Published in final edited form as:

Theor Popul Biol. 2018 July ; 122: 78–87. doi:10.1016/j.tpb.2018.03.004.

Inference from the stationary distribution of allele frequencies in a family of Wright-Fisher models with two levels of genetic variability

Jake M. Ferguson^a and Erkan Ozge Buzbas^b

^aCenter for Modeling Complex Interactions, University of Idaho

^bDepartment of Statistical Science, University of Idaho

Abstract

The distribution of allele frequencies obtained from diffusion approximations to Wright-Fisher models are useful in developing intuition about the population level effects of evolutionary processes. The statistical properties of the stationary distributions of K -allele models have been extensively studied under neutrality or under selection. Here, we introduce a new family of Wright-Fisher models in which there are two hierarchical levels of genetic variability. The genotypes composed of alleles differing from each other at the selected level have fitness differences with respect to each other and evolve under selection. The genotypes composed of alleles differing from each other only at the neutral level have the same fitness and evolve under neutrality. We show that with an appropriate scaling of the mutation parameter with respect to the number of alleles at each level, the frequencies of alleles at the selected and the neutral level are conditionally independent of each other, conditional on knowing the number of alleles at all levels. This conditional independence allows us to simulate from the the joint stationary distribution of the allele frequencies. We use these simulated frequencies to perform inference on parameters of the model with two levels of genetic variability using Approximate Bayesian Computation.

Keywords

Wright-Fisher model; K -allele models; diffusion approximation; balancing selection; Approximate Bayesian Computation

1. Introduction

Since 1940s, diffusion approximations to Wright-Fisher models have been a fruitful mathematical approach for understanding the effects of evolutionary processes on the long term behavior of idealized populations (Beatty 1986). Paul Joyce had made numerous contributions to theoretical population genetics using these models. He enjoyed the mathematical challenges these models offered, such as finding stationary distributions of allele frequencies or establishing statistical properties of estimators. But more importantly, Paul thought that the probabilistic results obtained in diffusion approximations provided a way to hone our intuition about the effects of evolutionary phenomena in shaping the genetic variability in idealized populations. The novel model presented in this paper and the

methods used to investigate the properties of the parameter estimates are a direct extension of his work on K -allele Wright-Fisher models with balancing selection.

A Wright-Fisher model is a discrete time stochastic population model in which the alleles at a genetic locus evolve under the effect of genetic and demographic processes. Examples of these processes include mutation, selection, recombination, drift, migration, and population structure. Two alleles evolving at a single locus under neutrality in a large population is a textbook example of the Wright-Fisher model (see e.g., Rice 2004, pp. 139–144) for introducing the mathematical basis of population-level evolution. Wright-Fisher models capture the mathematical definition of evolution in an idealized population as the change in allele frequencies over generations. They also allow precise probabilistic formulations of population allele frequency distributions in multi-locus genomes undergoing multiple processes. This flexibility to accommodate a wide range of processes has created a fruitful framework of idealized mathematical models for theoretical population geneticists to understand evolution. The quantities obtained in Wright-Fisher models can often be analytically manipulated to yield insightful results about various evolutionary consequences of genetic and demographic phenomena.

In this paper, we describe a new family of Wright-Fisher models in which there are two hierarchically structured levels of genetic variability. Genotypes composed of alleles defined at the first level of genetic variability have fitness differences and evolve under selection with respect to each other. We refer to an allele of the first level as a class of the second level. At the second level of genetic variability, each allele of the first level is further classified into a set of other alleles. The genotypes composed of alleles within a class have the same fitness and evolve neutrally with respect to each other (Figure 1). Throughout this paper, we work in the framework of K -allele Wright-Fisher models and develop statistical methods to estimate the parameters of the stationary distribution arising from the diffusion approximation to our two-level model. In some sections, we provide a generic treatment of the K -allele Wright-Fisher model but in others we focus on specific models based on symmetric mutation, which means that the rate of mutation from one allele to another is the same for all alleles. We show that conditional on knowing the number of alleles at all levels, the joint stationary distribution of the allele frequencies for the model with two levels of genetic variability factors as the product of the stationary distribution under selection and under neutrality. This factorization is due to an appropriate scaling and partitioning of the overall mutation rate at each of the two levels of genetic variability. This result implies that the allele frequencies for two classes are independent of each other conditional on knowing the number of alleles at all levels. In practice, this conditional independence is methodologically useful because it allows us to efficiently simulate the allele frequencies under neutrality and selection from their marginal distributions respectively. We use these simulated frequencies in the context of Approximate Bayesian Computation (ABC) to perform statistical inference on model parameters of the joint distribution of the model with two levels of genetic variability. Since our model is based on neutral and selected K -allele Wright-Fisher models, we review these models and some of their history in the following section.

2. *K*-allele Wright-Fisher models

2.1. A brief history of *K*-allele Wright-Fisher models

The stationary distribution of allele frequencies at a single locus arising from the diffusion process to Wright-Fisher models can be traced back to the early days of the modern evolutionary synthesis. The two allele models under generic selection were known to R.A. Fisher at least to some degree (Ewens 2004 [pp. 20-24]). However, Fisher did not explicitly write on these models and the exact form of the models he conceived is unclear. Sewall Wright was also familiar with the stationary distributions arising from diffusion approximations for two allele models (Wright 1931). In their classical population genetics reference Kimura and Crow (1970) provided a systematic study of the two allele models under neutrality and various selection schemes. They presented results on allele frequencies at large population and long time evolution limits for two-allele models.

For $K > 2$, the stationary distribution of allele frequencies in a fully parameterized *K*-allele model with selection and mutation has *K* mutation parameters, one for each allele, and $K(K+1)/2$ selection parameters, one for each genotype. Using the allele frequencies from a single locus for purposes of statistical inference on parameters such as the mutation rate or strength of selection leads to an overparameterized model since the number of parameters is greater than the number of alleles. Inference about the parameters of the fully-parameterized model can be performed if there are sufficiently many loci all evolving under identical evolutionary processes with the same parameter values, but this is a strong assumption on the evolution of a multi-locus genome.

For single-locus models, an approach to investigate the statistical properties of the system has been to focus on constrained models which reduce the number of parameters by making assumptions on selection and mutation mechanisms. A generic mutation model has K^2 parameters for *K* alleles. However, we assume a Wright-Fisher diffusion approximation with parent-independent mutation, where mutation from one allele to another is independent of parent's type resulting in *K* mutation parameters for *K* alleles. An example of a constrained model is a model assuming symmetric mutation, which means that all alleles mutate at the same rate. Thus, instead of *K* mutation parameters, one for each allele as in the fully parameterized model, there is only one mutation parameter in a symmetric mutation model. Constrained models capture biologically interesting selection schemes. For example, the full selection matrix with $K(K+1)/2$ parameters is constrained to a single parameter symmetric balancing selection model if we assume that all heterozygotes have equal fitness, all homozygotes have equal fitness, and the fitness of heterozygotes is higher than that of homozygotes.

An early analysis of the stationary distribution of allele frequencies arising from *K*-allele Wright-Fisher models with selection is given by Wright (1949, p.383), who investigated the symmetric balancing selection model. The same model was later studied by Kimura (1955, 1956), but a formal derivation of the stationary distribution was not given in these resources. Watterson (1977) was first to provide a systematic study of the stationary distribution under balancing selection. More recent work on estimating the strength of selection in a balancing selection model has maintained the simplifying assumption of symmetric balancing

selection so that the estimability of parameters was not an issue (Maruyama and Nei 1981, Donnelly, Nordborg, and Joyce 2001). One of our examples in this paper is on estimating the strength of selection in the stationary distribution of the allele frequencies in the two-level model under symmetric balancing selection.

In the mathematical tradition of theoretical population genetics, the works of Wright, Kimura, and Watterson were expanded to investigate the properties of these models deeply. Generalizations and rigorous mathematical accounts using measure theoretic approaches based on Fleming-Viot processes, which include Wright-Fisher models as a subset are available (Ethier and Nagylaki 1989, Ethier and Kurtz 1993, Donnelly and Kurtz 1999, Ethier and Griffiths 1987) and properties of new models are investigated continually (see for example, Spano, Jenkins, and Griffiths this issue). For a concise summary of many benchmark developments and results Ewens (2004) is an excellent accessible reference.

When alleles evolve under selection, the stationary distributions often have computationally intractable likelihoods and studying the statistical properties of parameter estimators is challenging. However, analytical results have been obtained under specific selection patterns of biological interest, such as balancing selection (Joyce, Krone, and Kurtz 2003). A number of recent approaches taking the advantage of novel numerical and computational methods have been used to develop estimators for the strength of selection and properties of these estimators have been investigated. Examples include a perfect sampling approach to simulate samples under non-neutral models (Fearnhead 2006), statistical inference methods to estimate the strength of balancing selection using an importance sampling approach (Donnelly, Nordborg, and Joyce 2001), Bayesian inference on the strength of balancing selection by Markov chain Monte Carlo (Joyce et al. 2012), and spectral methods to perform inference on the strength of selection using time series data (Steinrücken et al. 2014).

2.2. The diffusion approximation to K-allele Wright-Fisher models

For a population of N diploid individuals, we consider K -allelic types denoted by $A_1, A_2, \dots, A_K$ at a locus. At each generation, the population is described by the allele frequencies $\mathbf{x} = (x_1, x_2, \dots, x_K)$, subject to $\sum_{i=1}^K x_i = 1$, where x_i denotes the frequency of allele A_i .

Mutation from one allele type to another is independent of parent's type with rate $u = \sum_{i=1}^K u_i$ to one of the K types available in the population including itself. Under selection, a diploid individual with genotype $A_i A_j$ has a selection coefficient equal to $1 + s_{ij}$ where $s_{ij} > 0$, for all i, j except a reference genotype, say, $A_1 A_1$ which has a selection coefficient $s_{11} = 0$. The neutral model is a special case of the model with selection when $s_{ij} = 0$, for all i, j . The population at generation $t + 1$ is obtained by sampling N alleles independently of each other from the population at generation t , where each allele is sampled as follows: 1) sample a pair $A_i A_j$ with probability proportional to its selection coefficient, 2) choose an allele from the pair $A_i A_j$ with equal probability, 3) With probability u replace the chosen allele with one of the K types in the population, where each type occurs with probability u_i .

The long-term evolution of a population under the process described in the previous paragraph produces a stationary distribution of the allele frequencies. This stationary

distribution is obtained by rescaling the mutation and selection parameters as $\theta = \lim_{N \rightarrow \infty} 4Nu$ and $\sigma_{ij} = \lim_{N \rightarrow \infty} 2Ns_{ij}$ respectively. Using the coefficients $\nu = (\nu_1, \nu_2, \dots, \nu_K)$ with $\sum_{i=1}^K \nu_i = 1$, $\nu_i > 0$ for all i , and the matrix notation $\Sigma = \{\sigma_{ij}\}$ is for the selection parameter, the stationary distribution given by

$$f(\mathbf{x}|\theta, \Sigma, \nu) = C_{\mathbf{x}}^{-1}(\theta, \Sigma, \nu) e^{-\mathbf{x}\Sigma\mathbf{x}^T} \prod_{i=1}^K x_i^{\theta\nu_i-1}, \quad (1)$$

where

$$C_{\mathbf{x}}(\theta, \Sigma, \nu) = \int_{\mathcal{X}_K \times \dots \times \mathcal{X}_2 \times \mathcal{X}_1} e^{-\mathbf{x}\Sigma\mathbf{x}^T} \prod_{i=1}^K x_i^{\theta\nu_i-1} dx_K dx_{K-1} \dots dx_1 \quad (2)$$

is a parameter-dependent normalization constant for $f(\mathbf{x}|\theta, \Sigma, \nu)$ to sum to 1 and $\mathcal{X}_i$ is the space of frequencies for allele i .

Setting $\sigma_{ij} = 0$, for all i, j , in equation 1 recovers the stationary distribution under the neutral K -allele model, and this is a Dirichlet distribution given by

$$f(\mathbf{y}|\theta, \nu) = B_{\mathbf{y}}^{-1}(\theta, \nu) \prod_{i=1}^K y_i^{\theta\nu_i-1}, \quad (3)$$

where for clarity, we denote the frequency of neutral alleles by y_i , and the normalization constant is given by the beta function

$$B_{\mathbf{y}}(\theta, \nu) = [\prod_{i=1}^K \Gamma(\theta\nu_i)] / \Gamma(\sum_{i=1}^K \theta\nu_i),$$

where $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}$ is the gamma function. In the following sections, we will assume

that the mutation is symmetric for the neutral K allele model given in equation 3. This implies that $\nu_i = 1/K$ for all i and all models become conditional on knowing the number of allelic types K . Furthermore, we will use the stationary distribution under selection and under neutrality in the same context but we will allow the number of alleles differ in these models. To emphasize this point, we introduce the parameter vector κ which contains all the number of alleles in the neutral model and the model under selection. We use the notation $C_{\mathbf{x}}(\theta, \Sigma, \nu) = C_{\mathbf{x}}(\theta, \Sigma, \kappa)$ and $B_{\mathbf{y}}(\theta, \nu) = B_{\mathbf{y}}(\theta, \kappa)$ for constants, and the notation $f(\mathbf{x}|\theta, \Sigma, \kappa)$, $f(\mathbf{y}|\theta, \kappa)$ for the stationary distributions under selection and neutrality respectively. Figure 2 plots some examples of stationary probability density functions arising from diffusion approximations to 3-allele Wright-Fisher models under neutrality and selection.

3. A family of Wright-Fisher models with two levels of genetic variability

The stationary distributions in equations 1 and 3 are obtained under the assumption that the K allelic types at a locus evolve either under selection or under neutrality respectively. Typically, alleles at a locus are characterized using a genetic criterion such as: *two alleles are of different types if their DNA differ at least by one non-synonymous mutation from each other*. At highly variable coding regions of the genome, however, the properties of the functional products of genes such as proteins might depend on multiple genetic criteria. In these cases, characterizations of allelic classes at multiple levels are required to capture the effects of evolutionary and demographic processes that maintain the genetic variability at a locus. As a concrete example, we consider the loci that make up the Major Histocompatibility Complex in the human genome, known as the Human Leukocyte Antigen (HLA) (see <http://hla.alleles.org> for a wealth of information). HLA genes code for proteins that help to distinguish the immune system proteins from pathogen proteins, and thus have a major role in the proper functioning of the immune system. There are about 200 HLA genes on chromosome 6 and more than 16,000 alleles have been identified in HLA loci. This staggering level of variability has changed the classification of HLA alleles radically in the last decade. A hierarchical definition of classes of allelic types has become the standard. Currently an HLA allele is uniquely characterized at five levels of variability: 1. its allelic group, 2. a non-synonymous difference in the coding region, 3. a synonymous difference in the coding region, 4. a difference in the non-coding region, and 5. the changes in expression (Marsh et al. 2010). At each level of this system, a different criterion of genetic variability is applied to an allele. Further, alleles belonging to a class that is characterized at a higher level (according to numbering 1-5), inherit the properties of alleles belonging to a class that is characterized at a lower level.

This hierarchical model of specifying the genetic variability and the following characterization of alleles motivates modeling multiple classes of alleles within a Wright-Fisher framework. Investigating a Wright-Fisher model with multiple classes of alleles is useful for at least two purposes. First, it will help us to understand how the allele frequencies at a locus are affected by genetic processes that might only operate on subsets of alleles or genotypes. For the HLA example above, we might want to model selection operating on genotypes that are made of two alleles that differ from each other at level 2 due to non-synonymous differences. On the other hand, genotypes that are made of two alleles that differ only at level 3 or higher will evolve neutrally with respect to each other (for models investigating overdominance in HLA loci see Black and Hedrick 1997, Stoffels and Spencer 2008).

Second, a Wright-Fisher model with multiple classes of alleles will make informative use of allele frequencies in parameter estimation by taking into account the fact that each class jointly informs the estimators of the parameters of relevant processes, thereby decreasing the uncertainty in estimators. Continuing with the HLA example, the allele frequencies characterized at level 2 will carry information on genetic drift, mutation, and selection, but the allele frequencies characterized at level 3 will carry information on genetic drift and mutation only since the genotypes that are made of alleles within a class at level 3 evolve neutrally with respect to each other. In the next section we build a Wright-Fisher model

where there are two levels of genetic variability at a locus. At the first level, there is one class of alleles that evolves under selection. At the second level, for each of the alleles at the first level there is another class of alleles that evolves under neutrality. First, we define the stationary distribution of the joint allele frequencies for this model with two levels of genetic variability. Then we study the statistical properties of the estimators of the strength of selection and mutation parameters for a sample of allele frequencies from this distribution.

3.1. The joint stationary distribution of allele frequencies in the model with two levels of genetic variability

We first modify the standard notation $A_1, A_2, \dots, A_K$ for K -allele models that we introduced in section 2.2 to clearly denote the alleles in the model with two levels of genetic variability. We use A_{ij} to denote the neutral allele $j \in \{1, 2, \dots, k_i\}$ within selected allele $i \in \{1, 2, \dots, K\}$ evolving under selection. We keep A_i to denote the frequency of allele i . For example, genotypes $A_{ij}A_{ij'}$ have the same fitness with respect to each other for all pairs (j, j') because the alleles that make up this genotype belong to the same allele i at the level at which selection operates. On the other hand, the genotypes $A_{ij}A_{i'j'}$ and $A_mA_{m'}$ will potentially have fitness differences if at least one of $i \neq m$ or $i' \neq m'$ holds. We have the parameter vector $\kappa = (K, k_1, k_2, \dots, k_K)$. We denote the frequency of allele A_{ij} by y_{ij} so that the vector of allele frequencies within allele i is $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ik_i})$, subject to $\sum_{j=1}^{k_i} y_{ij} = 1$, for all $i = 1, 2, \dots, K$. For the frequency of A_i we use the notation from the K -allele model with selection: x_i with $\mathbf{x} = (x_1, x_2, \dots, x_K)$, subject to $\sum_{i=1}^K x_i = 1$. Our definitions of alleles and their frequencies may seem strange at first. One can also define the alleles at a single level with $i \times j$ allelic types $A_1, A_2, \dots, A_{i \times j}$ and work with their frequencies such that their frequencies sum to 1. However, using $\mathbf{x}$ and $\mathbf{y}_i$ has the following advantage. The alleles that form genotypes that are neutral with respect to each other have frequencies normalized to 1 within each class at the neutral level, and the allele frequencies at the level in which selection operates are normalized to 1 as well. These definitions allow us to define the two-level model as a function of a Wright-Fisher model under selection, and Wright-Fisher models under neutrality.

Our main technical result provided in Appendix 1 shows that the joint stationary distribution of the allele frequencies in the model with two levels of genetic variability is given by

$$f(\mathbf{x}, \mathbf{y} | \theta, \Sigma, \kappa) = f(\mathbf{x} | \theta_1, \theta_2, \dots, \theta_K, \Sigma, \kappa) \prod_{i=1}^K f(\mathbf{y}_i | \theta_{k_i}, \kappa). \quad (4)$$

Here, each $f(\mathbf{y}_i | \theta_{k_i}, \kappa)$ is the stationary distribution under the neutral k_i -allele model as in equation 3. Further, $f(\mathbf{x} | \theta_1, \theta_2, \dots, \theta_K, \Sigma, \kappa)$ is the stationary distribution under the K -allele model with selection similar to equation 1 in its exponential part (which captures the effect of selection), but with mutation parameters $(\theta_1, \theta_2, \dots, \theta_K)$ that are not necessarily symmetric. In equation 4, θ_i and θ_{k_i} are appropriately scaled mutation parameters. Based on

the assumption of symmetric mutation between alleles A_{ij} , they are given by $\theta_i = k_i(\theta/\tau - 1) + 1$, and $\theta_{k_i} = k_i\left(\frac{\theta}{\tau}\right)$, where $\tau = \sum_{i=1}^K k_i$.

Our conclusion from equation 4 is that the joint stationary distribution factors as the product of the stationary distribution of allele frequencies under the K -allele model with selection and the stationary distribution of allele frequencies under the neutral K -allele model.

A practical implication of this factorization is that to simulate draws from the stationary distribution of the model with two levels of genetic variability, the allele frequencies under the selection model and classes of allele frequencies under neutral models can be simulated independently of each other. Figure 3 shows examples of K -allele models under neutrality, selection, and the model with two levels of genetic variability. Simulating draws from the stationary distribution of the model with two levels of genetic variability allows us to perform inference on the parameters of this model by Approximate Bayesian Computation (ABC).

3.2. Approximate Bayesian Computation (ABC) for inference on the mutation parameter and the strength of selection

The stationary distribution obtained from the diffusion approximation to a Wright-Fisher model produces a model for the random population frequencies. In other words, the population frequencies are a random draw from the stationary distribution in equation 4 which generates a population of joint allele frequencies $\mathbf{x}$ and $(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K)$ at a locus. The sample allele frequencies is a multinomial random sample given the population frequencies and it consists of the sample counts of alleles under the selection model

$\mathbf{n}_x = (n_{x_1}, n_{x_2}, \dots, n_{x_K})$, and the sample counts of alleles under the neutral model

$\mathbf{n}_{y_i} = (n_{y_{i1}}, n_{y_{i2}}, \dots, n_{y_{ik_i}})$, $i = 1, 2, \dots, K$, where n_{x_i} and $n_{y_{ij}}$ are the number of A_i and A_{ij} respectively.

The *likelihood* of the sample allele frequencies is based on a multinomial sample where probability of success parameters for each allele type is the population allele frequency generated by the distribution in equation 4. By Bayes' theorem, the joint posterior distribution of θ and Σ given the sample counts $\mathbf{n}_x, \mathbf{n}_{y_i}$, $i = 1, 2, \dots, K$ is

$$\pi(\theta, \sum | \mathbf{n}_x, \mathbf{n}_{y_1}, \mathbf{n}_{y_2}, \dots, \mathbf{n}_{y_K}, \kappa) \propto P(\mathbf{n}_x, \mathbf{n}_{y_1}, \mathbf{n}_{y_2}, \dots, \mathbf{n}_{y_K} | \theta, \sum, \kappa) \pi(\theta) \pi(\sum),$$

where $P(\mathbf{n}_x, \mathbf{n}_{y_1}, \mathbf{n}_{y_2}, \dots, \mathbf{n}_{y_K} | \theta, \sum, \kappa)$, is the joint likelihood of the observed allele counts given the parameters, and $\pi(\theta)$, $\pi(\Sigma)$ are the prior distributions of the mutation parameter and the strength of selection assuming prior independence of θ and Σ . Performing inference on parameters θ and Σ is challenging. The parameter-dependent normalizing constant $C_x(\theta, \Sigma, \kappa)$, which is computationally intractable, appears in the joint population distribution of the allele frequencies and thus the calculation of the likelihood $P(\mathbf{n}_x, \mathbf{n}_{y_1}, \mathbf{n}_{y_2}, \dots, \mathbf{n}_{y_K} | \theta, \sum, \kappa)$

depends on it. Standard computational methods to sample a distribution such as the rejection algorithm or Markov chain Monte Carlo (MCMC) require evaluating the likelihoods up to a normalizing constant *independent of parameters*, and therefore, they are not directly applicable here. However, the result given in Appendix 1 and described in section 3.1 allows us to simulate random draws of population frequencies from the stationary distribution of the allele frequencies in a two-level model. This is achieved by simulating the allele frequencies from the model under selection and from models under neutrality all independently of each other. Using these frequencies as parameters of the multinomial distribution, we simulate samples $\mathbf{n}_x, \mathbf{n}_{y_1}, \mathbf{n}_{y_2}, \dots, \mathbf{n}_{y_K}$ as random draws from the multinomial model. These simulated samples are key to perform inference by ABC (Beaumont et al. 2002, Tavaré et al. 1997), which allows us to sample the posterior distribution of parameters.

ABC is a simulation-based inference method to produce an approximate sample from the posterior distribution of parameters. ABC utilizes a large number of data sets simulated from the data generating process under different parameter values to approximate the likelihood of the observed data. When different parameter values are drawn from the prior distribution, the likelihood is weighted by the prior and ABC samples approximate draws from the posterior distribution. In practice, ABC is often implemented as follows. First, a reference table of parameter values, generated from their prior distribution, and the data sets (one per parameter set) simulated under those parameter values is built. Second, the dimensionality of all simulated data sets is reduced by summarizing each data set with summary statistics. These statistics are then compared to the same summary statistics calculated from the observed data set. The parameter values that produce summary statistics that are close to the summary statistics calculated from the observed data set are accepted as approximate samples from the posterior distribution. Closeness is made precise with a tolerance parameter and a metric to measure the distance between summary statistics as described in the next paragraph. There are two approximations in ABC. The first approximation is due to dimension reduction from the full data set to summary statistics. Often, the summary statistics are not sufficient for parameters of the model and thus there is information loss by this dimension reduction. The statistical consequence is to replace the likelihood of the data with the likelihood of the summary statistics. The second approximation arises in all but very small sample spaces. For large discrete or continuous data spaces it is not practical to expect to simulate data sets that produce summary statistics which exactly match the summary statistics calculated from the observed data set. Therefore, a distance metric such as Euclidean distance and a small tolerance ε are employed to quantify the approximation in the acceptance step. In the remainder of this section we discuss how to efficiently simulate a large number of data sets under the model with two levels of genetic variability as well as which errors are involved when we use ABC with our model.

Allele frequencies under the neutral K -allele model with parameters $\theta_{v_i}, i = 1, 2, \dots, K$ are simulated from the Dirichlet distribution by standard methods. Simulation of allele frequencies under the K -allele model with selection can be performed by several methods. The efficiency of these methods depends on the complexity of the selection scheme in the sense of distinct number of unique parameters in the matrix of selection parameters and the

magnitude of these parameter for the strength of selection. Often, selection schemes in which the matrix Σ can be parameterized by a single parameter σ are of interest to population geneticists. Examples include the symmetric balancing selection in which all heterozygotes have the same advantage over all homozygotes; the symmetric homozygote advantage in which all homozygotes have the same advantage over all genotypes; or the directional selection where one genotype has higher fitness than all other genotypes. Efficient numerical methods using Fast Fourier Transforms to calculate the normalization constants $C_{\mathbf{x}}(\theta, \Sigma, \kappa)$ exist for these models (see Joyce et al. 2012). Once these constants are calculated on a fine grid of parameter values, random allele frequencies are generated by first evaluating the Cumulative Distribution Function on the grid and then using the numerical inverse as a look up table. This is the numerical inverse CDF method.

If the selection matrix Σ is structurally complex in the sense that the number of distinct selection parameters is large, then the rejection sampling is a good candidate to simulate the allele frequencies under selection. An example is using the Dirichlet proposal distribution, which corresponds to the stationary distribution under neutrality, with the same mutation parameters as the distribution under selection, which is the target distribution. This approach works well when the selection parameters are small but it does not work well when the selection is strong. Strong selection implies large values of parameters in the selection matrix and the population frequencies are far from what would be expected under neutrality. Hence, the frequencies proposed from the neutral model will often be rejected making the rejection algorithm inefficient. In the examples provided in section 4 we use small values for selection parameters and employ rejection sampling. Algorithm 1 shows our procedure to simulate samples from the joint distribution of the allele frequencies under the model with two levels of genetic variability given the parameters $\theta, \Sigma, \kappa = (K, k_i, i = 1, 2, \dots, K)$, $\tau = \sum_{i=1}^K k_i$, and the total sample size n .

Algorithm 1

-
- 1 Simulate a sample under selection:
    - a. Simulate $\mathbf{x}^* \sim \text{Dirichlet}(\theta_1, \theta_2, \dots, \theta_K)$, where $\theta_i = k_i[(\theta/\tau) - 1] + 1$.
    - b. Accept $\mathbf{x}^*$ as from distribution under selection with probability proportional to $e^{-\mathbf{x}^* \Sigma \mathbf{x}^*}$.
    - c. Simulate $\mathbf{n}_{\mathbf{x}}^* \sim \text{Multinomial}(n, \mathbf{x}^*)$, where n is the sample size and from the population and $\mathbf{n}_{\mathbf{x}}^* = (n_{x_1}^*, n_{x_2}^*, \dots, n_{x_K}^*)$.
  - 2 Simulate samples under neutrality conditional on $\mathbf{n}_{\mathbf{x}}^*$:

For $i=1:K$,

    - a. Simulate $\mathbf{y}_i^* \sim \text{Dirichlet}(\theta_{k_1}, \theta_{k_2}, \dots, \theta_{k_i})$, where $\theta_{k_i} = k_i(\theta/\tau)$.
    - b. Simulate $\mathbf{n}_{\mathbf{y}_i}^* \sim \text{Multinomial}(n_{x_i}^*, \mathbf{y}_i^*)$.
-

4. Inference on the mutation parameter and the strength of selection under the model with two levels of genetic variability

In this section, we present a simulation study designed to investigate the performance of estimates of the mutation parameter and the strength of selection by ABC. We consider a model in which the allele frequencies are available from multi-locus genomes where the alleles at each locus evolve under a model with two levels of genetic variability.

Equation 4 holds for any selection scheme and simulating allele frequencies evolving under highly parameterized selection matrices is straightforward in principle. However, these models are rarely useful for developing intuition about broad selection schemes because they are too complex to interpret the individual parameters meaningfully. Further, there are often not enough degrees of freedom in the allele frequencies to estimate all parameters of a full parameterized selection matrix precisely. Thus we focus on three selection schemes: directional selection, directional genic selection, and balancing selection, and vary the following parameters.

1. The mutation parameter and the strength of selection (θ, σ) ,
2. The number of loci in the genome L ,
3. The number of alleles at each level, K and k_i ($i = 1, 2, \dots, K$) alleles at each locus,
4. The sample size from each locus (n_x, n_y) .

For all models we assume symmetric recurrent mutation among alleles.

4.1. The selection models, the simulation parameters, and ABC

In the directional selection model, the matrix Σ has all elements equal to 0 except for $\sigma_{ii} = \sigma > 0$ for a given i . That is, only one homozygote has higher fitness than the rest of the genotypes all of which have equal fitness. The directional genic selection model is the same as directional selection with the addition that $\sigma_{ij} = \sigma_{ii}/2$ for all j . This is the case where the heterozygotes A_iA_j carrying only one A_i allele have half the fitness of the homozygote A_iA_i and genotypes carrying no A_i have 0 fitness.

Perhaps the most interesting selection scheme that we consider is the the symmetric balancing selection. This is the selection scheme that was studied by Wright (1949), Watterson (1977), Donnelly, Nordborg, and Joyce (2001), Buzbas and Joyce (2009). Here, Σ is a diagonal matrix with elements $\sigma_{ii} = -\sigma$, where $\sigma > 0$ on the diagonal and 0 elsewhere. This means that all heterozygotes have the equal fitness and this is higher than the (equal) fitness of all homozygotes. In this model, the joint estimation of the mutation and selection parameters precisely is a statistically challenging task because both mutation and selection promote the genetic variability simultaneously.

The data from simulations were obtained as follows. We first generated $M = 1000$ selection and mutation parameters $(\sigma^{(i)}, \theta^{(i)})$, $i = 1, 2, \dots, M$, where $\theta^{(i)}$ and $\sigma^{(i)}$ were drawn independently of each other from a uniform distribution on $[2, 8]$. The true parameter values

for the mutation rate and the selection strength were also generated independently of each other from the discrete uniform distribution on [2, 8]. For each of these M parameter sets, we simulated the population frequencies ($\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$) and then sampled the population multinomially to get $\mathbf{n}_x^{(i)}, \mathbf{n}_y^{(i)}$. The sample sizes $\mathbf{n}_x^{(i)}, \mathbf{n}_y^{(i)}$ were set to 50 and 100. For the number of loci L , we used 5 values (2, 5, 10, 20, and 40). For each locus we generated a random number of alleles under selection K and the neutral alleles $k_i, i = 1, 2, \dots, K$ independently of each other from a discrete uniform distribution on [2, 6].

We performed the statistical inference using ABC as described in section 3.2 and a posterior sample was obtained from the joint posterior distribution of (θ, σ) . For all of our models, we were fortunate to be able to eliminate the error in ABC due to replacing the data likelihood with the summary statistics likelihood that are not sufficient. This is due to the fact that the stationary distributions of the Wright-Fisher models with selection belong to the well-studied exponential family of distributions for which statistics jointly sufficient for (θ, σ) are well-known. The first class of sufficient statistics applies to alleles at both levels in our model and it is defined as the sum of the logarithm of the allele frequencies. We denote these statistics by $G_x = \sum_i^K \log x_i$ and $G_{y_i} = \sum_j^{k_i} \log y_{ij}$ for all i . The second class of sufficient statistics is related to the exponential function in the selection model. For example, in the directional selection model only x_i^2 is multiplied by a non zero selection parameter σ_{ii} and so x_i^2 is a sufficient statistic. In the directional genic selection, x_i^2 and $x_i x_j$ for all j are sufficient statistics. Finally, under the balancing selection model, the well-known homozygosity given by $\sum_i^K x_i^2$ is a sufficient statistic.

We implemented ABC by first sampling 10^6 values from the prior distributions on θ and σ assuming prior independence. We used uniform distributions for the prior distribution of (θ, σ) , which were also the proposal distributions in our rejection method. Ideally, the support of a uniform prior distribution on parameters should be as large as possible so as not to assign zero probability to any parameter value. However, using large arbitrary bounds is computationally not efficient under selection. This is due to the fact that simulating allele frequencies under strong selection (for large values of σ) requires a computationally infeasible number of proposals under neutrality. We wanted to keep the bounds for the prior distributions as large as possible, but at the same time to be able to simulate the allele frequencies under the desired selection scheme. At the end, we settled on the following intervals. Under balancing selection we used a uniform prior on $\sigma \in [0, 500]$, under directional selection we used $\sigma \in [0, 14]$, and for directional genic selection we used $\sigma \in [0, 20]$. Then we generated one set of population allele frequencies for each parameter pair (θ, σ) and kept the 10^3 parameter sets which produced summary statistics closest to the summary statistics calculated from the observed data set and rejecting the rest. We applied the post-sampling linear regression adjustment (Beaumont et al. 2002) to these sampled points and obtained the final sample from the joint posterior distribution of parameters.

4.2. Results and discussion

We measured the performance of estimators using the mean-squared error (MSE) based on the posterior samples and the test values of the parameters used to generate the population frequencies ($\mathbf{x}$, $\mathbf{y}$) for each simulation. The selection scheme had a substantial effect on the precision of the selection and mutation parameter estimates (Figure 4). Under the balancing selection, we obtained larger MSE for the selection parameter than either under the directed genic selection or the directed selection models. This result has been observed before (Buzbas and Joyce 2009), and intuitively can be explained as follows. In the case of symmetric balancing selection with symmetric mutation, the mode of the stationary distribution for each allele is $1/K$, indicating that both mutation and selection processes work toward decreasing variance in the allele frequencies (see Figure 2 top row, third plot). Thus given a sample of alleles, whether the elevated genetic variability in that sample is due to the effect of the balancing selection or the recurrent mutation is statistically difficult to distinguish. In contrast, in directional or genic selection, the frequency of *specific alleles*—those that contribute to genotypes that have high fitness—increase and not all the allele frequencies. The estimates of the mutation parameter are fairly similar across different selection schemes and their precision is high though the estimates under directional selection scheme had lower MSE than both balancing and genic selection.

As the number of alleles increased, MSE of the estimators decreased (Figure 5). This result shows that increasing the the number of alleles in a model with symmetric mutation increases the information about the parameters. An interpretation of this result for the mutation parameter that relies on properties of the Dirichlet family of distributions is as follows. Observations from a symmetric Dirichlet distribution with a large number of classes is more informative about the parameter of the distribution in comparison to observations from a symmetric Dirichlet distribution with small number of classes. Our results show that the same property holds for the selection parameter as well for the selection schemes that we investigated here.

Differences in MSE of estimators between low (50) and high (100) sample sizes tended to be greater for the selection parameter than the mutation parameter and the MSE of the selection parameter estimator was much larger than 0 under all selection schemes. We note that the maximum number of loci considered in this paper is 40 and not high by today's genomic standards and the sample size is small at 100. In all cases the MSE of the mutation parameter estimator was close to zero even with a fairly moderate number of loci sampled.

In conclusion, for the stationary distribution of allele frequencies arising in Wright-Fisher models with two levels of genetic variability, our findings indicate low MSE in estimates of mutation parameter and high MSE in estimates of the strength of selection. The results from this work extend one of the main findings in Paul Joyce's work on K -allele Wright-Fisher models to a broader class of Wright-Fisher models with two levels of genetic variability. In joint estimation of the strength of balancing selection and the mutation parameter, the strength of selection is the parameter that is more difficult to estimate with high MSE.

Acknowledgments

We thank to an anonymous reviewer for their careful reading of an earlier draft of this manuscript and for catching mistakes in the proof. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix 1

We denote the joint allele frequencies by $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is a vector of allele frequencies under the K -allele model with selection: x_i with $\mathbf{x} = (x_1, x_2, \dots, x_K)$, subject to $\sum_{i=1}^K x_i = 1$, and $\mathbf{y}$ is a list of length K , and each element of $\mathbf{y}$, denoted by $\mathbf{y}_i$ is a vector of length k_i with $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ik_i})$, $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ik_i})$, subject to $\sum_{j=1}^{k_i} y_{ij} = 1$, for all $i = 1, 2, \dots, K$. Our goal is to show that conditional on K and the partitioning of K by k_i defined by the vector parameter $\kappa = (K, k_i, i = 1, 2, \dots, K)$, the joint stationary distribution of the allele frequencies $(\mathbf{x}, \mathbf{y})$ for the model with two levels of genetic variability is equal to the product of the stationary distribution of the allele frequencies $\mathbf{x}$, the level at which selection operates, and the stationary distribution of the allele frequencies $\mathbf{y}$, the level at which the evolution is neutral.

Since by definition $(\mathbf{x}, \mathbf{y})$ satisfy $\sum_{i=1}^K x_i = 1$, $\sum_{j=1}^{k_i} y_{ij} = 1$, the frequency of the allele A_{ij} in the population is the product of the frequency of the selected allele x_i and the frequency of the neutral allele within its selected class y_{ij} and it is equal to $x_i y_{ij}$. We use this fact and the assumption of the symmetric mutation among alleles to obtain our result. We let the total number of y_{ij} alleles by $\tau = \sum_{i=1}^K k_i$. The joint stationary distribution of $(\mathbf{x}, \mathbf{y})$ that arises from diffusion approximation to Wright-Fisher model with full selection matrix Σ and mutation parameter θ is of the form

$$f(\mathbf{x}, \mathbf{y} | \theta, \Sigma, \kappa) = C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) e^{Q(\mathbf{x}, \mathbf{y}, \Sigma)} \prod_{i=1}^K \prod_{j=1}^{k_i} (x_i y_{ij})^{\theta/\tau - 1}. \quad (5)$$

Here, $C_{\mathbf{xy}}(\theta, \Sigma, \kappa)$ is a normalizing constant for $f(\mathbf{x}, \mathbf{y} | \theta, \Sigma, \kappa)$ to sum to 1 and it is given by

$$\int_{\mathcal{X} \times \mathcal{Y}} e^{Q(\mathbf{x}, \mathbf{y}, \Sigma)} \prod_{i=1}^K \prod_{j=1}^{k_i} (x_i y_{ij})^{\theta/\tau - 1} d\mathbf{x} d\mathbf{y}$$

where, we use the integral sign for multiple integrals on $\mathcal{X}$, the allele frequency space of $\mathbf{x}$, and on $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \dots \times \mathcal{Y}_K$ with $\mathcal{Y}_i$ denoting the allele frequency spaces of $\mathbf{y}_i$, $i = 1, 2, \dots, K$. We now focus on the function $Q(\mathbf{x}, \mathbf{y}, \Sigma)$ because our proof relies on showing that this function depends only on the allele frequencies $\mathbf{x}$ and not on $\mathbf{y}$. The full selection matrix is $\Sigma = \{\sigma_{ij}\}$, and for genotype $A_{ij}A_{lm}$, $i = 1, 2, \dots, K$, $m = 1, 2, \dots, K$, σ_{ij} is the selection

parameter. This selection parameter is the same for all j, m in a given selected class denoted by the indices i, l . The function $Q(\mathbf{x}, \mathbf{y}, \Sigma)$ is defined by the stationary distribution of allele frequencies under Wright-Fisher diffusion:

$$Q(\mathbf{x}, \mathbf{y}, \Sigma) = - \sum_{i=1}^K \sum_{l=1}^K \sum_{j=1}^{k_i} \sum_{m=1}^{k_l} \sigma_{il}(x_i y_{ij})(x_l y_{lm}). \quad (6)$$

We write this equation as

$$Q(\mathbf{x}, \mathbf{y}, \Sigma) = - \sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l \sum_{j=1}^{k_i} y_{ij} \sum_{m=1}^{k_l} y_{lm} \quad (7)$$

$$= - \sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l \sum_{j=1}^{k_i} y_{ij} \quad (8)$$

$$= - \sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l. \quad (9)$$

Equation 7 follows from noting that σ_{il}, x_i, x_l do not depend on j, m so they can be taken out of the two innermost sums, and y_{ij} does not depend on m so it can be taken out of the innermost sum. Equation 8 follows because the sum of allele frequencies y_{lm} over index m is equal to 1, and equation 9 follows because the sum of allele frequencies y_{ij} over index j is equal to 1.

Substituting $Q(\mathbf{x}, \mathbf{y}, \Sigma)$ from equation 9 into equation 5 we get the joint distribution of $(\mathbf{x}, \mathbf{y})$ under full selection matrix as

$$f(\mathbf{x}, \mathbf{y} | \theta, \Sigma, \kappa) = C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) e^{-\sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l} \prod_{i=1}^K \prod_{j=1}^{k_i} (x_i y_{ij})^{(\theta/\tau) - 1}. \quad (10)$$

We now distribute the power $(\theta/\tau) - 1$ to x_i and y_{ij} , factoring x_i out of the inner product, and adding and subtracting 1 to the power of x_i we get

$$f(\mathbf{x}, \mathbf{y} | \theta, \Sigma, \kappa) = C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) e^{-\sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l} \prod_{i=1}^K x_i^{k_i[(\theta/\tau) - 1] + 1 - 1} \prod_{j=1}^{k_i} y_{ij}^{(\theta/\tau) - 1}. \quad (11)$$

We now rearrange this density function and write it as the product of the stationary distribution of the allele frequencies $\mathbf{x}$ under selection and of $\mathbf{y}$ under neutrality. The stationary distribution of $\mathbf{x}$ under selection and generic mutation is given by

$$f(\mathbf{x}|\theta, \Sigma, \kappa) = C_{\mathbf{x}}^{-1}(\theta, \Sigma, \kappa) e^{-\sum_{i=1}^K \sum_{l=1}^K \sigma_{il} x_i x_l} \prod_{i=1}^K x_i^{(\theta_i - 1)}, \quad (12)$$

where $\sum_{i=1}^K \theta_i = \theta$. On the other hand, the stationary distribution of the allele frequencies under neutrality for each $\mathbf{y}_i$ with symmetric mutation is given by

$$f(\mathbf{y}_i|\theta, \kappa) = B_{\mathbf{y}_i}^{-1}(\theta, \kappa) \prod_{j=1}^{k_i} y_{ij}^{(\theta/k_i - 1)}. \quad (13)$$

Thus, our goal is to write equation 11 as a function of equation 12 and 13. We note that for $i = 1, 2, \dots, K$ the mutation parameters for $\mathbf{y}_i$ are given by $\theta_{k_i} = k_i(\theta/\tau)$, whereas the mutation parameters for $\mathbf{x}$ are given by $\theta_i = k_i[(\theta/\tau) - 1] + 1$. Using these notations and equations 12 and 13, we write the stationary distribution in equation 11 as

$$f(\mathbf{x}, \mathbf{y}|\theta, \Sigma, \kappa) = C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) C_{\mathbf{x}}(\theta, \Sigma, \kappa) \left[\prod_{i=1}^K B_{\mathbf{y}_i}(\theta_{k_i}, \kappa) \right] f(\mathbf{x}|\theta, \Sigma, \kappa) f(\mathbf{y}_i|\theta_{k_i}, \kappa) \quad (14)$$

where $C_{\mathbf{x}}(\theta, \Sigma, \kappa)$ is the normalizing constant under selection for the distribution of $\mathbf{x}$. Now, $f(\mathbf{x}, \mathbf{y}|\theta, \Sigma, \kappa)$ is a probability density function, and therefore it satisfies

$$\int_{\mathcal{X} \times \mathcal{Y}} f(\mathbf{x}, \mathbf{y}|\theta, \Sigma, \kappa) = 1,$$

so we have

$$1 = \int_{\mathcal{X} \times \mathcal{Y}} C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) C_{\mathbf{x}}(\theta, \Sigma, \kappa) \left[\prod_{i=1}^K B_{\mathbf{y}_i}(\theta_{k_i}, \kappa) \right] f(\mathbf{x}|\theta, \Sigma, \kappa) f(\mathbf{y}_i|\theta_{k_i}, \kappa) \quad (15)$$

$$= C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) C_{\mathbf{x}}(\theta, \Sigma, \kappa) \left[\prod_{i=1}^K B_{\mathbf{y}_i}(\theta_{k_i}, \kappa) \right] \int_{\mathcal{X}} f(\mathbf{x}|\theta, \Sigma, \kappa) \int_{\mathcal{Y}_i} f(\mathbf{y}_i|\theta_{k_i}, \kappa) \quad (16)$$

$$= C_{\mathbf{xy}}^{-1}(\theta, \Sigma, \kappa) C_{\mathbf{x}}(\theta, \Sigma, \kappa) \left[\prod_{i=1}^K B_{\mathbf{y}_i}(\theta_{k_i}, \kappa) \right], \quad (17)$$

where the first line follows by the right hand side of equation 14, the second line follows by rearranging the integrals, and the third line follows by the fact that $f(\mathbf{x}|\theta, \Sigma, \kappa)$ and $f(\mathbf{y}_i|\theta_{k_i}, \kappa)$ are probability density functions and thus each sums to 1. Therefore, we have

$$f(\mathbf{x}, \mathbf{y}|\theta, \Sigma, \kappa) = f(\mathbf{x}|\theta, \Sigma, \kappa) f(\mathbf{y}_i|\theta_{k_i}, \kappa), \quad (18)$$

which shows the desired result that conditional on $\kappa = (K, k_i, i = 1, 2, \dots, K)$, the joint stationary distribution of the allele frequencies for the model with two levels of genetic variability, is equal to the product of the stationary distribution of the allele frequencies $\mathbf{x}$, the level at which selection operates, and the stationary distribution of the allele frequencies $\mathbf{y}$, the level at which the evolution is neutral.

References

- Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002; 162:2025–2035. [PubMed: 12524368]
- Beatty J. The synthesis and the synthetic theory. In: Bechter W, editor *Integrating Scientific Disciplines*. Martinus Nijhoff Publishers; Dordrecht: 1986 125127 *Genetics*
- Black FL, Hedrick PW. Strong balancing selection at hla loci: Evidence from segregation in south amerindian families. *Proc Natl Acad Sci USA*. 1997; 94(23):12452–12456. [PubMed: 9356470]
- Buzbas EO, Joyce P. Maximum likelihood estimates under k -allele models with selection can be numerically unstable. *Annals of Applied Statistics*. 2009; 3:1147–1162.
- Buzbas EO, Joyce P, Abdo Z. Estimation of selection intensity under overdominance by Bayesian methods. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8 Article 32.
- Buzbas EO, Joyce P, Rosenberg NA. Inference on balancing selection for epistatically interacting loci. *Theoretical Population Biology*. 2011; 79:102–113. [PubMed: 21277883]
- Donnelly P, Kurtz T. A countable representation of the Fleming-Viot measure-valued diffusion. *Ann Prob*. 1999; 24:743–760.
- Donnelly P, Nordborg M, Joyce P. Likelihood and simulation methods for a class of nonneutral population genetics models. *Genetics*. 2001; 159:853–867. [PubMed: 11606558]
- Ethier SN, Griffiths RC. The infinitely-many-sites model as a measure-valued diffusion. *Ann Prob*. 1987; 15:515–545.
- Ethier SN, Kurtz TG. Fleming-Viot processes in population genetics. *SIAM Journal of Control Optimization*. 1993; 31(2):345–386.
- Ethier SN, Nagylaki T. Diffusion approximations of the two-locus Wright-Fisher model. *Journal of Mathematical Biology*. 1989; 27:17–28. [PubMed: 2708916]
- Ewens W. *Mathematical Population Genetics: I Theoretical Introduction* Second. Springer; London, UK: 2004
- Fearnhead P. The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theor Popul Biol*. 2006; 70:376–386. [PubMed: 16563450]
- Fisher RA. On the dominance ratio. *Proc Roy Soc Edinburgh*. 1922; 42:321–341.
- Joyce P, Genz A, Buzbas EO. Efficient simulation and likelihood methods for a class of non-neutral multi-allele models. *Journal of Computational Biology*. 2012; 19:650–661. [PubMed: 22697240]

- Joyce P, Krone SM, Kurtz TG. When can one detect over-dominant selection in the infinite-alleles model? *Ann App Prob*. 2003; 13(1):181–212.
- Kimura M. Stochastic Processes and Distribution of Gene Frequencies under Natural Selection. *Cold Spring Harbor Symp Quant Biol*. 1955; 20:33–55. [PubMed: 13433553]
- Kimura M. Rules for testing stability of a selective polymorphism. *Proc Natl Acad Sci US*. 1956; 42:336–340.
- Kimura M, Crow JF. *An Introduction to Population Genetics Theory* Harper and Row; Publishers, NY: 1970
- Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Viã M, Geraghty DE, Holdsworth R, Hurley CK, Lau M, Lee KW, Mach B, Maier M, Mayr WR, Muller CR, Parham P, Petersdorf EW, Sasazuki T, Strominger JL, Svejgaard A, Terasaki PI, Tiercy JM, Trowsdale J. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*. 2010; 75(4): 291–455. [PubMed: 20356336]
- Maruyama T, Nei M. Genetic variability maintained by mutation and over-dominant selection in finite populations. *Genetics*. 1981; 98:441–459. [PubMed: 17249094]
- Rice SH. *Evolutionary theory: mathematical and conceptual foundations* Sinauer; Sunderland, MA, US: 2004
- Steinrück M, Bhaskar A, Song YS. A novel spectral method for inferring general selection from time series genetic data. *Annals of Applied Statistics*. 2014; 8:2203–2222. [PubMed: 25598858]
- Stoffels RJ, Spencer HG. An asymmetric model of overdominance at major histocompatibility complex genes: Degenerate pathogen recognition and intersection advantage. *Genetics*. 2008; 178:1473–1489. [PubMed: 18245836]
- Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. *Genetics*. 1997; 145(2):505–518. [PubMed: 9071603]
- Watterson GA. Heterosis or neutrality? *Genetics*. 1977; 85:789–814. [PubMed: 863245]
- Wright S. *Evolution in Mendelian Populations*. *Genetics*. 1931; 16:97–159. [PubMed: 17246615]
- Wright S. *Adaptation and Selection*. In: Jepsen GL, Simpson GG, Mayr E, editors *Genetics, Paleontology, and Evolution* Princeton Univ. Press; Princeton, NJ: 1949 383

Figure 1.

Alleles (shown within squares) and their population frequencies (notation used throughout the paper) in the hierarchical structure of levels of genetic variability. The level at which selection acts has K allelic types, and the i th allele can further be classified by its k_i neutral variants. In all models, we assume that the number of allelic types represented by parameters $K, k_i, i = 1, 2, \dots, K$ are known. The population frequencies of alleles are normalized and they sum to 1 at the level that selection acts, and also within each neutral variant set of k_i allelic types.

Figure 2.

The stationary probability density functions for 3-allele Wright-Fisher models under neutrality and selection, as a function of the first (x_1) and the second (x_2) allele frequency. $\theta_i = \theta v_i$ denotes the mutation parameter for allele A_i and σ_{ij} is the selection parameter for genotype $A_i A_j$. In all plots, $\sigma_{ij} = 0$ for all (i, j) not shown.

I. K-allele Wright-Fisher model under neutrality			
Stationary Distribution of allele frequencies		Dirichlet($\theta_3 = [3, 3, 3]$)	
Alleles	Frequencies (mode)	Matrix of selection parameters	
		A ₁	A ₂ A ₃
A ₁	x ₁ = 1/3	A ₁	0
A ₂	x ₂ = 1/3	A ₂	0 0
A ₃	x ₃ = 1/3	A ₃	0 0 0

II. K-allele Wright-Fisher model under symmetric balancing selection			
Stationary Distribution of allele frequencies		ExpDirichlet($\theta_3 = [3, 3, 3], \sigma = -5$)	
Alleles	Frequencies (mode)	Matrix of selection parameters	
		A ₁	A ₂ A ₃
A ₁	x ₁ = 1/3	A ₁	−σ
A ₂	x ₂ = 1/3	A ₂	0 −σ
A ₃	x ₃ = 1/3	A ₃	0 0 −σ

III. A Wright-Fisher model with two levels of genetic variability under balancing selection											
Stationary Distribution of allele frequencies					ExpDirichlet($\theta_3 = [3, 3, 3], \sigma = -5$) × Dirichlet($\theta_3 = [3, 3, 3]$) × Dirichlet($\theta_2 = [2, 2]$) × Dirichlet($\theta_2 = [2, 2]$)						
Class	Frequencies (mode)	Alleles	Frequencies (mode)	Matrix of selection parameters							
				A ₁	A ₂	A ₃	B ₁	B ₂	C ₁ C ₂		
A	x ₁ = 1/3	A ₁	y ₁₁ = 1/3	A ₁	−σ	−σ	−σ				
		A ₂	y ₁₂ = 1/3	A ₂	−σ	−σ	−σ				
		A ₃	y ₁₃ = 1/3	A ₃	−σ	−σ	−σ				
B	x ₂ = 1/3	B ₁	y ₂₁ = 1/2	B ₁	0	0	0	−σ	−σ		
		B ₂	y ₂₂ = 1/2	B ₂	0	0	0	−σ	−σ		
C	x ₃ = 1/3	C ₁	y ₃₁ = 1/2	C ₁	0	0	0	0	0	−σ	−σ
		C ₂	y ₃₂ = 1/2	C ₂	0	0	0	0	0	−σ	−σ

Figure 3. Schematic descriptions of single-locus K -allele Wright-Fisher models: I) a neutral model with symmetric mutation. II) a model of symmetric balancing selection where heterozygotes have fitness advantage over homozygotes. III) The model introduced in this paper with two levels of genetic variability. There are three alleles (A,B,C) at the first level which are evolving under symmetric balancing selection (selection matrix given). Each of these alleles is further classified to 3,2, and 2 alleles, evolving under neutrality.

Figure 4.

Comparison of mean-squared error (MSE) of estimators for the strength of selection σ and the mutation parameter θ under each selection scheme (balancing, directional genic, and directional) with respect to the genome size considered (as a function of the number of loci). Each locus has 100 alleles. The points denote the mean of the MSE of estimators, and the bars denote the standard deviation of MSE of estimators. The plots show the number of loci slightly offset, so that the estimates and bars do not overlap.

Figure 5.

Comparison of mean-squared error (MSE) for mutation parameter θ and the strength of selection σ in each type of selection model (balancing, directional genic, and directional) for a given number of loci sampled with a different number of alleles sampled. Points correspond to the mean of the estimator MSE, bars correspond to the standard deviation of the estimator MSE. The plots have the number of loci slightly offset, so that estimates and bars do not overlap. Sample size gives the number of alleles sampled.