the partition can be distributed among different fitness states. The ABC inference pipeline alleviates these computational difficulties since it was specifically designed for cases where the probability of the statistic of interest is either not a closed-form expression or computationally costly to evaluate.

Furthermore, we have assumed that all alleles can adopt either a low- or high-fitness state, so that, with a sufficient fitness difference between the two states, sequences in the population will predominantly concentrate in the high-fitness state and, for such sequences, the newly arising mutations will be either neutral or deleterious. In this aspect, our fitness landscape conforms to a central tenet of the neutral theory which assumes that the contribution of beneficial mutations can be ignored [64]. Note, however, that the magnitude of the difference between the two fitness states is inferred from the data rather than imposed, enabling us to differentiate between the strictly neutral scenario and its generalization to both deleterious and neutral evolutionary dynamics. Finally, the Ewens sampling approach is based on the steady-state assumption: although specific sequences that make up the evolving population change, the “de-labeled” statistics such as the average number of distinct alleles in the population is time-independent [8, 46]. Note that a generalization to more than two fitness states is not likely to be qualitatively different since the population will always adopt the highest-fitness configuration in steady state, with the mutational load due to deleterious mutations into all lower-fitness fitness.

We have applied the ABC inference approach to study selective constraints on the genomic evolution of *D. melanogaster*. *D. melanogaster* is a key model organism in modern genetics and as a result evolution of fruit fly populations in the wild has received considerable attention in the population genetics community, both experimentally and computationally. In particular, a considerable number of fly genomes have been sequenced, aligned and functionally annotated to a common standard in a large-scale effort [66, 67]. Specifically, phase 3 of the *Drosophila* Population Genomics Project (DPGP3) has provided 197 haploid embryo genomes from a single *D. melanogaster* population in Zambia, Sub-Saharan Africa. *D. melanogaster* likely originated in the Sub-Saharan region [103], so that the genome sample is from the species’s ancestral range. This data provides a rich collection of polymorphisms and allelic counts in a single fruit fly population. The allelic counts serve as input to the ABC inference pipeline developed in this work. To carry out ABC analysis, we have parsed the *D. melanogaster* genome into 100-bp non-overlapping windows. The size of the windows was chosen to minimize the effects of recombination, which is not explicitly treated in the Ewens sampling framework, while still dealing, in each window, with a polymorphic sample that provides informative allelic counts.

Furthermore, linkage between mutations that belong to the same window is fully taken into account, going beyond the other major assumption of the neutral theory, that positive and negative selection at linked loci does not affect the dynamics of neutral alleles [64]. As a result, our approach is closer to the background selection framework, which explicitly treats the effects of recombination and linkage but puts emphasis on negative rather than positive selection [42, 76, 77, 80]. Similar to background selection, our computational procedure can be viewed as a baseline model, deviations from which would be indicative of positive selection events such as selective sweeps. We find that, consistent with previous studies (reviewed in Ref. [61]), a large fraction of the *Drosophila* genome appears to evolve under selective constraints. Similar to previous work [80], we find that purifying selection can explain the observed patterns of nucleotide diversity in the *Drosophila* population under consideration. The major role of deleterious mutations is expected given that deleterious and neutral mutations are typically much more numerous than beneficial ones [21]. We observe that sequences under selective constraints are not preferentially associated with coding regions or other functional elements, or with centromeric or telomeric positions (although selection does appear to be stronger at sub-telomeric regions, Fig. 3.11), and are instead distributed evenly throughout the genome. All sequences under selection are grouped into three distinct peaks, with weak, intermediate, and strong selection (Fig. 3.9B). The peaks of selection strength correlate with the total number of polymorphisms observed in a genomic window and with the frequencies of allelic counts, with sequences under weaker selection generally being more polymorphic (Fig. 3.13). These global constraints may reflect the need to maintain nucleosome positioning [104] or higher-order chromatin structure [105], or other universal constraints whose exact nature is currently unclear.

The ability to treat linkage and epistasis within genomic windows of arbitrary width (constrained only by computational considerations) provides a substantial advantage over the Poisson Random Field approach, [45, 87, 88, 89] which can infer the strength of selection (rather than merely detect its presence) but is unable to account for linkage between sites. In addition to providing a quantitative Bayesian estimate of selection strength, our inference pipeline yields simultaneous estimates of population-size-rescaled mutation rates and of the fractions of neutral, deleterious and beneficial mutations for each high- and low-fitness allele in a steady-state population. Overall, our θ estimates yield 2-3 fold higher values genome-wide compared to standard neutral estimates [99, 100] and our own ABC estimate without selection (Fig. 3.7). We conclude that ignoring selection against deleterious mutations leads to consistent underestimation of effective population sizes. Finally, we find that as a rule, there are many more deleterious than neutral mutations available to an allele in a high-fitness state. Interestingly, it is the alleles subjected to intermediate levels of selection that are the most robust to mutations (i.e., have the largest number of neutral mutations available to them) (Fig. 3.10C).

Our inference relies on the steady-state assumption and therefore our estimates may become inaccurate if the population is in the process of expansion or contraction. However, our framework should be able to account for past changes in the population size, such as bottlenecks, through adjusting the effective population size. *Drosophila* demographics is a potential compounding factor because not only derived fruit fly populations have been associated with severe bottlenecks [102], but the ancestral range population in Sub-Saharan Africa is also predicted to have undergone a significant bottleneck [90]. These past events should reduce the magnitude of the effective population size in our framework;

our predictions of θ (and using mutation rates per nucleotide from previous mutation-accumulation studies [97, 98]) yield $N_{\text{eff}} \approx 10^6$, reasonably consistent with the population size estimates from Ref. [90].

In summary, we have developed an ABC inference framework for simultaneous genome-wide prediction of selection strengths, mutation rates, and the fraction of viable alleles. The framework is based on the Ewens sampling formula, which we previously generalized to evolution under selection [82]. Applying this approach to the evolutionary dynamics of a single *Drosophila* population, we observe, in line with previous reports, that a major fraction of the fly genome evolves under purifying selection against the constant influx of deleterious mutations. Moreover, we have found that genomic sequences can be classified into three distinct classes on the basis of their selection strength and investigated the effect of selection on mutation rate estimates. The accuracy of our predictions has been verified against synthetic data, which allowed us to systematically test all the major assumptions inherent in the model and gauge their potential effect on the accuracy of genome-wide predictions. Our computational approach can be used in other organisms for which population-level genomic data is available, providing an alternative to the Poisson Random Field and neutral approaches for explicit inference of key population-genetic parameters.

Bibliography

[1] S. Wright, Genetics **16**, 97 (1931).

[2] R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).

[3] P. A. P. Moran, Math Proc Cambr Philos Soc **54**, 60 (1958).

[4] S. Wright, in *Genetics, Paleontology and Evolution*, edited by G. L. Jepson, G. G. Simpson, and E. May (Princeton University Press, 1949) pp. 365–389.

[5] G. A. Watterson, Genetics **85**, 789 (1977).

[6] J. F. C. Kingman, J Royal Stat Soc B **37**, 1 (1975).

[7] J. F. C. Kingman, Theor Pop Biol **11**, 274 (1977).

[8] W. J. Ewens, Theor Pop Biol **3**, 87 (1972).

[9] A. I. Podgornaia and M. T. Laub, Science **347**, 673 (2015).

[10] O. Puchta, B. Cseke, H. Czaja, D. Tollervey, G. Sanguinetti, and G. Kudla, Science **352**, 840 (2016).

[11] C. Li, W. Qian, C. J. Maclean, and J. Zhang, Science **352**, 837 (2016).

[12] K. S. Sarkisyan *et al.*, Nature **533**, 397 (2016).

[13] Y. H. Chan, S. V. Venev, K. B. Zeldovich, and C. R. Matthews, Nat Comm **8**, 14614 (2017).

[14] M. Lunzer, S. P. Miller, R. Felsheim, and A. M. Dean, Science **310**, 499 (2005).

[15] P. A. Romero and F. H. Arnold, Nat Rev Mol Cell Biol **10**, 866 (2009).

[16] M. Lunzer, G. B. Golding, and A. M. Dean, PLoS Genet **6**, e1001162 (2010).

[17] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, Nature **490**, 535 (2012).

[18] P. A. Lind, O. G. Berg, and D. I. Andersson, Science **330**, 825 (2010).

[19] R. T. Hietpas, J. D. Jensen, and D. N. A. Bolon, Proc Nat Acad Sci USA **108**, 7896 (2011).

[20] R. Sanjuan, A. Moya, and S. F. Elena, Proc Nat Acad Sci USA **101**, 8396 (2004).

[21] A. Eyre-Walker and P. D. Keightley, Nat Rev Genet **8**, 610 (2007).

[22] G. Sella, D. A. Petrov, M. Przeworski, and P. Andolfatto, PLoS Genet **5**, e1000495 (2009).

[23] S. Wright, Proc Nat Acad Sci USA **23**, 307 (1937).

[24] M. Kimura, Quant Biol **20**, 33 (1955).

[25] W.-H. Li, Genetics **90**, 349 (1978).

[26] W.-H. Li, Proc Nat Acad Sci USA **74**, 2509 (1977).

[27] W.-H. Li, Genetics **92**, 647 (1979).

[28] W. J. Ewens, *Mathematical Population Genetics: I. Theoretical Introduction*, 2nd ed. (Springer, 2004).

[29] M. Slatkin, Genet Res Cambr **64**, 71 (1994).

[30] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions* (Wiley, New York, NY, 1997).

[31] H. Crane, Stat Sci **31**, 1 (2016).

[32] S. Feng, Stat Sci **31**, 20 (2016).

[33] G. A. Watterson, Genetics **88**, 405 (1978).

[34] M. N. Grote and T. P. Speed, Ann Appl Prob **12**, 637 (2002).

[35] P. Joyce and S. Tavare, J Math Biol **33**, 602 (1995).

[36] P. Joyce, J Appl Prob **32**, 609 (1995).

[37] P. Joyce, S. M. Krone, and T. G. Kurtz, Ann Appl Prob **13**, 181 (2003).

[38] K. Handa, Elect Comm in Prob **10**, 223 (2005).

[39] T. Huillet, J Comp Appl Math **206**, 755 (2007).

[40] S. N. Ethier and T. G. Kurtz, Stochastic Models in Biology, Lecture Notes in Biomathematics **70**, 72 (1987).

[41] S. N. Ethier and T. G. Kurtz, Stoch Proc Appl **54**, 1 (1994).

[42] B. Charlesworth, M. T. Morgan, and D. Charlesworth, Genetics **134**, 1289 (1993).

[43] R. Hudson and N. Kaplan, in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding (Chapman and Hall, New York, NY, 1994) pp. 140–153.

[44] M. M. Desai, L. E. Nicolaisen, A. M. Walczak, and J. B. Plotkin, Theor Pop Biol **81**, 144 (2012).

[45] S. A. Sawyer and D. L. Hartl, Genetics **132**, 1161 (1992).

[46] W. J. Ewens and W.-H. Li, J Math Biol **10**, 155 (1980).

[47] R. C. Griffiths, Journal of Mathematical Biology **17**, 1 (1983).

[48] A. Genz and P. Joyce, Comp Sci Stat **35**, 181 (2003).

[49] P. Joyce, A. Genz, and E. O. Buzbas, J Comp Biol **16**, 650 (2012).

[50] P. R. Haddrill, L. Loewe, and B. Charlesworth, Genetics **185**, 1381 (2010).

[51] R. Ronen, N. Udpa, E. Halperin, and V. Bafna, Genetics **195**, 181 (2013).

[52] A. F. Feder, S. Kryazhimskiy, and J. B. Plotkin, Genetics **196**, 509 (2014).

[53] R. Vitalis, M. Gautier, K. Dawson, and M. A. Beaumont, Genetics **196**, 799 (2014).

[54] S. H. Martin, M. Möst, W. J. Palmer, C. Salazar, W. O. McMillan, F. M. Jiggins, and C. D. Jiggins, Genetics **203**, 525 (2016).

[55] J. G. Schraiber, S. N. Evans, and M. Slatkin, Genetics **203**, 493 (2016).

[56] J. H. Gillespie, *Population Genetics: A Concise Guide* (The Johns Hopkins University Press, Baltimore, MD, 2004).

[57] M. Kimura, Genetics **47**, 713 (1962).

[58] M. Kimura and T. Ohta, Genetics **61**, 763 (1969).

[59] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (The Blackburn Press, Caldwell, NJ, 1970).

[60] G. Sella and A. E. Hirsh, Proc Nat Acad Sci USA **102**, 9541 (2005).

[61] G. Sella, Theor Pop Biol **75**, 30 (2009).

[62] I. M. Rouzine, A. Rodrigo, and J. M. Coffin, Microbiol Mol Biol Rev **65**, 151 (2001).

[63] S. Kullback and R. A. Leibler, Ann Math Stat **22**, 79 (1951).

[64] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, UK, 1983).

[65] T. Ohta, Nature **246**, 96 (1973).

[66] J. B. Lack, C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool, Genetics **199**, 1229 (2015).

[67] J. B. Lack, J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool, Mol. Biol. Evol. **33**, 3308 (2016).

[68] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, T. C. Kaufman, B. R. Calvi, and the FlyBase Consortium, Nucleic Acids Res **47**, D759 (2019).

[69] P. Andolfatto, Nature **437**, 1149 (2005).

[70] D. S. Lawrie, P. W. Messer, R. Hershberg, and D. A. Petrov, PLoS Genet **9**, e1003527 (2013).

[71] J. D. Wall, P. Andolfatto, and M. Przeworski, Genetics **162**, 203 (2002).

[72] N. H. Barton, Phil. Trans. R. Soc. B **365**, 2559 (2010).

[73] J. M. Smith and J. Haigh, Genet. Res. **23**, 23 (1974).

[74] N. L. Kaplan, R. R. Hudson, and C. H. Langley, Genetics **123**, 887 (1989).

[75] J. C. Fay and C.-I. Wu, Genetics **155**, 1405 (2000).

[76] R. R. Hudson and N. L. Kaplan, Genetics **141**, 1605 (1995).

[77] B. Charlesworth, Genetics **190**, 5 (2012).

[78] W. Stephan, Phil. Trans. R. Soc. B **365**, 1245 (2010).

[79] J. C. Fay, Trends Genet. **27**, 343 (2011).

[80] J. M. Comeron, PLoS Genet **10**, e1004434 (2014).

[81] M. Nordborg, B. Charlesworth, and D. Charlesworth, Genet. Res. **67**, 159 (1996).

[82] P. Khromov, C. D. Malliaris, and A. V. Morozov, PLoS ONE **13**, e0190186 (2018).

[83] M. A. Beaumont, W. Zhang, and D. J. Balding, Genetics **162**, 2025 (2002).

[84] K. Csillery, M. G. B. Blum, O. E. Gaggiotti, and O. Francois, Trends Ecol. Evol. **25**, 410 (2010).

[85] M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, PLoS Comput Biol **9**, e1002803 (2013).

[86] M. Kreitman, Annu. Rev. Genomics Hum. Genet. **1**, 539 (2000).

[87] D. L. Hartl, E. N. Moriyama, and S. A. Sawyer, Genetics **138**, 227 (1994).

[88] C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl, Genetics **159**, 1779 (2001).

[89] C. D. Bustamante, R. Nielsen, and D. L. Hartl, Theor. Popul. Biol. **63**, 91 (2003).

[90] S. Sheehan and Y. S. Song, PLoS Comput Biol **12**, e1004845 (2016).

[91] J. P. Chinnici, Genetics **69**, 71 (1971).

[92] J. P. Chinnici, Genetics **69**, 85 (1971).

[93] N. F. Abdullah and B. Charlesworth, Genetics **76**, 447 (1974).

[94] L. D. Brooks and R. W. Marks, Genetics **114**, 525 (1986).

[95] J. M. Comeron, R. Ratnappan, and S. Bailin, PLoS Genet **8**, e1002905 (2012).

[96] A. H. Chan, P. A. Jenkins, and Y. S. Song, PLoS Genet **8**, e1003090 (2012).

[97] D. R. Schrider, D. Houle, M. Lynch, and M. W. Hahn, Genetics **194**, 937 (2013).

[98] P. D. Keightley, R. W. Ness, D. Halligan, and P. R. Haddrill, Genetics **196**, 313 (2014).

[99] G. Watterson, Theor. Popul. Biol. **7**, 256 (1975).

[100] F. Tajima, Genetics **105**, 437 (1983).

[101] D. Hartl and A. Clark, *Principles of Population Genetics*, 4th ed. (Sinauer Associates, 2007).

[102] K. Thornton and P. Andolfatto, Genetics **172**, 1607 (2006).

[103] D. Lachaise, M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner, in *Evolutionary Biology, vol. 22*, edited by M. K. Hecht, B. Wallace, and G. T. Prance (Springer US, Boston, MA, 1988) pp. 159–225.

[104] R. V. Chereji and A. V. Morozov, Brief. Funct. Genomics **14**, 50 (2015).

[105] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, Nat. Rev. Genet. **14**, 390 (2013).