Allele frequencies and selection coefficients in locally adapted populations

Journal of Theoretical Biology 565 (2023) 111463

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Allele frequencies and selection coefficients in locally adapted populations

Richard M. Sibly [a,*], Robert N. Curnow [b]

[a] School of Biological Sciences, University of Reading, UK
[b] Department of Mathematics and Statistics, University of Reading, UK

ARTICLE INFO

Keywords:
Migration-selection balance
Dispersal
Population genetics
Ecological genomics
F_{ST}

ABSTRACT

Understanding the role of natural selection in driving evolutionary change requires accurate estimates of the strength of selection acting at the genetic level in the wild. This is challenging to achieve but may be easier in the case of populations in migration-selection balance. When two populations are at equilibrium under migration-selection balance, there exist loci whose alleles are selected different ways in the two populations. Such loci can be identified from genome sequencing by their high values of F_{ST}. This raises the question of what is the strength of selection on locally-adaptive alleles. To answer this question we analyse a 1-locus 2-allele model of a population distributed between two niches. We show by simulation of selected cases that the outputs from finite-population models are essentially the same as those from deterministic infinite-population models. We then derive theory for the infinite-population model showing the dependence of selection coefficients on equilibrium allele frequencies, migration rates, dominance and relative population sizes in the two niches. An Excel spreadsheet is provided for the calculation of selection coefficients and their approximate standard errors from observed values of population parameters. We illustrate our results with a worked example, with graphs showing the dependence of selection coefficients on equilibrium allele frequencies, and graphs showing how F_{ST} depends on the selection coefficients acting on the alleles at a locus. Given the extent of recent progress in ecological genomics, we hope our methods may help those studying migration-selection balance to quantify the advantages conferred by adaptive genes.

1. Introduction

A complete understanding of the role of natural selection in driving evolutionary change requires accurate estimates of the strength of selection acting at the genetic level in the wild. Until recent advances in molecular population genetics, measuring natural selection at the genetic level has been challenging (Linnen and Hoekstra, 2009; Thurman and Barrett, 2016). Thurman and Barrett (2016) located 79 papers that used molecular techniques to study natural selection acting at the genetic level in natural populations, from which the importance of genomic data is clear, though variation in time and space can complicate tracking the strength of selection (Rudman et al., 2018). Several methods for inferring the strength of selection from gene frequency data have been developed (Tataru et al., 2017). Starting with the Wright-Fisher model of the effects of random genetic drift in a randomly mating population of finite size, several approaches have used the diffusion approximation to estimate the effects of various combinations of mutation, migration and selection on how allele frequencies change over the generations. For example, Vitalis et al. (2014) introduced a method extending the diffusion approximation of genetic drift in the migration-drift equilibrium island model to allow for the effects of selection. When applied to analysis of selection on the lactase-producing gene LCT, Vitalis et al. (2014)'s method showed that the strongest selection coefficients occurred in Europe and the Indus Valley, where scaled selection coefficients ranged up to 100.

These studies started with a model of genetic drift and required complex mathematical development. Some simplification can be achieved if the starting point is instead a large population, so that drift can be ignored. The advantages of this approach have been investigated by Jewett et al. (2016), who concluded that at least for time-series data, "ignoring drift leads to estimates of selection coefficients that are nearly as accurate as estimates that account for the true population history, even when population sizes are small and drift is high. This result is of interest because inference methods that ignore drift are widely used in evolutionary studies and can be many orders of magnitude faster than methods that account for population sizes." Hoekstra et al. (2004)

* Corresponding author.
E-mail addresses: r.m.sibly@reading.ac.uk (R.M. Sibly), r.n.curnow@reading.ac.uk (R.N. Curnow).

https://doi.org/10.1016/j.jtbi.2023.111463
Received 19 July 2022; Received in revised form 22 December 2022; Accepted 8 March 2023
Available online 11 March 2023

followed this approach of ignoring drift. Starting from models of populations in migration-selection balance (Haldane, 1930; Wright, 1931), they derived for a simple 2-allele 2-population model an equation that allows estimation of the selection coefficient against a deleterious allele from its equilibrium frequency in each of two populations, the level of dominance and the migration rates between the populations. Using the equation, Hoekstra et al. (2004) estimated the selection coefficient acting on the *Mc1r* gene, which codes for coat colour, in populations of pocket mice living on black lava and on neighbouring light rocks. Here we use a similar approach to study locally adapted populations at equilibrium, held in a balance with selection acting in different directions in different populations. This results in genetic differentiation between the populations if migration rates between the populations are sufficiently low.

Genetic differentiation between populations is generally summarised by F_{ST} (Whitlock, 2011), and Beaumont and Nichols (1996) showed how F_{ST} can be used in genomic studies to identify loci responsible for local adaptations by their relatively high values of F_{ST} (see, e.g., (Flanagan and Jones, 2017; Graham et al., 2018; Savolainen et al., 2013)). Loci not under selection have relatively low values of F_{ST} and information from these loci can provide estimates of migration rates between niches. As an example consider Graham et al. (2018)'s study of high and low altitude populations of the speckled teal (*Anas flavirostris*) in South America, which used genome sequencing and F_{ST} analysis to identify 'outlier' genes selected in opposite ways in the two populations. The outlier genes had F_{ST} values in the range 0.44–0.77, and in some cases it was possible to identify the functions of the outlier genes that are adaptive at high altitude. The remaining non-outlier genes had F_{ST} values around 0.05 which, when further analysed, suggested $< 3\%$ migration rate from low to high altitude, and less from high to low. This genomic study of local adaptation identified loci responsible for local adaptations by their high values of F_{ST}, and genomic variation in unselected regions of the genome to estimate migration rates between niches. High values of F_{ST} occur in genes responsible for local adaptation in large populations at equilibrium, but F_{ST} may also be high in populations not at equilibrium, in which evolution is still under way (Lotterhos and Whitlock, 2014; Lotterhos and Whitlock, 2015). The methods developed in the present paper are for populations at equilibrium.

Intuitively it is clear that F_{ST} is increased by the strength of selection in each population and decreased by migration between them, but a method is needed to quantify the relationship. As a first step, we show here how Hoekstra et al. (2004) 's equation for a deterministic 1-locus 2-allele model of a large population distributed between two niches can be extended to show how selection coefficients can be obtained for both niches. We account for selection acting on migrants as well as residents, which was not addressed by Hoekstra et al. (2004). In section 2 we describe a finite and an infinite population genetic model and compare the resulting evolutionary trajectories towards equilibrium. From this it appears that final evolutionary outcomes do not depend on population size. The rest of the paper investigates the interdependence of the factors affecting evolutionary outcomes in the infinite-population model. In section 3 we show how selection coefficients depend on equilibrium allele frequencies, migration rates, dominance and relative population sizes in two niches. A worked example of the use of the method is provided, together with visualisations of the relationships between selection coefficients and equilibrium allele frequencies. Section 4 considers the implications for F_{ST} and presents visualisations of the dependence of F_{ST} on selection coefficients and migration rates. Given current interest in studying local adaptation using ecological genomics, we hope our methods may help quantify the advantages conferred by adaptive genes.

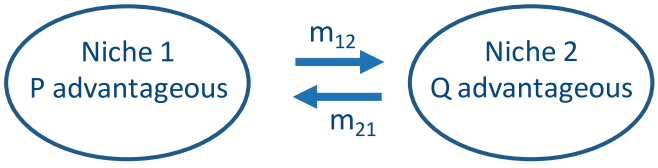

Fig. 1. Conceptual overview of the models. For clarity the niches, of sizes N_1 and N_2, are shown distinct, but in nature may be contiguous or overlap. m_{12} and m_{21} specify the proportion of individuals in one niche that migrate to the other each generation after viability selection and population regulation have occurred. When analysing the models we suppose that Q is disadvantageous in niche 1 (i.e., s_1 is negative) but advantageous in niche 2 (i.e., s_2 is positive), while PP homozygotes have fitness 1 in both niches. There are no sex differences in fitnesses or migration rates.

2. Comparison of evolutionary trajectories of finite and infinite population genetic models

2.1. Population genetic models

In this section we begin by describing the finite and infinite population genetic models on which the paper is based. In both, a single locus with two alleles P and Q is modelled in an environment consisting of two niches with some migration between niches prior to mating, as depicted in Fig. 1. The locus determines ecological adaptation to one niche or the other. The fitness of the three genotypes in each niche is shown in Table 1.

Generations are discrete and individuals die after mating. Here we analyse two types of Wright-Fisher model: a model with finite population sizes coded in SLiM (Haller and Messer, 2019), and a deterministic infinite-population model described below. All models include mutation. In the SLiM model parents mate randomly in each patch each generation and are chosen with probability proportional to their fitness. N_1 and N_2 individuals are created in this way and some of them migrate to form part of the next generation's population in the other patch as shown in Fig. 1. The SLiM code for the SLiM model is given in Supplementary Materials. The deterministic infinite-population model is described in subsection 2.2.

2.2. Recurrence equations giving the frequencies of the P allele in successive generations in the deterministic infinite-population model

Life histories in this model occur in the following order. At the start of each generation individuals in each niche mate at random, and all mating individuals obtain the same number of offspring. The number of offspring of each genotype that survive in each niche is the product of its initial frequency and its fitness. Population regulation then returns the population number in each niche to its initial value, N_1 in niche 1 and N_2 in niche 2 (only the ratio N_1:N_2 is relevant in the infinite population model, but for clarity of exposition we here retain notation distinguishing population sizes in the two patches). Finally some individuals migrate between niches as shown in Fig. 1, which leads to the start of the next generation. The recurrence equations derived below give the frequencies of the P allele in successive generations in each niche. The relative frequencies of P and Q at the start of a generation in the two niches are given in Table 1. These are first modified by mutation so that:

$$p_1' = p_1(1-\mu) + q_1\mu \tag{1a}$$

$$p_2' = p_2(1-\mu) + q_2\mu \tag{1b}$$

where μ is the mutation rate per genome per generation. The relative frequencies of P in the two niches after selection and population regulation are then:

$$p_1'' = (p_1'^2 + (1+hs_1)p_1'q_1')/(1 + 2hp_1'q_1's_1 + s_1q_1'^2) \tag{2a}$$

$$p_2'' = (p_2'^2 + (1+hs_2)p_2'q_2')/(1 + 2hp_2'q_2's_2 + s_2q_2'^2) \tag{2b}$$

The relative frequencies of P after migration are:

Table 1
The fitness of the three genotypes in each niche. p_1 and q_1 represent the relative frequencies of the P and Q alleles at the start of a generation in niche 1, their frequencies in niche 2 are p_2 and q_2. Carriers of the QQ genotype obtain fitnesses $1 + s_1$ and $1 + s_2$ in niches 1 and 2 respectively, PP homozygotes have fitness 1 in both niches. Parameter h indicates the level of dominance of the Q allele.

	Niche 1, size N_1			Niche 2, size N_2		
Genotype	PP	PQ	QQ	PP	PQ	QQ
Fitness	1	$1 + hs_1$	$1 + s_1$	1	$1 + hs_2$	$1 + s_2$
Relative frequency	p_1^2	$2p_1q_1$	q_1^2	p_2^2	$2p_2q_2$	q_2^2

$$p_1'' = ((1 - m_{12})N_1p_1'' + m_{21}N_2p_2'')/c_1 \quad (3a)$$

$$p_2'' = ((1 - m_{21})N_2p_2'' + m_{12}N_1p_1'')/c_2 \quad (3b)$$

where p_1'' and p_2'' are obtained from Eq. (2), $c_1 = (1-m_{12})N_1 + m_{21}N_2$ and $c_2 = (1 - m_{21})N_2 + m_{12}N_1$.

The equations show how the frequencies of P in the two niches in the next generation, p_1''' and p_2''', can be derived from the frequencies in the present generation, p_1 and p_2.

2.3. Dependence of evolutionary outcomes on population size

Evolutionary trajectories of the finite and infinite population genetic models described in subsections 2.1 and 2.2 can now be compared. Sample outputs from the SLiM finite population model and the deterministic infinite-population model are shown in Fig. 2 for cases in which $N_1 = N_2 = N$. These outputs are from single simulations but the final equilibrium values were found to be very similar in repeated simulations. After chance variations in the times of occurrence of the first successful Q mutations in finite population simulations, evolutionary trajectories from initial mutation to final outcome were similar between models with different population sizes and were completed within a few hundred generations. Equilibrium frequency values of P alleles (blue lines) are as expected always higher in patch 1 than in patch 2. In smaller populations there is more variation between generations, as a result of genetic drift, but overall there is good agreement between the models in their final frequencies. In the SLiM model times until first successful mutation were sometimes appreciably longer than shown: the lengthy time to the equilibrium in the lower panel example for $N_1 = N_2 = 1000$ was due to the loss of earlier mutations through drift. Fixation sometimes occurred at $N = 100$ because of large fluctuations in allele frequencies due to drift: a mutation producing a new Q allele was then needed before populations could proceed to equilibrium.

To evaluate the effect of population size on final frequencies, final SLiM frequencies were plotted against infinite-population final frequencies as shown in Fig. 3. Note that the correspondence is very good when population size is 10, 000 (black symbols). The aberrant red point resulted from a simulation at $N = 100$ in which P was fixed after 100, 000 generations. The effect of increasing h from 0.5 to 1 is as expected to increase the frequency of P in both patches, as can also be seen in Fig. 2.

In sum, for the parameters investigated final evolutionary outcomes do not depend on population size. The rest of the paper reveals the interdependence of the factors affecting evolutionary outcomes in the deterministic infinite-population model.

3. Relationship between allele frequencies and selection coefficients

In this section we show how in the deterministic infinite-population model, selection coefficients can be derived from equilibrium allele frequencies, migration rates, dominance and the relative population sizes in the two niches. Formulae are given with which to calculate standard errors, and a worked example of the use of the method is provided, together with visualisations of the relationships between selection coefficients and equilibrium allele frequencies. The inverse relationship between allele frequencies and selection coefficients is also considered.

3.1. Calculation of selection coefficients from equilibrium values of p_1 and p_2, migration rates and population sizes, ignoring mutation

Mutation rates are very small in comparison with the selection coefficients of interest here, and equilibrium frequencies are within 0.001 for the infinite-population model with and without mutation in Fig. 3.

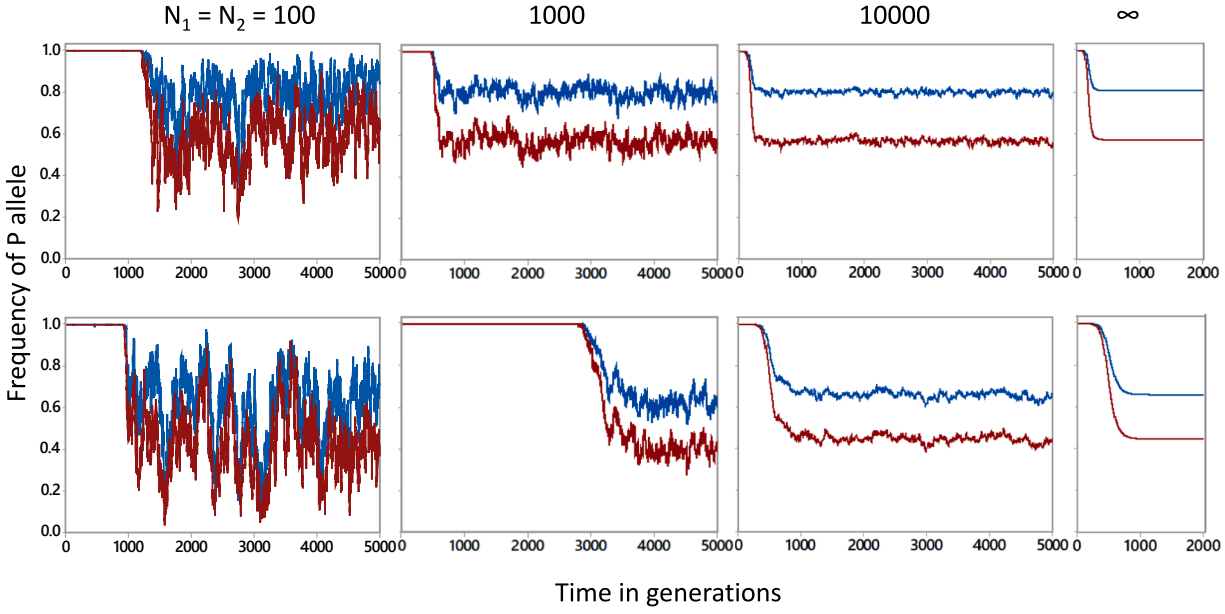

Fig. 2. Frequency of the P allele in two patches over multiple generations in relation to population size. Population sizes in each patch are shown at the top of each column. Top row: h = 1, bottom row: h = 0.5. Blue lines show frequency in patch 1, red lines in patch 2. $s_1 = -0.1$, $s_2 = 0.1$; m = 0.05; mutation rate = 10^{-6} per generation except 10^{-5} for N = 100: higher mutation rates were used for N = 100 so that the first successful mutation occurred in a reasonable number of generations. Initial frequency of P was 1.0 in both patches.

On this basis we replace p_1' by p_1 and p_2' by p_2 in the recurrence equations derived in subsection 2.2, and this allows us to analyse what happens when populations reach equilibrium. At equilibrium in large populations, $p_1''' = p_1$ and $p_2''' = p_2$, so from equations 3:

$$p_1'' = \frac{c_1p_1 - m_{21}N_2p_2''}{(1-m_{12})N_1} = \frac{c_2p_2 - (1-m_{21})N_2p_2''}{m_{12}N_1} \tag{4}$$

If m_{12}, m_{21}, N_1, N_2 and equilibrium values of p_1 and p_2 are known, then Eq. (4) can be rearranged to obtain:

$$p_2'' = \frac{m_{12}(c_1p_1 + c_2p_2) - c_2p_2}{(m_{12}+m_{21}-1)N_2}, = K_2, \text{ say} \tag{5a}$$

Similarly

$$p_1'' = \frac{m_{21}(c_1p_1 + c_2p_2) - c_1p_1}{(m_{12}+m_{21}-1)N_1}, = K_1, \text{ say.} \tag{5b}$$

Equating p_2'' in Eqs. (5a) and (2b), and remembering p_1' has been replaced by p_1, we obtain:

$$s_2 = \frac{p_2 - K_2}{K_2q_2^2 + hp_2q_2(2K_2-1)} \tag{6a}$$

And similarly

$$s_1 = \frac{p_1 - K_1}{K_1q_1^2 + hp_1q_1(2K_1-1)} \tag{6b}$$

where K_1 and K_2 are given by Eq. (5).

Approximate standard errors for the estimates of s_1 and s_2 can be obtained by Taylor series expansion about the infinite-population values of p_1 and p_2. These standard errors quantify the uncertainty in estimates of s_1 and s_2 that result from uncertainty in estimates of allele frequencies, migration rates and population sizes. Assuming the covariances between $p_1, p_2, m_{12}, m_{21}, N_1$ and N_2 are small, the variance of $s_2(p_1, p_2, m_{12}, m_{21}, N_1, N_2)$, $V(s_2)$, can be written as a Taylor series. If third order terms and above can be ignored, then:

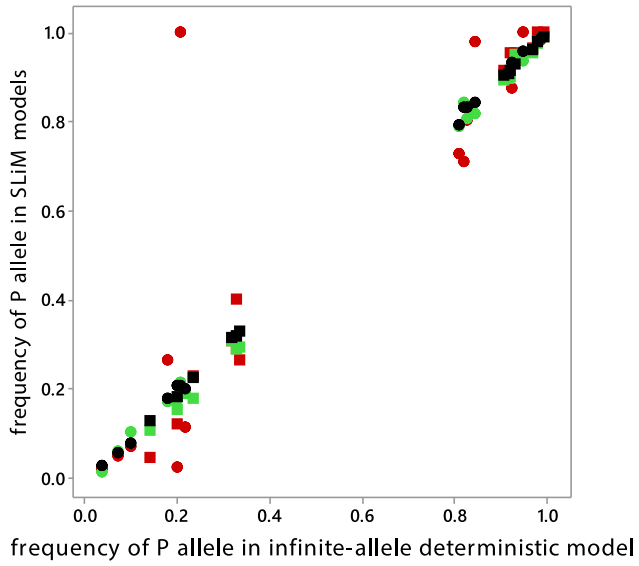

Fig. 3. Correspondence between outputs of SLiM and infinite-population models. Black, green and red symbols correspond to N = 10,000, 1,000 and 100 respectively. Dominance indicated by squares if h = 1; circles if h = 0.5. Outputs recorded after 10,000 generations except 100,000 for N = 100 (more generations were allowed for N = 100 to let successful mutations occur). m = 0.01, mutation rate 10^{-6}. Values of (s_1, s_2) were either $s_1 = -0.1$ and $s_2 = 0.1$, 0.2, 0.3 or 0.9; or $s_2 = 0.1$ and $s_1 = -0.2$, -0.3 or -0.9. Left-hand 21 points are outcomes in patch 2, right-hand 21 points in patch 1.

$$V(s_2) \approx \left(\frac{\partial s_2}{\partial p_1}\right)^2 V(p_1) + \left(\frac{\partial s_2}{\partial p_2}\right)^2 V(p_2) + \left(\frac{\partial s_2}{\partial m_{12}}\right)^2 V(m_{12}) + \left(\frac{\partial s_2}{\partial m_{21}}\right)^2 V(m_{21}) + \left(\frac{\partial s_2}{\partial N_1}\right)^2 V(N_1) + \left(\frac{\partial s_2}{\partial N_2}\right)^2 V(N_2) \tag{7}$$

This gives the square of the standard error of s_2, $V(s_2)$, in terms of the squares of the standard errors of $p_1, p_2, m_{12}, m_{21}, N_1$ and N_2; $V(p_1)$, $V(p_2)$, $V(m_{12})$, $V(m_{21})$, $V(N_1)$, $V(N_2)$ respectively. The last four of these will need to be estimated experimentally. $V(s_1)$ is obtained similarly.

In finite populations p_1 varies about its infinite-population value as a result of genetic drift, with variance $V_{drift}(p_1)$, say, here called drift variance, and is also independently subject to sampling variance. Using the finite sampling correction without replacement, the formula for the sampling variation is $p_1(1-p_1)(N_1 - N_{gen})/(2N_{gen}(N_1 - 1))$, where N_{gen} is the number of sampled genomes. Division by 2 is necessary because there are twice as many alleles as genomes if sampling takes place after random mating but before selection. $V(p_1)$ is obtained as the sum of the drift and sampling variances of p_1. $V(p_2)$ is obtained similarly. A worked example of calculation of selection coefficients and their approximate standard errors using an Excel spreadsheet is provided in Supplementary Materials.

3.2. Visualising the relationship between selection coefficients and equilibrium allele frequencies

Until now theory has been illustrated with frequencies of the P allele – which may be thought of as wild type – but in this subsection it is convenient to focus instead on the Q allele, remembering that $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. The selection coefficients corresponding to particular equilibrium allele frequencies can be calculated using equations 6 and these are illustrated in Fig. 4. To attain a stable equilibrium with $s_2 > 0$ and $s_1 < 0$ it is necessary that there be more Q alleles in niche 2 than in niche 1, i.e., $q_2 > q_1$, and this is why points only occur in the corresponding triangle of q_1 q_2 space in Fig. 4. The upper panels of Fig. 4 show that to maintain equilibrium, for a given value of q_1, selection for Q in niche 2, s_2, has to increase increasingly with q_2. Thus relatively strong selection for Q is needed in niche 2 if Q is to be maintained there at high levels, as shown by the rightward upturn in the surfaces in the upper panels. The lower panels show the converse situation in niche 1 (note reversal of q_1 and q_2 axes). To maintain equilibrium, selection against Q in niche 1 has to increase as q_1 decreases. Relatively strong selection against Q is needed in niche 1 if it is to be maintained there at low levels, as shown by the rightward downturn in the surfaces in the lower panels as q_1 tends to zero. The effects of level of dominance, here shown at the extremes of h = 0.5 and h = 1, appear relatively minor, though there are differences close to fixation in niche 1. There is a leftward downturn as q_1 tends to 1 in the lower lefthand panel but not in the lower righthand panel. This is because to maintain equilibrium close to fixation, stronger selection against Q is needed in niche 1 if h = 0.5 than if h = 1.

3.3. Calculating equilibrium frequencies of alleles in each population when selection coefficients are known

We now turn to the inverse problem of calculating equilibrium frequencies of alleles in each population when selection coefficients are known. From Eq. (4):

$$p_1 = \frac{m_{12}+m_{21}-1}{c_1m_{12}}N_2p_2'' + \frac{1-m_{12}}{c_1m_{12}}c_2p_2 \tag{8}$$

Substituting for p_2'' from Eq. (2b), and remembering we are setting $p_1 = p_1'$ and $p_2 = p_2'$ as explained above, gives an equation for p_1 in terms of p_2, m_{12}, m_{21}, h, s_1 and s_2,:

$$p_1 = \frac{m_{12}+m_{21}-1}{c_1m_{12}}N_2\left(\frac{p_2(1+hs_2q_2)}{1+2hs_2p_2q_{2+}+s_2q_2^2}\right) + \frac{1-m_{12}}{c_1m_{12}}c_2p_2 \tag{9a}$$

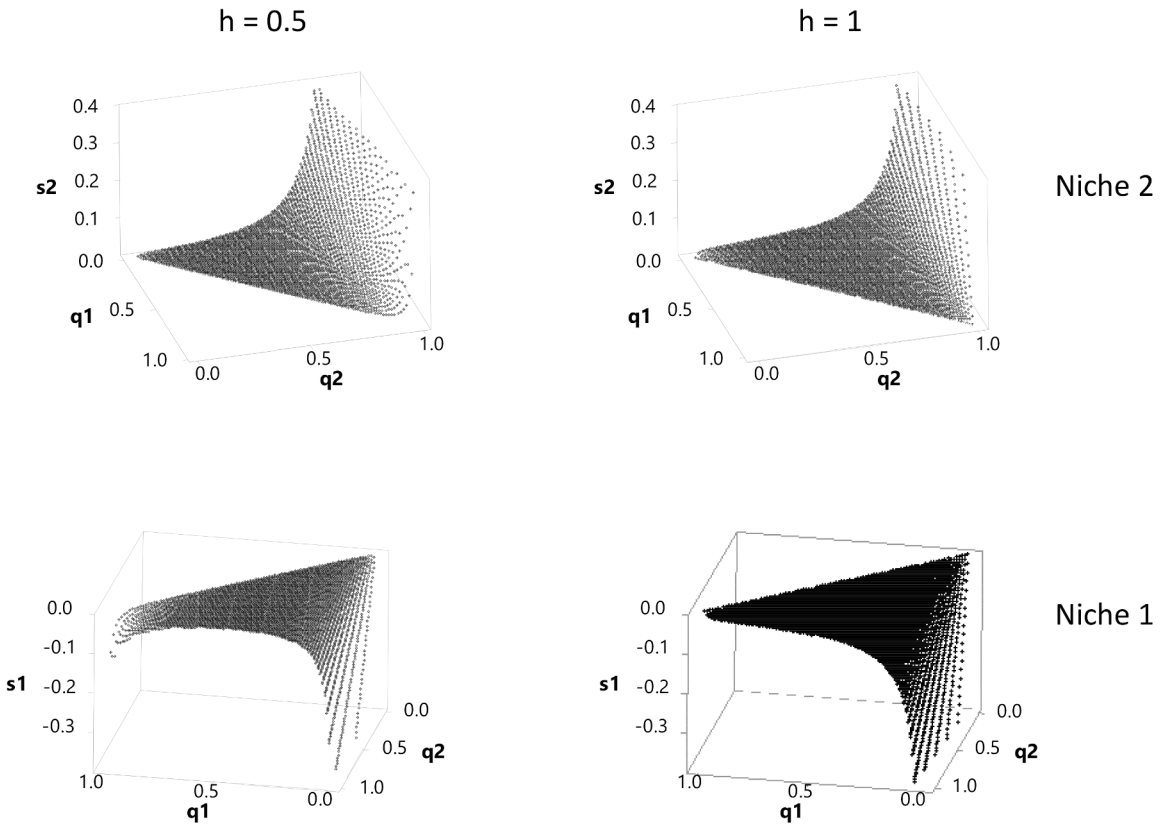

Fig. 4. Three-dimensional plots showing selection coefficients s_2 and s_1 needed to maintain particular equilibrium allele frequencies q_1 and q_2. Upper panels show s_2 in niche 2, lower panels show s_1 in niche 1. Left hand panels are for level of dominance h = 0.5, right hand panels h = 1. Selection coefficients were calculated over a grid of q_1 q_2 values, using equations 6 for the case $m_{12} = m_{21} = 0.01$, $N_1 = N_2$.

The analogous equation for p_2 is:

$$p_2 = \frac{m_{12} + m_{21} - 1}{c_2 m_{21}} N_1 \left(\frac{p_1(1 + hs_1q_1)}{1 + 2hs_1p_1q_{1+} + s_1{q_1}^2}\right) + \frac{1 - m_{21}}{c_2 m_{21}} c_1 p_1 \qquad (9b)$$

Substituting p_2 from Eq. (9b) into Eq. (9a) yields an equation in p_1, m_{12}, m_{21}, N_1, N_2, h, s_1 and s_2 which can be solved to obtain equilibrium values of p_1 for given values of m_{12}, m_{21}, N_1, N_2, h, s_1 and s_2. Equilibrium values of p_2 can be obtained similarly.

Equations (9) do not provide explicit expressions for the equilibrium values of p_1 and p_2. An alternative to using an equation solver to obtain equilibrium values is to simulate the evolutionary process for specified values of h, m_{12}, m_{21}, N_1, N_2, s_1 and s_2 using the recurrence equations (2) and (3). Checks showed that the simulated evolutionary outcomes satisfied equations (9).

Equations (9) show how equilibrium frequencies of alleles in each population can be calculated when selection coefficients are known.

4. Implications for F_{ST} and visualisations of the dependence of F_{ST} on selection coefficients and migration rate

The interdependence of the factors affecting evolutionary outcomes has been shown in equations 6 and 9. In this section we explore the implications for F_{ST}; how equilibrium values of F_{ST} are related to the migration rates between the niches, and the strengths of selection s_1 and s_2 within them.

4.1. Calculation of F_{ST}

F_{ST} was calculated from allele frequencies using the equation $F_{ST} = {\sigma_S}^2/{\sigma_T}^2$, where ${\sigma_S}^2$ and ${\sigma_T}^2$ represent the variances of an allele's frequency between subpopulations, and in the total population, respectively (Holsinger and Weir, 2009). The variance of p, the frequency of an allele in a population of size n, is given by the binomial distribution as npq. So $F_{ST} = (p_T q_T - \frac{1}{2}p_1 q_1 - \frac{1}{2}p_2 q_2)/p_T q_T$, where subscripts T, 1 and 2 refer to the total population and the populations in niches 1 and 2 respectively. F_{ST} can therefore alternatively be thought of in terms of the average frequency of heterozygotes in the two populations compared with the frequency of heterozygotes if there was random mating between all the individuals in the two populations. F_{ST} was assigned the value 0 when one allele became fixed in both niches because F_{ST} tends to 0 as any one allele tends to fixation.

4.2. Visualisation of the function $F_{ST} = F_{ST}(m, s_1, s_2)$ when $m_{12} = m_{21} = m$

In this subsection we show how equilibrium values of F_{ST} are related to the migration rates between the niches, and the strengths of selection s_1 and s_2 within them. For simplicity we assume that migration rates are the same in each direction, so that $m_{12} = m_{21} = m$, and that population sizes in the two niches are equal so that $N_1 = N_2$. We begin by showing how genetic differentiation, measured by F_{ST}, is reduced by migration when selection is maximal, total selection one way in niche 1 and the other way in niche 2 (Fig. 5). Maximal selection produces maximum values of F_{ST}, and these decline as migration rates increase, from a maximum of 1 when populations are isolated, to around 0.4 when migration rates are 10% and down to zero when migration rate is 50% (Fig. 5). The maximum values of F_{ST} are a little higher for h = 0.5 than for h = 1 (dominance) when migration rates are less than 15%, but otherwise similar.

To visualise the effects of varying values of s_1 and s_2 on F_{ST} we calculated equilibrium values of F_{ST} over a grid of values of m, s_1 and s_2 to obtain visualisations of the function $F_{ST} = F_{ST}(m, s_1, s_2)$, and these are

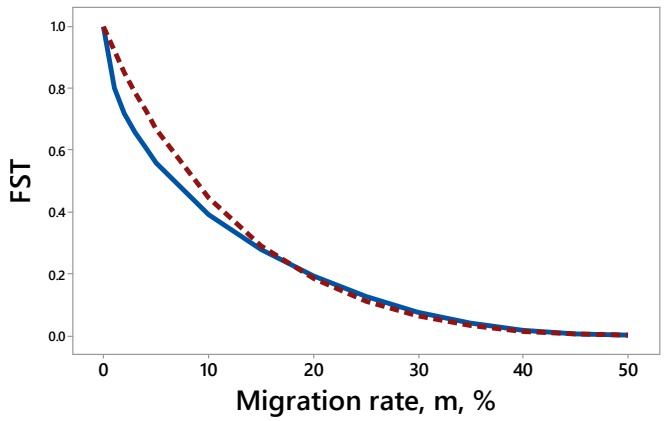

Fig. 5. The maximum values of F_{ST} in relation to migration rate, m. Maximum values of F_{ST} occur when selection on QQ is total one way in niche 1 and the other way in niche 2, and this was approximated by setting s_1 at -1 and s_2 at 10^9. Solid blue curve is for h = 1, dashed red curve h = 0.5. Values of F_{ST} were calculated for populations at equilibrium as determined by simulation of evolutionary trajectories of allele frequencies. Equilibrium judged by eye was generally achieved within 100 generations.

presented in Fig. 6 for two levels of dominance h. The panels in the rows of Fig. 6 represent visualisations of the function $F_{ST} = F_{ST}(m, s_1, s_2)$ for three values of migration rate m. The highest values of F_{ST} are achieved when migration rates are lowest (left-hand panels of Fig. 6). Within each panel the highest value of F_{ST} occurs when selection is at its strongest, for Q in niche 2 ($s_2 = 1$) and against Q in niche 1 ($s_1 = -1$). If selection is too low, then either P or Q go to fixation and F_{ST} goes to zero. At other values of s_1 and s_2 the equilibrium is a polymorphism and $F_{ST} > 0$. The shapes of the F_{ST} surfaces for h = 1 (top row of Fig. 6) and h = 0.5 (bottom row) are qualitatively similar but they differ a little in quantitative detail. The F_{ST} surfaces shown in Fig. 6 are not perfectly symmetrical about the $s_2 = -s_1$ plane, because $s_1 = -1$ represents total selection against QQ in niche 1 but $s_2 = +1$ does not represent total selection for QQ in niche 2; that is achieved when $s_2 = \infty$.

5. Discussion

Methods to estimate selection coefficients are needed to further understand evolutionary processes in wild populations. Here, using a simple two-niche two-allele model, we show in equations 6 how selection coefficients s_1 and s_2 can be estimated from measurements of migration rates, population sizes, dominance and the equilibrium values of allele frequencies in two habitats. In a worked example it is shown that the standard errors of the selection coefficients can also be estimated from measurements of the other population parameters. These estimates rely, as in all Wright-Fisher analyses, on the simple model life history analysed being an adequate representation of reality. The method of estimating selection coefficients using equation 6 requires knowledge of migration rates, allele frequencies and population sizes in the two habitats, and the level of dominance. Migration rates can sometimes be estimated by marking individuals or using genetic markers (e.g., (Sunde et al., 2020)), and allele frequencies are routinely measured in genomic studies. Effects of levels of dominance are discussed below.

Because F_{ST} is generally reported rather than allele frequencies, we analysed the relationship between F_{ST} and selection coefficients, and presented the results graphically in Figs. 5 and 6 as visualisations of the function $F_{ST} = F_{ST}(m, s_1, s_2)$, where both migration rates m between the two niches and the populations sizes within them are equal. These visualisations show how equilibrium values of F_{ST} are related to the migration rates between the niches, and the strengths of selection s_1 and s_2 within them. The graphs provide quantitative detail as to how F_{ST} declines as migration rates increase (Figs. 5 and 6). If F_{ST} and migration rates are known, Fig. 6 shows that some inferences are possible as to the values of selection coefficients.

All the results presented here show some dependence on levels of dominance. Although there is interest in the evolution of dominance, surprisingly little is known of values of levels of dominance in natural populations (Billiard et al., 2021; Huber et al., 2018; Thurman and Barrett, 2016) except that overdominance is infrequent (Thurman and Barrett, 2016) but see also Brookfield (2020)). Here in Figs. 2–6 we present evolutionary outcomes for what are, in the absence of overdominance, the extreme values 0.5 and 1. Fig. 4 suggests that to

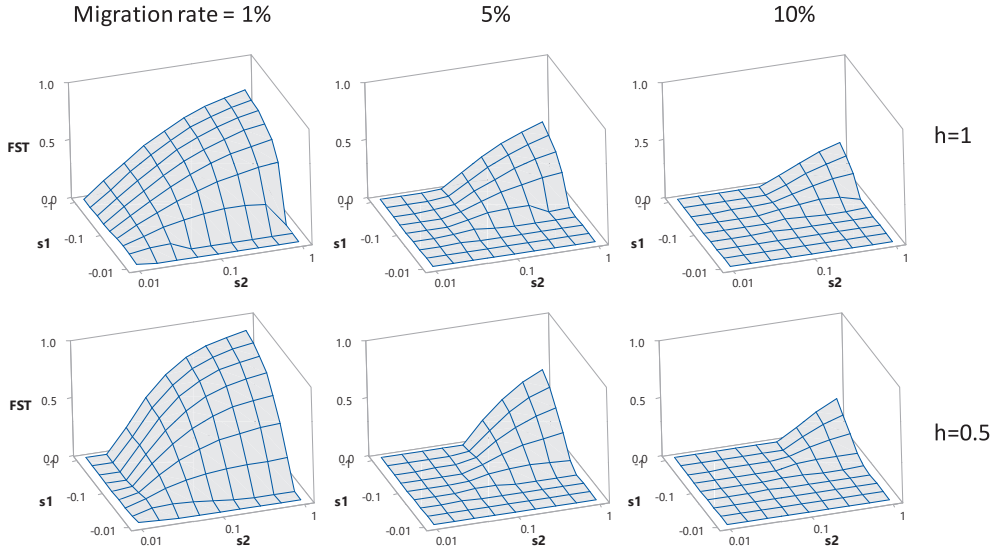

Fig. 6. Values of F_{ST} in relation to selection coefficients s_1 and s_2 for three rates of migration between niches m, and two levels of dominance h. Top row: Q is dominant (i.e. h = 1); bottom row h = 0.5. Values of F_{ST} were calculated for populations at equilibrium as determined by simulation of the evolutionary process.

maintain equilibrium close to fixation, stronger selection against Q is needed in niche 1 if h = 0.5 than if h = 1, but otherwise it seems that in the absence of overdominance, levels of dominance have only small effects on equilibrium values.

Inferences in practice will need to take account of several caveats. Our calculations are for populations at equilibrium, but in the real world selection pressures, migration rates and population sizes may vary over time, so that allele frequencies, F_{ST} and other variables vary too. Fig. 2 gives an indication of how populations approach equilibrium in the presence of drift, but further analysis would be valuable, perhaps using approaches such as those reviewed by Tataru et al. (2017). The effects of drift on standard errors are shown in equation 7. Loci may become differentiated between populations not because they are themselves selected but because they are physically close to loci that are selected (Petry, 1983), the phenomenon of linked selection (see, e.g., (Aeschbacher et al., 2017; Burri, 2017)).

Here we have derived equations, for populations in migration-selection balance, that show the relationships between equilibrium allele frequencies, selection coefficients, migration rates, population sizes and dominance. These are illustrated by graphs that show quantitatively how, at a given locus, equilibrium allele frequencies are related to the selection coefficients that hold them in equilibrium, and how F_{ST} increases with the selection coefficients acting on the alleles at the locus. We hope our methods may help those studying migration-selection balance to quantify the advantages conferred by adaptive genes.

CRediT authorship contribution statement

Richard M. Sibly: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Robert N. Curnow:** Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Data archiving

There are no data in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Mark Beaumont, Renaud Vitalis, Simon Aeschbacher, Ilik Saccheri and two referees for comments and suggestions on an earlier version of the MS, and to Ben Haller for writing some of the SLiM code.

Appendix A. Supplementary data

Supplementary data (SLiM code for the finite population model described in the text. Worked example of calculation of selection coefficients and their approximate standard errors using an Excel spreadsheet. Excel spreadsheet showing how selection coefficients and their approximate standard errors may be calculated from observed values and standard errors of allele frequencies, migration rates, population sizes and level of dominance, on the assumption that populations are in equilibrium.) to this article can be found online at https://doi.org/10.1016/j.jtbi.2023.111463.

References

Aeschbacher, S., Selby, J.P., Willis, J.H., Coop, G., 2017. Population-genomic inference of the strength and timing of selection against gene flow. Proc. Natl. Acad. Sci. U. S. A. 114, 7061–7066. https://doi.org/10.1073/pnas.1616755114.

Beaumont, M.A., Nichols, R.A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. B-Biol. Sci. 263, 1619–1626. https://doi.org/10.1098/rspb.1996.0237.

Billiard, S., Castric, V., Llaurens, V., 2021. The integrative biology of genetic dominance. Biol. Rev. 96, 2925–2942. https://doi.org/10.1111/brv.12786.

Brookfield, J.F.Y., 2020. Genetic variation: Harmful recessive mutations have unexpected effects on variation. Curr. Biol. 30, R16-R18, 10.1016/j.cub.2019.11.040.

Burri, R., 2017. Interpreting differentiation landscapes in the light of long-term linked selection. Evol. Lett. 1, 118–131. https://doi.org/10.1002/evl3.14.

Flanagan, S.P., Jones, A.G., 2017. Constraints on the F-ST-heterozygosity outlier approach. J. Hered. 108, 561–573. https://doi.org/10.1093/jhered/esx048.

Graham, A.M., Lavretsky, P., Munoz-Fuentes, V., Green, A.J., Wilson, R.E., McCracken, K.G., 2018. Migration-selection balance drives genetic differentiation in genes associated with high-altitude function in the speckled teal (Anas flavirostris) in the Andes. Genome Biol. Evol. 10, 14–32. https://doi.org/10.1093/gbe/evx253.

Haldane, J.B.S., 1930. A mathematical theory of natural and artificial selection (Part VI. Isolation). Proc. Cambr. Philos. Soc. 26, 220–230.

Haller, B.C., Messer, P.W., 2019. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. Mol. Biol. Evol. 36, 632–637. https://doi.org/10.1093/molbev/msy228.

Hoekstra, H.E., Drumm, K.E., Nachman, M.W., 2004. Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. Evolution 58, 1329–1341.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat. Rev. Genet. 10, 639-650.

Huber, C.D., Durvasula, A., Hancock, A.M., Lohmueller, K.E., 2018. Gene expression drives the evolution of dominance. Nat. Commun. 9 https://doi.org/10.1038/s41467-018-05281-7.

Jewett, E.M., Steinrucken, M., Song, Y.S., 2016. The effects of population size histories on estimates of selection coefficients from time-series genetic data. Mol. Biol. Evol. 33, 3002–3027. https://doi.org/10.1093/molbev/msw173.

Linnen, C.R., Hoekstra, H.E., 2009. Measuring natural selection on genotypes and phenotypes in the wild. Cold Spring Harb. Symp. Quant. Biol. 74, 155–168.

Lotterhos, K.E., Whitlock, M.C., 2014. Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests. Mol. Ecol. 23, 2178–2192. https://doi.org/10.1111/mec.12725.

Lotterhos, K.E., Whitlock, M.C., 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol. Ecol. 24, 1031–1046. https://doi.org/10.1111/mec.13100.

Petry, D., 1983. The effect on neutral gene flow of selection at a linked locus. Theor. Popul Biol. 23, 300–313. https://doi.org/10.1016/0040-5809(83)90020-5.

Rudman, S.M., Barbour, M.A., Csillery, K., Gienapp, P., Guillaume, F., Hairston, N.G., Hendry, A.P., Lasky, J.R., Rafajlovic, M., Rasanen, K., Schmidt, P.S., Seehausen, O., Therkildsen, N.O., Turcotte, M.M., Levine, J.M., 2018. What genomic data can reveal about eco-evolutionary dynamics. Nat. Ecol. Evol. 2, 9–15. https://doi.org/10.1038/s41559-017-0385-2.

Savolainen, O., Lascoux, M., Merila, J., 2013. Ecological genomics of local adaptation. Nat. Rev. Genet. 14, 807–820. https://doi.org/10.1038/nrg3522.

Sunde, J., Yildirim, Y., Tibblin, P., Forsman, A., 2020. Comparing the performance of microsatellites and RADseq in population genetic studies: Analysis of data for pike (Esox lucius) and a synthesis of previous studies. Front. Genet. 11 https://doi.org/10.3389/fgene.2020.00218.

Tataru, P., Simonsen, M., Bataillon, T., Hobolth, A., 2017. Statistical inference in the Wright-Fisher model using allele frequency data. Syst. Biol. 66, E30–E46. https://doi.org/10.1093/sysbio/syw056.

Thurman, T.J., Barrett, R.D.H., 2016. The genetic consequences of selection in natural populations. Mol. Ecol. 25, 1429–1448. https://doi.org/10.1111/mec.13559.

Vitalis, R., Gautier, M., Dawson, K.J., Beaumont, M.A., 2014. Detecting and measuring selection from gene frequency data. Genetics 196, 799. https://doi.org/10.1534/genetics.113.152991.

Whitlock, M.C., 2011. G '(ST) and D do not replace F-ST. Mol. Ecol. 20, 1083–1091. https://doi.org/10.1111/j.1365-294X.2010.04996.x.

Wright, S., 1931. Evolution in Mendelian populations Genetics 16, 97-159, 10.1007/bf02459575.