Studying models of balancing selection using phase-type theory

Kai Zeng[*,1], **Brian Charlesworth**[†] **and Asger Hobolth**[‡]

[*]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom, [†]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom, [‡]Department of Mathematics, Aarhus University, DK-8000 Aarhus C, Denmark

ABSTRACT Balancing selection (BLS) is the evolutionary force that maintains high levels of genetic variability in many important genes. To further our understanding of its evolutionary significance, we analyse models with BLS acting on a biallelic locus: an equilibrium model with long-term BLS, a model with long-term BLS and recent changes in population size, and a model of recent BLS. Using phase-type theory, a mathematical tool for analysing continuous time Markov chains with an absorbing state, we examine how BLS affects polymorphism patterns in linked neutral regions, as summarised by nucleotide diversity, the expected number of segregating sites, the site frequency spectrum, and the level of linkage disequilibrium (LD). Long-term BLS affects polymorphism patterns in a relatively small genomic neighbourhood, and such selection targets are easier to detect when the equilibrium frequencies of the selected variants are close to 50%, or when there has been a population size reduction. For a new mutation subject to BLS, its initial increase in frequency in the population causes linked neutral regions to have reduced diversity, an excess of both high and low frequency derived variants, and elevated LD with the selected locus. These patterns are similar to those produced by selective sweeps, but the effects of recent BLS are weaker. Nonetheless, compared to selective sweeps, non-equilibrium polymorphism and LD patterns persist for a much longer period under recent BLS, which may increase the chance of detecting such selection targets. An R package for analysing these models, among others (e.g., isolation with migration), is available.

KEYWORDS balancing selection; phase-type theory; demographic changes; linkage disequilibrium; site frequency spectrum; selective sweep

Balancing selection refers to a type of natural selection that maintains genetic variability in populations (Fisher 1922; Charlesworth 2006; Fijarczyk and Babik 2015). Genes known to be under balancing selection are often involved in important biological functions. Examples include the major histocompatibility complex (MHC) genes in vertebrates (Spurgin and Richardson 2010), plant self-incompatibility genes (Castric and Vekemans 2004), mating-type genes in fungi (van Diepen *et al.* 2013), genes underlying host-pathogen interactions (Bakker *et al.* 2006; Hedrick 2011), inversion polymorphisms (Dobzhansky 1970), and genes underlying phenotypic polymorphisms in many different organisms (e.g., Johnston *et al.* 2013; Küpper *et al.* 2016; Kim *et al.* 2019). More recently, it has been proposed that a related process, known as associative overdominance, may play a significant role in shaping diversity patterns in genomic regions with very low recombination rates (Becher *et al.* 2020; Gilbert *et al.* 2020). These facts highlight the importance of studying balancing selection.

Understanding how balancing selection affects patterns of genetic variability is a prerequisite for detecting genes under this type of selection. The best studied models involve long-term selection acting at a single locus (Strobeck 1983; Hudson and Kaplan 1988; Takahata 1990; Takahata and Nei 1990; Vekemans and Slatkin 1994; Nordborg 1997; Takahata and Satta 1998; Innan and Nordborg 2003). It is well known that, in addition to maintaining diversity at the selected locus, long-term balancing selection increases diversity at closely linked neutral sites. This reflects an increased coalescence time for the gene tree connecting the alleles in a sample from the current population. When this tree is sufficiently deep, it is possible for the ages of the alleles to exceed the species' age, leading to trans-species polymorphism. Furthermore, long-term balancing selection alters the site frequency spectrum (SFS) at linked neutral sites, causing

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 2^{nd} April, 2021

[1]Corresponding author: Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom. Email: k.zeng@sheffield.ac.uk.

an excess of intermediate frequency derived variants. These properties underlie most of the methods used for scanning large-scale genomic data for targets of balancing selection (Andres *et al.* 2009; Leffler *et al.* 2013; DeGiorgio *et al.* 2014; Bitarello *et al.* 2018; Cheng and DeGiorgio 2019; Siewert and Voight 2020).

A significant limitation of most previous studies is the assumption that the population is at statistical equilibrium under selection, mutation and genetic drift. In reality, most populations have experienced recent changes in population size. Our ability to analyse data from these populations is limited by the lack of an effective way of making predictions about the joint effects of demographic changes and balancing selection on patterns of genetic variability. Moreover, many cases of balancing selection involve variants that have only recently spread to intermediate frequencies, rather than having been maintained for periods much longer than the neutral coalescence time (e.g. Eanes 1999; Kwiatkowski 2005; Corbett-Detig and Hartl 2012). Indeed, several theoretical studies have suggested that adaptation may occur through the frequent emergence of short-lived balanced polymorphisms (Sellis *et al.* 2011; Connallon and Clark 2014). Because of their young age, there may not be sufficient time for the diversity patterns predicted for long-term balancing selection to emerge. As a result, targets of recent balancing selection are unlikely to be detected by existing methods. This may explain why genome scans have only reported a relatively small number of potential selection targets (Andres *et al.* 2009; Leffler *et al.* 2013; DeGiorgio *et al.* 2014; Bitarello *et al.* 2018; Cheng and DeGiorgio 2019).

Multiple authors have suggested that the emergence of a recent balanced polymorphism will generate diversity patterns that resemble those generated by incomplete selective sweeps caused by positive selection favouring a beneficial mutation (Charlesworth 2006; Sellis *et al.* 2011; Fijarczyk and Babik 2015). In fact, methods designed for detecting sweeps can identify these signals (e.g., Zeng *et al.* 2006). However, there is currently no theoretical framework for studying recent balanced polymorphism, which precludes a detailed comparison with incomplete selective sweeps. Acquiring this knowledge will help us devise methods for distinguishing between balancing selection and positive selection, which will in turn allow us to test hypotheses about the importance of balancing selection in adaptation.

We tackle these problems by using phase-type theory. Briefly, a phase-type distributed random variable describes the time until a finite state continuous time Markov process enters one of its absorbing states. Thus, a phase-type distribution is similar to the distribution of the hitting time (or first passage time) for a diffusion process (Karlin and Taylor 1981; Ross 1996). As an example, imagine that we have taken a sample of n alleles from the population. Going backwards, the time it takes for the process to reach the most recent common ancestor follows a phase-type distribution. Phase-type theory refers to a set of mathematical tools for analysing the properties (e.g., mean and variance) of this type of random variable (Bladt and Nielsen 2017). In a recent study, Hobolth *et al.* (2019) used a time-homogeneous version of the theory to study several population genetic models at statistical equilibrium. Here, we extend this approach by deriving several useful results under a time-inhomogeneous framework. We use the new theory to analyse three models of balancing selection: an equilibrium model of long-term balancing selection, a model with long-term balancing selection and changes in population size, and a model of recent balancing selection. The analysis of the last model is accompanied by a comparison with a comparable selective sweep model.

For each of these models, we calculate summary statistics that are useful for understanding the effects of selection on diversity patterns in nearby genomic regions. Specifically, for a sample of alleles collected from a linked neutral site, we obtain (1) the expected pairwise coalescence time (proportional to nucleotide diversity π), (2) the expected level of linkage disequilibrium (LD) between the selected locus and the focal neutral site, (3) the total branch length of the gene tree (proportional to the total number of segregating sites $), and (4) the site frequency spectrum (SFS). Our results extend previous studies of the equilibrium model by providing a unifying framework for obtaining these statistics. The analysis of the non-equilibrium models provides useful insights that can be used for devising new genome scan methods or parameter estimation methods. We conclude the study by discussing the usefulness of phase-type theory in population genetics.

An equilibrium model of balancing selection

Consider a diploid, randomly mating population. The effective population size N_e is assumed to be constant over time. An autosomal locus with two alleles A_1 and A_2 is under balancing selection. The intensity of selection is assumed to be sufficiently strong and constant over time that the frequencies of the two alleles remain at their equilibrium values indefinitely. Denote the equilibrium frequencies of A_1 and A_2 by $\hat{p}_1$ and $\hat{p}_2$, respectively ($\hat{p}_1 + \hat{p}_2 = 1$). This set-up can accommodate any model of long-term balancing selection (with or without reversible mutation between A_1 and A_2), as long as it produces stable allele frequencies. A random sample of n alleles have been taken from a linked neutral locus. The recombination frequency between this locus and the selected locus is denoted by r. In the following four subsections, we use time-homogeneous phase-type theory to calculate the four statistics mentioned at the end of the Introduction. This introduces the methodology and notation, and sets the stage for extending the analysis to non-equilibrium models in later sections. A similar model has been investigated previously using different approaches (Strobeck 1983; Hudson and Kaplan 1988; Nordborg 1997). However, these do not provide analytical expressions for the SFS.

The mean coalescence time for a sample size of two

An allele at the neutral locus is associated with either A_1 or A_2 at the selected site (i.e., a neutral allele is on the same haplotype as either A_1 or A_2). The sample is therefore in one of three possible states (Figure 1). In state 1, both alleles are associated with A_2. In state 2, one allele is associated with A_1, and the other is associated with A_2. In state 3, both alleles are associated with A_1. Take state 1 as an example. An allele currently associated with A_2 was associated with A_1 in the previous generation either because there was an A_1 to A_2 mutation during gamete production, or because the parent was an A_1A_2 heterozygote and there was a recombination event. Define v_{21} as the *backward* mutation rate (see Supplementary Text S.1). The first event occurs with probability v_{21}, and the second event occurs with probability $r\hat{p}_1$. The probability that the focal allele becomes associated with A_1 in the previous generation is $m_{21} = v_{21} + r\hat{p}_1$. The two alleles in state 1 may share a common ancestor in the previous generation. Because the frequency of A_2 is $\hat{p}_2$, a total of $2N_e\hat{p}_2$ alleles were associated with A_2 in the previous generation. The chance that the two alleles coalesce is $1/(2N_e\hat{p}_2)$.

Under the standard assumption that the probability of occurrence of more than one event in one generation is negligible, the probability that the two alleles in state 1 remain unchanged for z generations is:

$$\left(1-2m_{21}-\frac{1}{2N_e\hat{p}_2}\right)^z \approx e^{-\left(2m_{21}+\frac{1}{2N_e\hat{p}_2}\right)z} = e^{-\left(2M_{21}+\frac{1}{\hat{p}_2}\right)t} \tag{1}$$

where $M_{21} = 2N_e m_{21} = \mu_{21} + \rho\hat{p}_1$, $\mu_{21} = 2N_e v_{21}$, $\rho = 2N_e r$, and $t = z/(2N_e)$.

We have scaled time in units of $2N_e$ generations, and will use this convention throughout unless stated otherwise. Using this timescale, when in state 1, the waiting time to the next event follows an exponential distribution with rate parameter $2M_{21} + (1/\hat{p}_2)$. Given that an event has occurred, the probability that it is caused by one of the two alleles becoming associated with A_1 is $2M_{21}/(2M_{21}+1/\hat{p}_2)$, and the probability that it is caused by the coalescence of the two alleles is $(1/\hat{p}_2)/(2M_{21}+1/\hat{p}_2)$. As illustrated in Figure 1, the first possibility moves the process from state 1 to state 2, whereas the second possibility terminates the process by moving it into the absorbing state where the most recent common ancestor (MRCA) is reached (state 4).

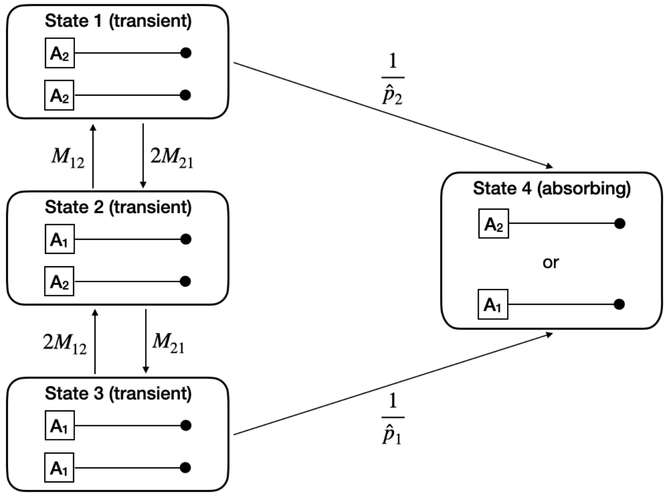


Figure 1 Transition rates between the states of the equilibrium balancing selection model for a sample size of two. A_1 and A_2 are the variants at the locus under balancing selection, with equilibrium frequencies $\hat{p}_1$ and $\hat{p}_2$, respectively. The *backward* mutation rate between A_i and A_j is v_{ij} per generation. The thin horizontal lines represent haplotypes, and the neutral locus is represented by a black dot. The recombination frequency between the two loci is r. Time is scaled in units of $2N_e$ generations. The rate at which a neutral allele associated with A_i becomes associated with A_j is $M_{ij} = \mu_{ij} + \rho\hat{p}_j$, where $\mu_{ij} = 2N_e v_{ij}$ and $\rho = 2N_e r$. Two neutral alleles associated with A_i coalesce at rate $1/\hat{p}_i$.

We can derive the transition rates between all four states of the process using similar arguments (Figure 1). This model is analogous to a two-deme island model in which $2N_e\hat{p}_1$ and $2N_e\hat{p}_2$ are the sizes of the two demes, and M_{12} and M_{21} are the scaled migration rates (e.g., Hudson and Kaplan 1988; Slatkin 1991; Nordborg 1997). Hereafter, we refer to the sub-population consisting of alleles associated with A_1 or A_2 as allelic class 1 or 2, respectively.

We can analyse this model efficiently using time-homogeneous phase-type theory (Hobolth *et al.* 2019). To this end, we define an intensity (rate) matrix as:

$$\mathbf{\Lambda} = \begin{bmatrix} -2M_{21}-\frac{1}{\hat{p}_2} & 2M_{21} & 0 & \frac{1}{\hat{p}_2} \\ M_{12} & -M_{12}-M_{21} & M_{21} & 0 \\ 0 & 2M_{12} & -2M_{12}-\frac{1}{\hat{p}_1} & \frac{1}{\hat{p}_1} \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

The first three rows in $\mathbf{\Lambda}$ are for states 1, 2, and 3, respectively. In row i ($i \in \{1,2,3\}$), the $j-$th element is the rate of jumping from state i to state j ($j \neq i$ and $j \in \{1,2,3,4\}$), and the diagonal element is the negative of the sum of all the other elements in this row. All elements of the last row of $\mathbf{\Lambda}$ are zero because state 4 is absorbing, so that the rate of leaving it is zero.

We can write $\mathbf{\Lambda}$ in a more compact form:

$$\mathbf{\Lambda} = \begin{bmatrix} \boldsymbol{S} & \boldsymbol{s} \\ \vec{0} & 0 \end{bmatrix} \tag{3}$$

where $\boldsymbol{S}$ represents the 3-by-3 sub-matrix in the upper left corner of $\mathbf{\Lambda}$, $\boldsymbol{s}^T = (\frac{1}{\hat{p}_2}, 0, \frac{1}{\hat{p}_1})$ consists of the first three elements in the last column of $\mathbf{\Lambda}$ (the superscript T denotes matrix transposition), and $\vec{0}$ is a row vector of zeros. Thus, $\boldsymbol{S}$ contains the transition rates between the transient states, and $\boldsymbol{s}$ contains the rates of jumping to the absorbing state. $\boldsymbol{S}$ and $\boldsymbol{s}$ are referred to as the sub-intensity matrix and the exit rate vector, respectively.

Assume that i and $2-i$ alleles in the sample are associated with A_1 and A_2, respectively. The time it takes for the process to reach the most recent common ancestor (MRCA) of the pair of alleles is a random variable that follows a phase-type distribution (Bladt and Nielsen 2017; Hobolth *et al.* 2019). To calculate the expected value of this random variable, denoted by $T_{i,2-i}$, we define the Green's matrix $\boldsymbol{U} = \{u_{ij}\}$, where u_{ij} is the expected amount of time the process spends in state j prior to reaching the MRCA, provided that the initial state is i ($i, j \in \{1,2,3\}$). As shown in Supplementary Text S.2, $\boldsymbol{U}$ can be calculated as:

$$\boldsymbol{U} = -\boldsymbol{S}^{-1} \tag{4}$$

(see also Theorem 3.1.14 in Bladt and Nielsen (2017)). Take $T_{0,2}$ as an example. The sample is in state 1. The expected amount of time the coalescent process spends in state k before reaching the MRCA is u_{1k} ($k \in \{1,2,3\}$). Thus, $T_{0,2} = \sum_{k=1}^{3} u_{1k}$. More generally, we have

$$T_{i,2-i} = \sum_{k=1}^{3} u_{i+1,k}. \tag{5}$$

It is also possible to use phase-type theory to obtain the probability density function and all the moments of the coalescence time (Hobolth *et al.* 2019).

Define the initial condition vector as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$, where α_i is the probability that the sample is in state i ($\sum_1^3 \alpha_i = 1$). Thus, for $T_{0,2}$, $\boldsymbol{\alpha} = (1,0,0)$. Further let $\boldsymbol{D}^T = (1,1,1)$. We can rewrite (5) as:

$$T_{i,2-i} = \boldsymbol{\alpha}\boldsymbol{U}\boldsymbol{D}. \tag{6}$$

As we will see later, expressing the results this way allows us to accommodate non-equilibrium situations. The vector $\boldsymbol{D}$ is known as the reward vector. Its k-th element D_k is the rate at which the quantity of interest accrues per unit time while the process stays in state k. Thus, the total contribution to $T_{i,2-i}$ made by state k is $u_{i+1,k}D_k$.

It is possible to obtain $\boldsymbol{U}$ analytically for the general model with reversible mutation between A_1 and A_2, as specified by (2). However, its terms are complicated, and are not shown. For sites that are not very tightly linked to the selected locus, movements of lineages between the two allelic classes are primarily driven by recombination (i.e., $\rho \gg \mu_{ij}$). Furthermore, with only two alleles at the selected locus, the general model is most appropriate for cases where the selected locus contains a small handful of nucleotides. In this case μ_{ij} is of the order of the average nucleotide diversity at neutral sites (e.g., about 0.02 in *Drosophila melanogaster* or about 0.001 in humans). For most applications, therefore, it is sufficient to work with a simplified model with $\mu_{ij} = 0$. In this case, we have $\hat{p}_1 M_{12} = \hat{p}_2 M_{21}$ (i.e., there is conservative migration; Nagylaki (1980)), which leads to:

$$\boldsymbol{U} = \begin{bmatrix} \frac{\hat{p}_2+2\hat{p}_1\hat{p}_2^3\rho}{1+2\hat{p}_1\hat{p}_2\rho} & 2\hat{p}_1\hat{p}_2 & \frac{2\hat{p}_1^3\hat{p}_2\rho}{1+2\hat{p}_1\hat{p}_2\rho} \\ \hat{p}_2^2 & 2\hat{p}_1\hat{p}_2 + \frac{1}{\rho} & \hat{p}_1^2 \\ \frac{2\hat{p}_1\hat{p}_2^3\rho}{1+2\hat{p}_1\hat{p}_2\rho} & 2\hat{p}_1\hat{p}_2 & \frac{\hat{p}_1+2\hat{p}_1^3\hat{p}_2\rho}{1+2\hat{p}_1\hat{p}_2\rho} \end{bmatrix}. \tag{7}$$

Summing the three rows, we have:

$$\begin{cases} T_{0,2} = 1 - \frac{\hat{p}_1(\hat{p}_1-\hat{p}_2)}{1+2\hat{p}_1\hat{p}_2\rho} \\ T_{1,1} = 1 + \frac{1}{\rho} \\ T_{2,0} = 1 + \frac{(\hat{p}_1-\hat{p}_2)\hat{p}_2}{1+2\hat{p}_1\hat{p}_2\rho} \end{cases} \tag{8}$$

The results in (8) are the same as those derived by Nordborg (1997). The additional insight obtained here is given by (7). For instance, regardless of whether the initial state is 1 or 3, the process spends, on average, an equal amount of time in state 2 before coalescence (i.e., $u_{12} = u_{32}$ in (7)). The results presented in Figure S1 further confirm that the simplified model should suffice in most cases, because the general model is well approximated by the simplified model for large enough ρ.

Let $\pi_{i,2-i}$ be the expected diversity when i and $2-i$ alleles in the sample are associated with A_1 and A_2, respectively. Under the infinite sites model (Kimura 1969), $\pi_{i,2-i} = 2\theta T_{i,2-i}$, where $\theta = 2N_e v$ and v is the mutation rate per generation at the neutral site. To put the discussion in context, we note that the expected coalescence time for two alleles is 1 under the neutral model with constant population size. From (8), we can see that $T_{1,1}$ is independent of $\hat{p}_1$ and $\hat{p}_2$, and is always greater than 1. For $T_{0,2}$, it is < 1 or > 1 when $\hat{p}_2$ is < 0.5 or > 0.5, respectively. Similarly, $T_{2,0}$ is < 1 or > 1 when $\hat{p}_1$ is < 0.5 or > 0.5, respectively. These trends hold even when there is reversible mutation between A_1 and A_2 (Figure S1).

In reality, the selected variants are often unknown, and detecting targets of balancing selection typically relies on investigating how diversity levels change along the chromosome (Charlesworth 2006; Fijarczyk and Babik 2015). It is therefore useful to consider the expected coalescence time for two randomly sampled alleles at the neutral site, defined as:

$$T = \hat{p}_1^2 T_{2,0} + 2\hat{p}_1\hat{p}_2 T_{1,1} + \hat{p}_2^2 T_{0,2} = 1 + \frac{\hat{p}_1\hat{p}_2(\rho+2)}{\rho(1+2\hat{p}_1\hat{p}_2\rho)} \tag{9}$$

where the results in (8) are used. The nucleotide site diversity is given by $\pi = 2T\theta$. Figure 2 shows that the diversity level is highest when $\hat{p}_1 = \hat{p}_2 = 0.5$. This is also true when there is reversible mutation between A_1 and A_2 (Figure S2). The simplified model is inherently symmetrical. For example, the curve for $\hat{p}_1 = 0.25$ is identical to that for $\hat{p}_1 = 0.75$. In all cases, marked effects on diversity are only seen in the immediate genomic neighbourhood of the selected site where ρ is of order 1 or less.

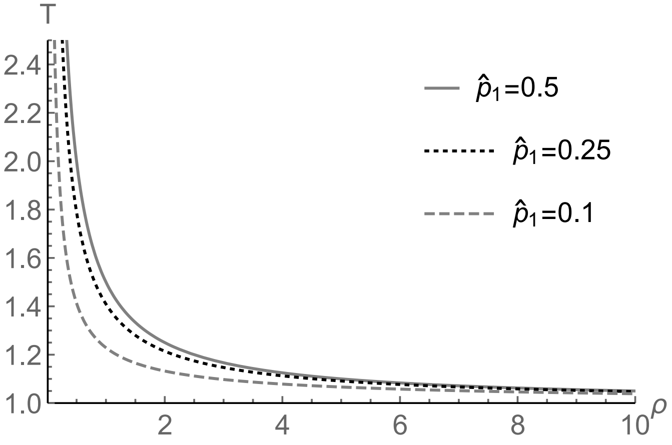


Figure 2 The expected pairwise coalescence time as a function of ρ. The simplified model with $\mu_{12} = \mu_{21} = 0$ is considered. $\hat{p}_1$ is the equilibrium frequency of A_1 at the selected locus.

LD between the selected locus and a linked neutral site

The expected pairwise coalescence time obtained in the previous section can be used to calculate a measure of LD between the two loci (Charlesworth *et al.* 1997). Assume that the neutral locus is segregating for two variants B_1 and B_2. Let the frequencies of B_1 in allelic class 1 and 2 be x and y, respectively. Thus, the frequency of B_1 in the population is $q_1 = \hat{p}_1 x + \hat{p}_2 y$, and that of B_2 is $q_2 = 1 - q_1$. Let $\delta = x - y$. The coefficient of LD between the two loci is given by $D = \hat{p}_1\hat{p}_2\delta$ (see p. 410 of Charlesworth and Charlesworth 2010). The corresponding correlation coefficient is $R^2 = D^2/(\hat{p}_1\hat{p}_2 q_1 q_2)$. It is impossible to derive a simple expression for $\mathbb{E}[R^2]$. A widely-used alternative can be written as:

$$\sigma^2 = \frac{\mathbb{E}[D^2]}{\mathbb{E}[\hat{p}_1\hat{p}_2 q_1 q_2]} = \frac{\hat{p}_1^2\hat{p}_2^2\mathbb{E}[\delta^2]}{\hat{p}_1\hat{p}_2\mathbb{E}[q_1 q_2]} = \frac{\hat{p}_1\hat{p}_2\mathbb{E}[\delta^2]}{\mathbb{E}[q_1 q_2]} \tag{10}$$

where we have used the fact that $\hat{p}_1$ and $\hat{p}_2$ are assumed to be constant (Ohta and Kimura 1971; Strobeck 1983; McVean 2002). Note that $\pi = 2\mathbb{E}[q_1 q_2]$ is the expected diversity at the neutral site.

As discussed in the previous section, we have $\pi = 2\theta T$ under the infinite sites model. To relate $E[\delta^2]$ to the expected pairwise coalescence times, we first define the expected diversity within allelic class 1 and allelic class 2 as $\pi_{A1} = 2\mathbb{E}[x(1-x)]$ and $\pi_{A2} = 2\mathbb{E}[y(1-y)]$, respectively. Again, under the infinite sites model, we have $\pi_{A1} = 2\theta T_{2,0}$ and $\pi_{A2} = 2\theta T_{0,2}$. In addition, let the weighted within allelic class diversity be $\pi_A = \hat{p}_1\pi_{A1} + \hat{p}_2\pi_{A2}$. Note that $\pi - \pi_A = 2\mathbb{E}[q_1 q_2 - \hat{p}_1 x(1-x) - \hat{p}_2 y(1-y)] = 2\hat{p}_1\hat{p}_2\mathbb{E}[\delta^2]$. Inserting these results into the right-most term of (10), we have:

$$\sigma^2 = \frac{\pi - \pi_A}{\pi} = \frac{T - T_A}{T} \tag{11}$$

where $T_A = \hat{p}_1 T_{2,0} + \hat{p}_2 T_{0,2}$ is the weighted average within allelic class coalescence time. Note that σ^2 has the same form as the fixation indices (e.g., F_{ST}) widely used in studies of structured populations. This close relationship between LD and the fixation indices was first pointed out by Charlesworth *et al.* (1997), who referred to σ^2 as F_{AT}. Our treatment here clarifies the relevant statements in this previous study. It also provides a genealogical interpretation of the results of Strobeck (1983).

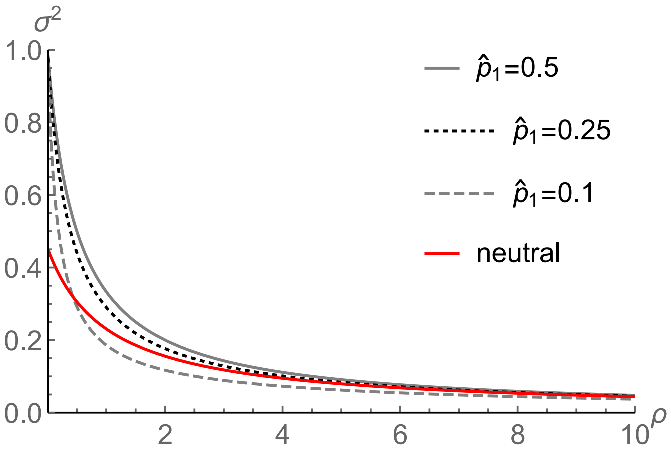


Figure 3 The level of LD between the selected and neutral loci as a function of ρ. The simplified model with $\mu_{12} = \mu_{21} = 0$ is considered. The neutral expectation for σ^2 is also included.

Figure 3 shows σ^2 as a function of ρ generated under the simplified model with $\mu_{12} = \mu_{21} = 0$. The level of LD between the selected and neutral loci is highest when $\hat{p}_1 = \hat{p}_2 = 0.5$, and decreases as $\hat{p}_1$ moves close to either 0 or 1 (note that the model is symmetrical such that, for $0 < z < 1$, the curve for $\hat{p}_1 = z$ is identical to that for $\hat{p}_1 = 1 - z$). As expected, reversible mutation between A_1 and A_2 lowers LD by increasing the rate at which lineages move between the two allelic classes (Figure S3). These results mirror those described above for diversity levels. Together they show that the effect of balancing selection on linked diversity and LD patterns is largest when the equilibrium frequencies of the selected variants are close to 50%.

It is informative to compare LD patterns under balancing selection with those under neutrality (i.e., $\sigma^2 = (5 + \rho)/(11 + 13\rho + 2\rho^2)$; Ohta and Kimura 1971). With balancing selection and $\hat{p}_1 = 0.5$, elevated LD is observed when $\rho < 4$ (Figure 3). With $\hat{p}_1 = 0.1$, LD is higher than neutral expectation when $\rho < 0.5$, and it becomes lower than the neutral level when $\rho > 0.5$. Considering crossing over alone, the scaled recombination rate per site is of the order of 0.002 in humans, and 0.01 in *Drosophila*. These values go up substantially if we also take into account gene conversion (e.g., Campos and Charlesworth 2019). Thus, even when the effect of balancing selection is at its maximum, the region affected is small. The effect becomes rather insubstantial when the equilibrium frequency is close to 0 or 1, suggesting that such selection targets are probably extremely difficult to detect.

Total branch length

We now consider the situation when a sample of n alleles is available, with n_1 of them associated with A_1 and n_2 with A_2 ($n_1 + n_2 = n$). Let L_{n_1,n_2} be the expected total branch length of the gene tree that describes the ancestry of the sample with respect to a neutral site linked to the selected locus. Note that, when $n = 2$, $L_{n_1,n_2} = 2T_{n_1,n_2}$. The results in Figure 2 imply that close genetic linkage to a locus under balancing selection will result in an increase in the total branch length. Because the expected number of segregating sites in the sample is given by $\theta L_{n_1,n_2}$ under the infinite sites model, we expect to see more polymorphic sites in regions surrounding targets of long-term balancing selection. This theoretical expectation underlies several tests for balancing selection (Hudson *et al.* 1987; DeGiorgio *et al.* 2014).

To illustrate the calculation, consider a sample with three alleles. It can be in one of four possible states, with states 1, 2, 3, and 4 corresponding to situations where 0, 1, 2, and 3 of the sampled alleles are associated with A_1. Going backwards in time, the coalescent process can move between these states via recombination or mutation between allelic classes. For instance, in state 1 all three alleles are associated with A_2, and the process moves to state 2 at rate $3M_{21}$. When more than one allele is in the same allelic class, coalescence may occur. Again, take state 1 as an example. There are three alleles in allelic class 2, so that the rate of coalescence is $\binom{3}{2}/\hat{p}_2 = 3/\hat{p}_2$. A coalescent event reduces the number of alleles to two, and thus moves the process to one of the three transient states depicted in Figure 1, referred to as states 5, 6, and 7 here. The transition rates between these states, as well as the rates of entering the absorbing state (i.e., the MRCA), are identical to those discussed above (i.e., (2)).

A diagram showing the transition rates between the states in this model can be found in Figure S4. The intensity matrix $\boldsymbol{\Lambda}$ for this model can be defined in the same way as described above, and is displayed in Supplementary Text S.3. $\boldsymbol{\Lambda}$ has a block structure:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{S}_3 & \boldsymbol{S}_{32} & \underline{\underline{0}} \\ \underline{\underline{0}} & \boldsymbol{S}_2 & \boldsymbol{s}_2 \\ \vec{0} & \vec{0} & 0 \end{bmatrix} \quad (12)$$

where $\underline{\underline{0}}$ is a matrix of zeros. $\boldsymbol{S}_3$ is a 4-by-4 matrix and contains the transition rates between states 1 - 4, all with three alleles. $\boldsymbol{S}_{32}$ is a 4-by-3 matrix and contains the rates of coalescent events that move the process from a state with three alleles to one with only two alleles (i.e., from states 1 - 4 to states 5 - 7). Finally, $\boldsymbol{S}_2$ and $\boldsymbol{s}_2$ are the same as the corresponding elements defined in (3). The sub-intensity matrix $\boldsymbol{S}$ is the 7-by-7 sub-matrix in the upper left corner of $\boldsymbol{\Lambda}$, and contains the transition rates between all the transient states.

Taking advantage of the block structure, we can calculate the Green's matrix more efficiently as:

$$\boldsymbol{U} = -\boldsymbol{S}^{-1} = -\begin{bmatrix} \boldsymbol{S}_3 & \boldsymbol{S}_{32} \\ \underline{\underline{0}} & \boldsymbol{S}_2 \end{bmatrix}^{-1} = \begin{bmatrix} -\boldsymbol{S}_3^{-1} & \boldsymbol{S}_3^{-1}\boldsymbol{S}_{32}\boldsymbol{S}_2^{-1} \\ \underline{\underline{0}} & -\boldsymbol{S}_2^{-1} \end{bmatrix}. \quad (13)$$

Recall that $\boldsymbol{U} = \{u_{ij}\}$ and u_{ij} is the expected amount of time the process spends in (transient) state j prior to reaching the MRCA, provided that the initial state is i. If, for instance, we want to calculate $L_{0,3}$, we first note that the sample is in state 1. The process spends, on average, $\sum_{j=1}^{4} u_{1j}$ in states 1 - 4. Because these states have three alleles, the coalescent genealogy must have three lineages. Thus, these four states contribute $3\sum_{j=1}^{4} u_{1j}$ to $L_{0,3}$. Similarly, states 5 - 7, which contain two alleles, contribute $2\sum_{k=5}^{7} u_{1k}$. Putting these together, we have:

$$L_{0,3} = 3\sum_{j=1}^{4} u_{1j} + 2\sum_{k=5}^{7} u_{1k}. \quad (14)$$

More generally, if the sample is in state i, we can define the initial condition vector as $\boldsymbol{\alpha} = \boldsymbol{e}_i$, where $i \in \{1, 2, 3, 4\}$ and $\boldsymbol{e}_i$ is a 1-by-7 vector whose elements are 0 except that the i-th element is 1. If we further define the reward vector as $\boldsymbol{D}^T = (3, 3, 3, 3, 2, 2, 2)$, we have:

$$L_{i,3-i} = \boldsymbol{\alpha U D}. \tag{15}$$

Note that this has the same form as (6). It is also possible to use phase-type theory to obtain the distribution and all the moments of the total branch length (Hobolth *et al.* 2019).

The approach can be easily extended to an arbitrary sample size n. As discussed above (see (9)), for data analysis, it is useful to consider the expected total branch length for a random sample of n alleles, defined as:

$$L = \sum_{i=0}^{n} \binom{n}{i} \hat{p}_1^i \hat{p}_2^{n-i} L_{i,n-i}. \tag{16}$$

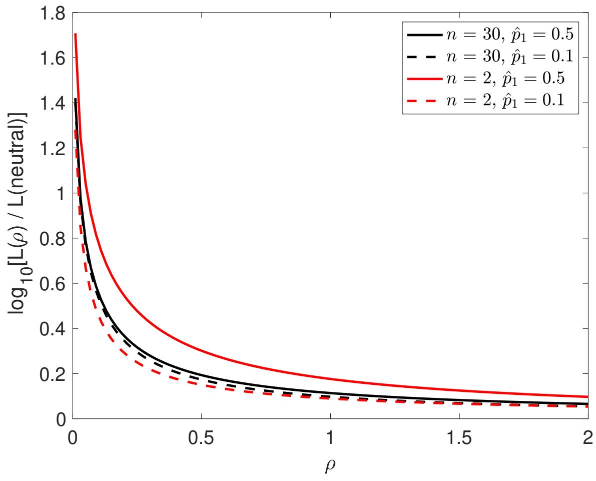


Figure 4 The expected total branch length L for several combinations of sample size (n) and equilibrium frequency of the selected variant A_1 ($\hat{p}_1$). The value of L under balancing selection is divided by its neutral expectation. The y-axis is on the $\log_{10}$ scale.

In Figure 4, we display L for several combinations of sample sizes and variant frequencies at the selected locus. To make the diversity-elevating effect more visible, we divide L by its neutral expectation (i.e., $2\sum_{i=1}^{n-1}\frac{1}{i}$). It is evident that, as n becomes larger, the sensitivity of L to $\hat{p}_1$ decreases, to the extent that, when $n = 30$, L is effectively independent of $\hat{p}_1$. In addition, the strongest signal of elevated diversity appears when $n = 2$ and $\hat{p}_1 = 0.5$, but becomes less pronounced as n increases. To interpret these observations, recall that, when $n = 2$, $\pi = \theta L$, whereas for larger n, θL is the expected number of segregating sites in the sample, denoted by $\mathbb{S}$. In data analysis, the nucleotide site diversity π is typically estimated from samples containing many alleles, and is known to be most sensitive to intermediate frequency variants (Tajima 1989). On the other hand, $\mathbb{S}$ is determined primarily by low frequency variants in the sample. Thus, these results suggest that $\mathbb{S}$ is less informative about balancing selection than π. However, the contrast between $\mathbb{S}$ and π can be used as an index of the departure of the SFS from its expectation at neutral equilibrium (Tajima 1989). This clearly points to the importance of considering the SFS, which is done in the next subsection.

This way of obtaining the total branch length is an alternative to the recursion method used in previous studies (Hudson and Kaplan 1988; DeGiorgio *et al.* 2014). The advantage of the current approach is that it can be extended to accommodate non-equilibrium dynamics such as population size changes and recent selection (see below). The dimension of the sub-intensity matrix $\boldsymbol{S}$ is $d = (n+1) + n + ... + 3 = \frac{1}{2}(n-1)(n+4)$. The numerical complexity increases rapidly because numerical matrix inversion requires $O(n^6)$ operations. However, by making use of the block structure (e.g., (13)), the number of operations is reduced to $O(n^5)$. Thus, this approach is computationally feasible for samples of dozens of alleles.

The site frequency spectrum (SFS)

Again, consider a sample of n alleles at the neutral site, with n_1 and n_2 of them associated with A_1 and A_2, respectively. The i-th element of the SFS is defined as the expected number of segregating sites where the derived variant appears i times in the sample ($0 < i < n$). Note that this definition is different from the standard definition for a panmictic population in that it is conditional on n_1 and n_2. Consider the gene tree for the sample. We refer to a lineage (branch) that is ancestral to i alleles in the sample as a lineage of size i ($0 < i < n$). Under the infinite sites model, mutations on a lineage of size i segregate at frequency i in the sample. Let $\phi_i(n_1, n_2)$ be the expected total length of all lineages of size i in the gene tree. The SFS under the infinite sites model can be expressed as $X_i(n_1, n_2) = \theta\phi_i(n_1, n_2)$ (e.g., Polanski and Kimmel 2003). We can calculate $\phi_i(n_1, n_2)$ using phase-type theory with additional book keeping.

To illustrate the calculation, consider a sample of three alleles. Going backwards in time, before the first coalescent event, all the lineages are size one. After the first coalescent event, one lineage is size two, and the other is size one. Thus, the transient states of the coalescent process can be represented by 4-tuples of the form $(a_{1,1}, a_{1,2}, a_{2,1}, a_{2,2})$ where $a_{i,j}$ is the number of lineages of size j that are currently associated with A_i. We have listed all the transient states in Table 1. The first four states contain three lineages, and the last four contain two lineages. We can determine the transition rates between the states using the same arguments that lead to Figures 1 and S4; the intensity matrix $\boldsymbol{\Lambda}$ is displayed in Supplementary Text S.4. Note that $\boldsymbol{\Lambda}$ has the same form as (12), so that we can obtain $\boldsymbol{U}$ using (13).

Table 1 The transient states for a sample size of three

ID	state	ID	state	ID	state	ID	state
1	$(0,0,3,0)$	2	$(1,0,2,0)$	3	$(2,0,1,0)$	4	$(3,0,0,0)$
5	$(0,0,1,1)$	6	$(1,0,0,1)$	7	$(0,1,1,0)$	8	$(1,1,0,0)$

As an example, if $n_1 = 2$ and $n_2 = 1$, the starting state is 3, so that only the elements in the third row of $\boldsymbol{U}$ are relevant. Because states 1 - 4 contain three size one lineages, they contribute $3\sum_{i=1}^{4} u_{3i}$ to $\phi_1(2, 1)$, but nothing to $\phi_2(2, 1)$. The last four states contain one size one lineage and one size two lineage. Thus, they contribute $\sum_{k=5}^{8} u_{3k}$ to both $\phi_1(2, 1)$ and $\phi_2(2, 1)$. Putting these results together, we have:

$$\begin{cases} \phi_1(2,1) = 3\sum_{i=1}^{4} u_{3i} + \sum_{k=5}^{8} u_{3k} \\ \phi_2(2,1) = \sum_{i=5}^{8} u_{3i} \end{cases} \tag{17}$$

Define the initial condition vector $\boldsymbol{\alpha} = (0,0,1,0,0,0,0,0)$, $\boldsymbol{\phi}(2,1) = (\phi_1(2,1), \phi_2(2,1))$ and

$$\boldsymbol{D}^T = \begin{bmatrix} 3 & 3 & 3 & 3 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}. \qquad (18)$$

We have $\mathbb{E}[\boldsymbol{\phi}(2,1)] = \boldsymbol{\alpha}\boldsymbol{U}\boldsymbol{D}$, which is again in the same form as (6).

We can obtain the other $\boldsymbol{\phi}(i, 3-i)$ by defining the appropriate $\boldsymbol{\alpha}$. In addition to the mean, it is also possible to use phase-type theory to obtain the variance of the SFS, as well as the covariance between different elements of the SFS (Hobolth *et al.* 2019). These results are applicable to any sample size $n \geq 2$. We defer showing results regarding the SFS until a later section where a model of recent balancing selection is analysed.

Obtaining the SFS by working directly with the continuous time Markov process has been shown to be numerically more stable and accurate than approaches that rely on solving the diffusion equation numerically (Kern and Hey 2017). However, a limitation is that the size of the state space increases rapidly with n (Andersen *et al.* 2014). This is true even after exploiting the block structure of the sub-intensity matrix $\boldsymbol{S}$. For instance, when $n = 16$, the dimension of the largest sub-matrix in $\boldsymbol{S}$ is 922, but it increases to 3493 when $n = 20$. However, the flexibility of phase-type theory, especially its ability to accommodate complex non-equilibrium models, makes it a useful tool, as we show next.

A model with strong balancing selection and changes in population size

So far we have only considered a model of balancing selection at statistical equilibrium. In this section, we switch our attention to a non-equilibrium model in which the population size changes in a stepwise manner. Specifically, we consider a diploid, randomly mating population. Looking back in time, its evolutionary history consists of H non-overlapping epochs, such that the effective population size is $N_{e,h}$ in epoch h ($h \in \{1, 2, ..., H\}$). The duration of epoch h is $[t_{h-1}, t_h)$, where $t_0 = 0$ (the present) and $t_H = \infty$. Thus, epoch H, the most ancestral epoch, has an infinite time span, over which the population is at statistical equilibrium. We assume that an autosomal locus is under balancing selection in epoch H, with two alleles A_1 and A_2 at equilibrium frequencies $\hat{p}_1$ and $\hat{p}_2$, respectively. Based on the results shown in the previous sections, we only consider the simplified model without reversible mutation between A_1 and A_2. In addition, we assume that selection is sufficiently strong, and the changes in population size are sufficiently small, that the frequencies of the two alleles remain at $\hat{p}_1$ and $\hat{p}_2$ in the more recent epochs. A similar approach has been applied successfully to modelling the joint effects of background selection and demographic changes (Zeng 2013; Nicolaisen and Desai 2013; Zeng and Corcoran 2015).

Total branch length

As before, consider a neutral site linked to the selected locus, with a sample of n alleles, of which n_1 and n_2 are associated with A_1 and A_2, respectively. Consider the expected total branch length, L_{n_1,n_2}. Here time is scaled in units of $2N_{e,1}$ generations (twice the effective population size in the current epoch). We first note that the current model has the same states as the equilibrium model analysed above (e.g., see Figure S4 for $n = 3$). The main difference between the two models lies in the transition rates between states.

We define the scaled recombination rate as $\rho = 2N_{e,1}r$. The rate at which an allele in allelic class i moves to allelic class j is $M_{ij} = \rho\hat{p}_j$. These have the same form as above (cf. Figure 1). In epoch h, the total number of alleles associated with A_1 in the population is $2N_{e,h}\hat{p}_1$. The probability that two alleles associated with A_1 in the current generation coalesce in the previous generation is $1/(2N_{e,h}\hat{p}_1)$. In other words, the probability that they remain un-coalesced for z generations is:

$$\left(1 - \frac{1}{2N_{e,h}\hat{p}_1}\right)^z \approx \exp\left\{-\frac{z}{2N_{e,h}\hat{p}_1}\right\} = \exp\left\{-\frac{g_h}{\hat{p}_1}t\right\} \qquad (19)$$

where $g_h = N_{e,1}/N_{e,h}$ and $t = z/(2N_{e,1})$. Thus, the coalescent rate between a pair of alleles in allelic class 1 is $g_h/\hat{p}_1$ in epoch h. Similarly, the rate for two alleles in allelic class 2 is $g_h/\hat{p}_2$.

In epoch h, the transition rates between the states are constant, and we can define an associated sub-intensity matrix, $\boldsymbol{S}_h$. We have already noted that the states in the current model are the same as those in the equilibrium model. $\boldsymbol{S}_h$ is very similar to the sub-intensity matrix for the equilibrium model (e.g., (12); see also Supplementary Text S.3). The only differences are (1) ρ is now defined as $2N_{e,1}r$ and (2) terms involving $1/\hat{p}_i$ should be replaced by $g_h/\hat{p}_i$.

Overall, the model has the following parameters: $\hat{p}_1$, ρ, t_1, g_1, t_2, g_2, ..., t_{H-1}, g_{H-1}, and g_H. Among these, $\hat{p}_1$ and ρ are shared across all the epochs, whereas epoch h has two epoch-specific parameters t_h and g_h (note that $t_H = \infty$). We have H sub-intensity matrices: $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, ..., $\boldsymbol{S}_H$. In Supplementary Text S.5, we introduce time-inhomogeneous phase-type theory and prove the following result:

Theorem 1. *Consider a continuous time Markov chain with finite state space $\{1, 2, ..., K, K+1\}$, where states $1, ..., K$ are transient, and state $K+1$ is absorbing. Assume that the time interval $[0, \infty)$ is subdivided into H non-overlapping epochs. The duration of epoch h is $[t_{h-1}, t_h)$, where $1 \leq h \leq H$, $t_0 = 0$, and $t_H = \infty$. The sub-intensity matrix for epoch h is denoted by $\boldsymbol{S}_h$. Then the Green's matrix is:*

$$\boldsymbol{U} = \sum_{h=1}^{H} \left[\prod_{i=1}^{h-1} e^{\boldsymbol{S}_i d_i}\right] \boldsymbol{U}_h \qquad (20)$$

where $d_h = t_h - t_{h-1}$, $\boldsymbol{U}_h = e^{\boldsymbol{S}_h d_h}\boldsymbol{S}_h^{-1} - \boldsymbol{S}_h^{-1}$, and $e^{\boldsymbol{S}_h d_h} = 0$ if $d_h = \infty$.

$\boldsymbol{U}_h = \{u_{ij,h}\}$ in (20) is the Green's matrix for epoch h. Its element $u_{ij,h}$ is the expected amount of time the process stays in state j in this epoch if it enters the epoch in state i. Intuitively, $u_{ij,h}$ equals the amount of time the process spends in state j had the duration of epoch h been $[0, \infty)$ (represented by $-\boldsymbol{S}_h^{-1}$) minus the amount of time it spends in state j in $[d_h, \infty)$ (represented by $-e^{\boldsymbol{S}_h d_h}\boldsymbol{S}_h^{-1}$). Let $\prod_{i=1}^{h-1} e^{\boldsymbol{S}_i d_i} = \{p_{ij,h-1}\}$. The element $p_{ij,h-1}$ is the probability that the process starts from state i at $t_0 = 0$ and is in state j by the end of epoch $h-1$ at time t_{h-1}. Thus, the overall Green's matrix $\boldsymbol{U}$ is the weighted mean of the epochs' contributions, with the weights being the probabilities that the process enters the epochs in a particular state.

Applying this theorem requires the evaluation of matrix exponentials. Although this can be done analytically for certain models (e.g., Waltoft and Hobolth 2018), it is not feasible for the models considered here. We instead employ numerical methods (Al-Mohy and Higham 2010; Moler and Van Loan 2003), as implemented in the `expm` function in Matlab or the `expm` package in R. The computational cost for obtaining $e^{\boldsymbol{S}_h d_h}$ is typically $O(d^3)$,

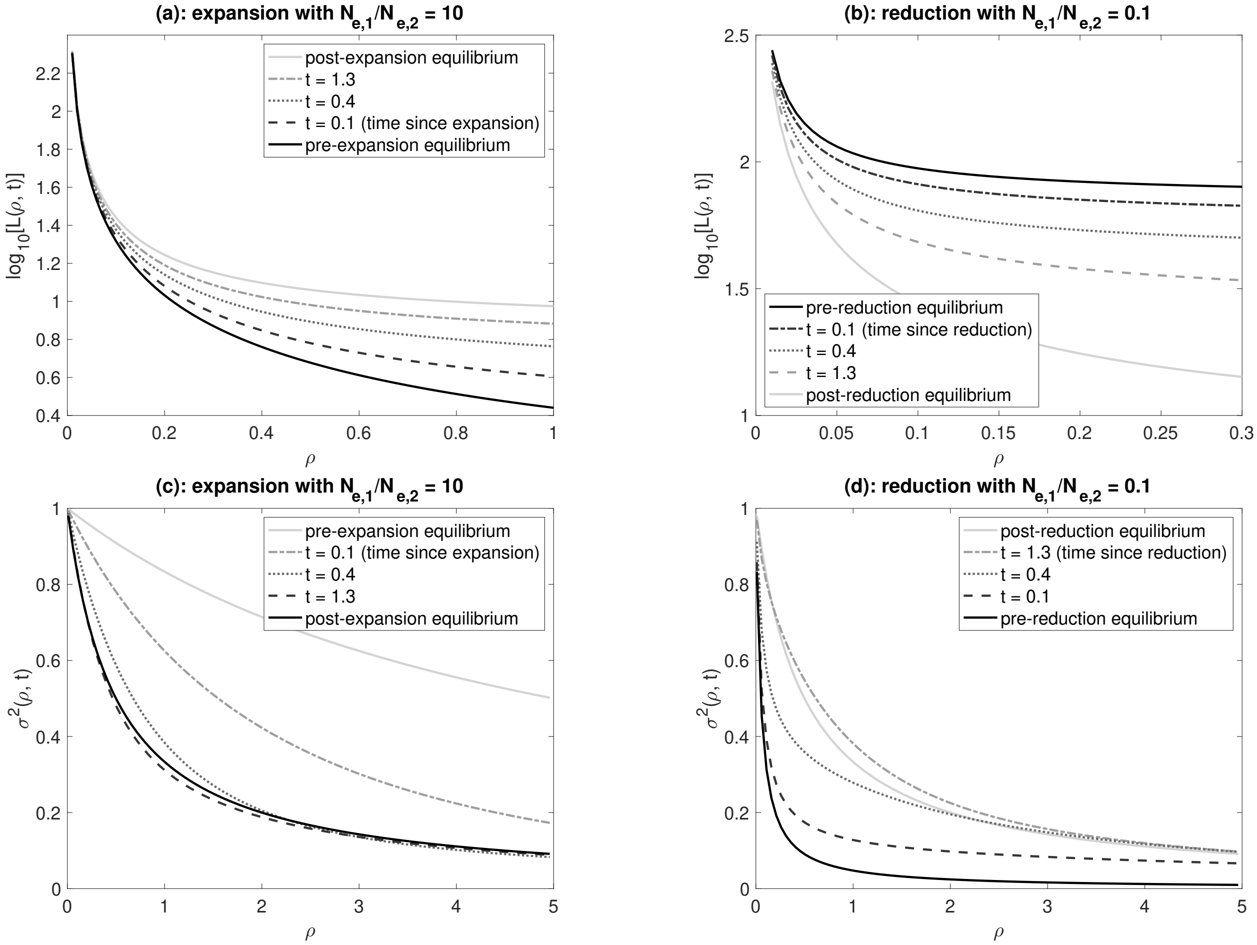


Figure 5 Expected total branch length and LD as a function of ρ and t. The population experienced a one-step change in population size at time t before the present. The population size in the present and ancestral epochs are $N_{e,1}$ and $N_{e,2}$, respectively. Time is scaled in units of $2N_{e,1}$ generations. The selected alleles A_1 and A_2 are at equilibrium frequencies $\hat{p}_1 = \hat{p}_2 = 0.5$. The sample size is $n = 20$.

where d is the dimension of $\boldsymbol{S}_h$. Once $\boldsymbol{U}$ has been calculated, the expected total branch length is given by $L_{n_1,n_2} = \boldsymbol{\alpha U D}$ (see (15)).

In Figures 5a and b, we show L, the expected total branch length, for a random sample of $n = 20$ alleles (see (16)), under either a one-step population size increase or a one-step population size reduction. The population size change occurred at time t before the present. Because L is insensitive to $\hat{p}_1$ when n is relatively large (Figure 4), we only consider $\hat{p}_1 = 0.5$ (the results are qualitatively very similar with $n = 2$; not shown). Neutral diversity levels in genomic regions closely linked to the selected site are affected by recent population size changes to a much smaller extent than regions farther afield. This is because, for small ρ, migration of lineages between allelic classes is slow, such that the tree size is mainly determined by the divergence between allelic classes rather than drift within allelic classes. The importance of the divergence component increases with decreasing ρ. In particular, when there has been a recent reduction in population size, this effect protects against the loss of neutral polymorphisms in a larger genomic region (Figure 5b). Consequently, all else being equal, strong balancing selection affects a bigger stretch of the genome and produces a higher peak of diversity in smaller populations, potentially making them easier to detect. A similar observation has been made in models with self-fertilisation and background selection (Nordborg *et al.* 1996).

Note that, although we have focused on calculating the total branch length, Theorem 1 can also be used to calculate the SFS. This can be done by defining an appropriate state space (e.g., Table 1) and a suitable reward matrix (e.g., (18)). We will demonstrate these calculations later when we analyse a model of recent balancing selection.

LD between the selected locus and a linked neutral site

The measure of LD can be calculated by replacing T and T_A in (11) with $T(t)$ and $T_A(t)$. In Figures 5c and d, we can see that σ^2 converges to its new equilibrium level at a much higher rate than the level of diversity, which is a well-known effect (e.g., McVean 2002). Interestingly, σ^2 appears to approach its new equilibrium in a non-monotonic way. For instance, in Figure 5c, LD levels at $t = 0.4$ are temporarily higher than the equilibrium value (the solid black curve), but become lower than the equilibrium value at $t = 1.3$. In Figure 5d, we can see that the level of LD is higher, and extends further, after the population size reduction (see also Figure S5). These results further suggest that balancing selection may be easier to detect in smaller populations.

Simulations

The theory developed above assumes that the frequencies of A_1 and A_2 remain constant at $\hat{p}_1$ and $\hat{p}_2$, respectively. This is true only when the population size is infinite. With a finite population size, allele frequencies fluctuate around their equilibrium values due to genetic drift. To investigate the effects of stochastic allele frequency fluctuation on the accuracy of our model predictions, we conducted simulations using `mbs` (Teshima and Innan 2009). Briefly, each simulation replicate contained two steps: (1) forward simulation to obtain allele frequency trajectories for the selected variants given the demographic history; (2) coalescent simulation for a sample of n alleles at a linked neutral site, conditioning on the trajectories obtained in step 1 (see Supplementary Text S.6 for more details). Because the theory does not depend on a specific selection model, we used an overdominance model whereby the fitnesses of the three genotypes A_1A_1, A_1A_2, and A_2A_2 are $1 - s_1$, 1, and $1 - s_2$, respectively. The equilibrium frequencies are $\hat{p}_1 = \frac{s_2}{s_1+s_2}$ and $\hat{p}_2 = \frac{s_1}{s_1+s_2}$.

To check the results presented in Figure 5, we let $s_1 = s_2 = s$, such that $\hat{p}_1 = \hat{p}_2 = 50\%$. To simulate the population expansion model in 5a, we assumed that $N_{e,1} = 20,000$ (the effective population size of the current epoch) and $N_{e,2} = 2,000$ (the effective population size of the ancestral epoch). For the population reduction model in 5b, we used $N_{e,1} = 2,000$ and $N_{e,2} = 20,000$.

As shown in Figure S6, the theoretical predictions are highly accurate. Here selection was strong, as measured by $\gamma_{min} = 2N_{e,min}s = 250$, where $N_{e,min} = \min(N_{e,1}, N_{e,2})$. To further check the robustness of our results, we reduced s, such that $\gamma_{min} = 20$. The substantial reduction in the intensity of selection leads to a significantly higher level of fluctuation in the frequencies of the selected variants (Figure S7). Encouragingly, the theoretical predictions remain accurate (Table S1).

A model of recent balanced polymorphism

We now turn our attention to the effects of the recent origin of a balanced polymorphism on patterns of genetic variability. Consider a diploid panmictic population with constant effective population size N_e. At an autosomal locus, a mutation from A_1 (the wild type) to A_2 (the mutant) arises. The fitnesses of the genotypes A_1A_1, A_1A_2, and A_2A_2 are $w_{11} = 1 - s_1$, $w_{12} = 1$, and $w_{22} = 1 - s_2$ ($s_1 > 0$ and $s_2 > 0$; i.e., there is heterozygote advantage). As above, we ignore reversible mutation between A_1 and A_2. In what follows, we first use a forward-in-time approach to obtain equations for describing the increase in the frequency of A_2 in the population. We then use the backward-in-time coalescent approach to calculate various measures of sequence variability in linked genomic regions. Wherever appropriate, we present results from a comparable selective sweep model, so that the two models can be compared.

Frequency of the mutant allele in the population

Let the frequencies of A_1 and A_2 in the current generation be p_1 and p_2, respectively. Let p_2' be the frequency of A_2 in the next generation. Using the standard theory (reviewed in Chap. 2 of Charlesworth and Charlesworth (2010)), the change in allele frequency in one generation due to selection is given by

$$\Delta p_2 = p_2' - p_2 = \frac{p_1p_2(w_{2\cdot} - w_{1\cdot})}{\bar{w}} \quad (21)$$

where $w_{1\cdot} = p_1w_{11} + p_2w_{12}$, $w_{2\cdot} = p_1w_{12} + p_2w_{22}$, and $\bar{w} = p_1w_{1\cdot} + p_2w_{2\cdot}$. Assuming that both $s_1 \ll 1$ and $s_2 \ll 1$, $\Delta p_2 \approx p_1p_2(w_{2\cdot} - w_{1\cdot}) = p_1p_2(p_1s_1 - p_2s_2)$. At equilibrium, $\Delta p_2 = 0$, such that the frequencies are $\hat{p}_1 = \frac{s_2}{s_1+s_2}$ and $\hat{p}_2 = \frac{s_1}{s_1+s_2}$.

When $p_2 \ll 1$, $\Delta p_2 \approx s_1p_2$. This is the same as when A_2 is under positive selection with fitnesses of the three genotypes being $w_{11} = 1$, $w_{12} = 1 + s_1$, and $w_{22} = 1 + 2s_1$, respectively (i.e., there is semi-dominance). Thus, we expect the initial signals generated by the increase in p_2 to be similar to those from an incomplete selective sweep, referred to here as the "corresponding sweep model".

The similarity between the two selection models means that we can borrow useful results from the selective sweep literature. In particular, after A_2 has been generated by mutation, its frequency must increase rapidly for it to escape stochastic loss. Following an approach first proposed by Maynard Smith (1976), we assume that p_2 increases instantly to $\epsilon = \frac{1}{\gamma_1}$, where $\gamma_1 = 2N_es_1$ (see also Desai and Fisher 2007). Thereafter, p_2 changes deterministically until its rate of change becomes very slow near the equilibrium point, when the coalescent process (considered in the next sub-section) is effectively the same as at equilibrium. Measuring time in units of $2N_e$ generation, $p_2(t)$ satisfies:

$$\frac{\mathrm{d}p_2}{\mathrm{d}t} = p_1p_2(p_1\gamma_1 - p_2\gamma_2) \quad (22)$$

where $\gamma_2 = 2N_es_2$. The solution to this differential equation is

$$\gamma_1 \ln(1 - p_2) + \gamma_2 \ln p_2 - (\gamma_1 + \gamma_2) \ln\left[\gamma_1 - (\gamma_1 + \gamma_2)p_2\right] = \gamma_1\gamma_2(t + c) \quad (23)$$

where c is a constant such that $p_2(0) = \epsilon$. We can obtain the frequency of A_2 at time t by solving for p_2 numerically.

It is instructive to compare the dynamics of $p_2(t)$ with those for the corresponding sweep model defined above. We assume that the frequency of the positively selected variant A_2 increases instantly to ϵ and grows deterministically until $1 - \epsilon$. Let $p_2^*(t)$ be the frequency of A_2 at scaled time t after its frequency arrived at ϵ. It can be shown that:

$$p_2^*(t) = \frac{\epsilon}{\epsilon + (1 - \epsilon)e^{-\gamma_1 t}} \quad (24)$$

(Crow and Kimura 1970; Stephan *et al.* 1992).

A recent study explicitly considered the stochastic phases when the frequency of the positively selected variant A_2 is below ϵ or greater than $1 - \epsilon$ (Charlesworth 2020a). These two phases contribute relatively little to the fixation time under the current model with strong selection and semi-dominance (see Table 1 of Charlesworth 2020a). Furthermore, when the frequency of A_2 is very close to 0 or 1, the coalescent process is effectively the same as under neutrality. Thus, ignoring these two stochastic phases is reasonable for our purposes.

In Figure 6, we display three balancing selection models, all with $\gamma_1 = 500$, but different γ_2 values, so that they have different equilibrium allele frequencies. For comparison, the corresponding sweep model with $\gamma_1 = 500$ is also presented. As can be seen, the allele frequency trajectories for the balancing selection models and the corresponding sweep model are similar only for a rather short period. After that, $p_2(t)$ increases at a much slower pace than $p_2^*(t)$. As shown below, these observations explain the differences between a recent balanced polymorphism and the spread of a beneficial mutation with respect to their effects on diversity patterns in nearby genomic regions.

Total branch length

We extend the coalescent approach developed above for the equilibrium model, in order to calculate the expected total branch length L for a random sample of n alleles at a linked neutral site (see (16)). The frequency of A_2 at the time of sampling is

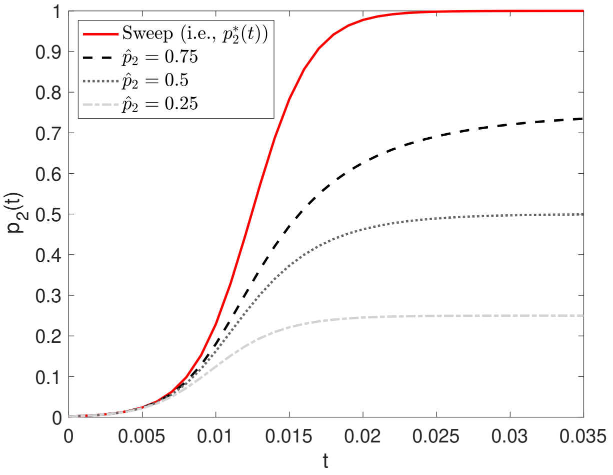


Figure 6 The frequency of the mutant allele A_2 as a function of t (time since its frequency reached ϵ). $\gamma_1 = 500$. γ_2 is adjusted such that the equilibrium frequency $\hat{p}_2$ is 0.25, 0.5, and 0.75, respectively. The trajectory under the corresponding sweep model is included for comparison.

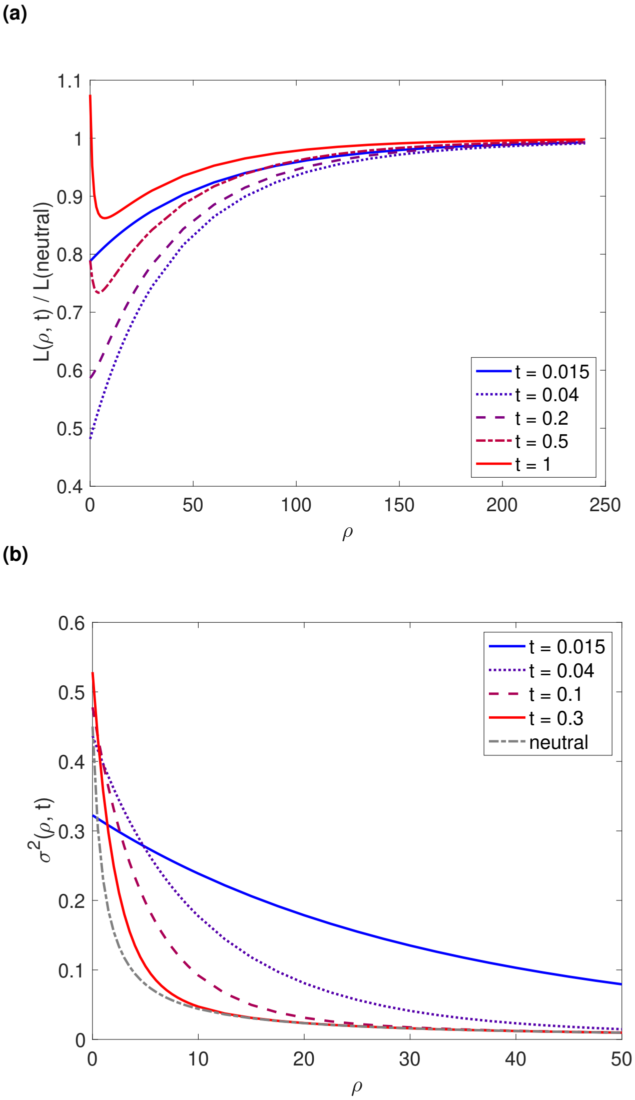
$p_2(t)$ where t is the time since the frequency of A_2 reached ϵ, expressed in units of $2N_e$ generations. At time τ before the present ($0 \leq \tau < t$), the frequency of A_2 is given by $p_2(t - \tau)$. For $\tau \geq t$, the process reduces to a standard neutral coalescent model with constant population size. To make use of Theorem 1, we divide $[p_2(t), \epsilon)$ into $H - 1$ equal-sized bins, such that the h-th bin is $[p_{2,h-1}, p_{2,h})$, where $p_{2,h} = p_2(t) + \frac{h}{H-1}(\epsilon - p_2(t))$ ($h \in \{0, 1, 2, ..., H-1\}$). Let τ_h be the solution to $p_2(t - \tau_h) = p_{2,h}$ given by (23). The corresponding time interval for bin h is $[\tau_{h-1}, \tau_h)$, which is shorter when the frequency of A_2 is changing at a faster rate. Thus, we have H epochs, with the first $H - 1$ in $[0, t)$ and epoch H covering the whole of $[t, \infty)$ (Figure S8).

Consider epoch h with $h < H$. The state space in this epoch is the same as that discussed above for the equilibrium model (see the arguments leading to (12)). Thus, the sub-intensity matrix for this epoch, $\boldsymbol{S}_h$, can be obtained in a similar way (cf., Figure S4). The only complication is that the frequency of A_2 changes within the epoch. However, if the time interval is sufficiently small, we can treat the frequency of A_2 as if it were constant. Here we set the frequency of A_2 in epoch h to its harmonic mean $q_{2,h}$, which can be calculated as:

$$\frac{1}{q_{2,h}} = \frac{1}{\tau_h - \tau_{h-1}} \int_{\tau_{h-1}}^{\tau_h} \frac{1}{p_2(t - \tau)} \mathrm{d}\tau. \tag{25}$$

We can then obtain $\boldsymbol{S}_h$ by simply replacing $\hat{p}_1$ and $\hat{p}_2$ in the sub-intensity matrix for the equilibrium model with $q_{1,h}$ and $q_{2,h}$, where $q_{1,h} = 1 - q_{2,h}$.

Note that, although the space state is the same for the epochs in $[0, t)$, this is not true for the transition from epoch $H - 1$ to epoch H. At the end of epoch $H - 1$, if more than one allele is associated with A_2, they coalesce into a single ancestral allele instantly. If the resulting ancestral allele is the only allele left, the process is terminated. Otherwise, if there are also n_1 alleles associated with A_1 at the time, then the $n_1 + 1$ alleles enter epoch H and coalesce at rate $\binom{n_1+1}{2}$. Thus, we need a mapping matrix $\boldsymbol{E}_{H-1,H}$, which is defined below (S22) in Supplementary Text S.5, to take into account the differences between the two epochs. For instance, for a sample of two alleles, the state space in $[0, t)$ has three transient states: (0, 2), (1, 1), and (2, 0), where the two numbers of each tuple represent the number of alleles associated with A_1 and A_2, respectively. However, epoch H has only one transient state, representing two uncoalesced alleles. If the process is in state (0, 2) at the end of $[0, t)$, it terminates with the instant coalescence of the two alleles. If the process is in any of the other two states, it enters epoch H with the same starting condition. Thus $\boldsymbol{E}_{H-1,H}^T = (0, 1, 1)$, where 0 in the first element means it is impossible to enter epoch H via state 1 in epoch $H - 1$, and the 1s mean that, if the process is in state 2 or 3 by the end of epoch $H - 1$, the process begins epoch H in state 1.


(a)

(b)

Figure 7 Nucleotide site diversity and LD in genomic regions surrounding a recently-emerged variant under balancing selection. The parameters are $\gamma_1 = 500$ and $\hat{p}_2 = 0.75$ (as in Figure 6). The discretisation scheme has $H = 76$ bins. In (a), the expected total branch length for a sample of $n = 2$ alleles is calculated for various value of t, the time since the frequency of A_2 reached ϵ. To make the effects more visible, L is divided by its neutral expectation. σ^2 in (b) measures the level of LD between the selected locus and a linked neutral site. For comparison, the neutral expectation of σ^2 is also included.

In all, the model has the following parameters: γ_1, γ_2, t, and ρ. By increasing the number of bins in the discretisation scheme (i.e., H; Figure S8), we can get arbitrarily accurate approximations. The results presented below are based on values of H such that the size of the frequency bins is about 1%. This is a rather conservative choice; using larger bins does not significantly change the results. Once the sub-intensity matrices are defined (i.e., $\boldsymbol{S}_h$ for $1 \leq h \leq H$), we can obtain $\boldsymbol{U}$ using Theorem 1 (see also Supplementary Text S.5) and $L = \boldsymbol{\alpha U D}$ (see (15)).

Figure 7a shows how neutral diversity levels are affected by a recent balanced polymorphism, using the balancing selection model with $\hat{p}_2 = 0.75$ considered in Figure 6. Initially, the rapid increase in the frequency of A_2 produces a drop in neutral diversity in nearby regions (the solid blue line). The maximum extent of reduction appears when $p_2(t)$ is close to its equilibrium value (the dotted line; $p_2(0.04) = 0.742$). After that, the diversity level starts to recover. Here, the increase in diversity level is fastest for regions closely linked to the selected site, because coalescence is slow when ρ is small. This leads to a U-shaped diversity pattern that persists for some time, which is followed by a rather slow approach to the equilibrium value (Figure S9). These dynamics are qualitatively the same when we consider a larger sample size with 20 alleles, although the reduction in diversity is less pronounced (Figure S10). Similar patterns are also observed for the other two balancing selection models in Figure 6 (Figure S11). The main difference is that models with a smaller $\hat{p}_2$ tend to result in a smaller reduction in neutral diversity. For instance, for the model with $\hat{p}_2 = 0.25$, the maximum reduction in nucleotide site diversity in very tightly linked regions is less than 6% (as opposed to a more than 50% reduction in Figure 7a), potentially making them very difficult to detect from data.

LD between the selected locus and a linked neutral site

It is straightforward to use the method developed in the previous subsection to calculate σ^2. From Figure 7b, we make two observations. First, LD builds up quickly and extends to a large genomic region when the frequency of A_2 is increasing rapidly (blue solid curve vs the neutral curve). This suggests the formation of long haplotypes around the selected locus, which can be used to help detect selection targets, as is done in extended haplotype tests (e.g., Voight *et al.* 2006; Ferrer-Admetlla *et al.* 2014). Second, the level of LD starts to decline before the reduction in diversity is maximal (the dotted curves in Figures 7a and b), suggesting that LD based detection methods will have already lost a substantial amount of their statistical power by this time. This implies that LD and diversity patterns complement each other when it comes to detecting targets of recent balancing selection.

Differences between balancing selection and selective sweeps in their effects on the total branch length and LD

We can analyse selective sweep models using the discretisation scheme outlined in Figure S8. In Figure 8a, we compare the balancing selection model shown in Figure 7 to its corresponding sweep model, with respect to their effects on L (the expected total branch length). Because the frequency of the beneficial allele increases much more rapidly (Figure 6), it causes a more pronounced reduction in diversity than the balanced polymorphism of the same age. Fixation of the beneficial allele occurs at $t = 0.025$. After that, diversity returns to its neutral level over a time period of the order of $2N_e$ generations, which is much faster than the time it takes for diversity to reach its equilibrium level under balancing selection (Figure S9). The patterns are similar when a larger sample size is considered (Figure S12).

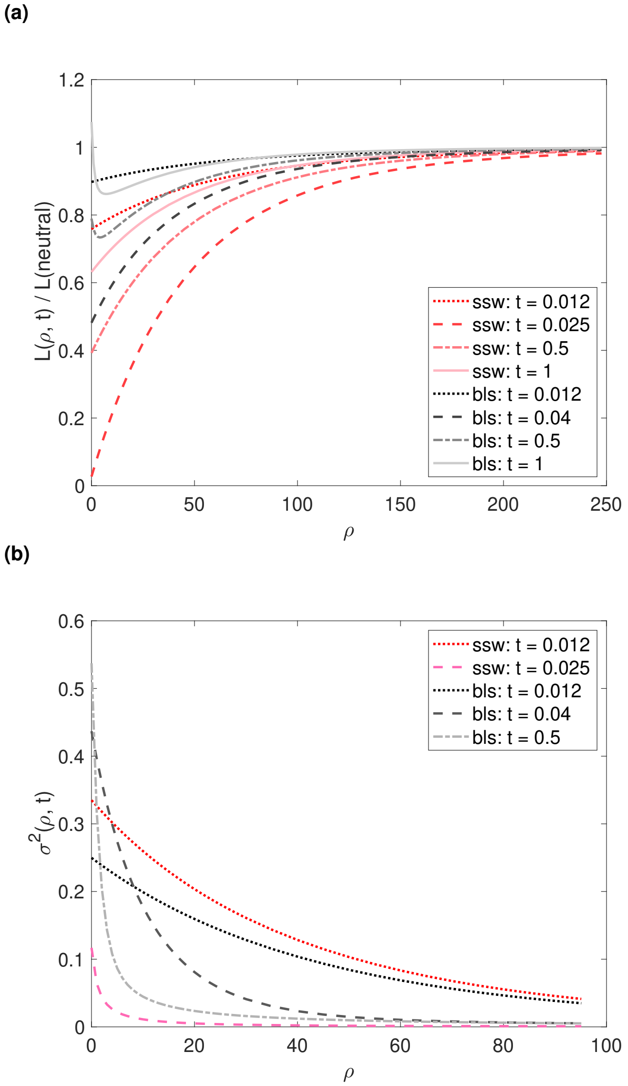


Figure 8 Comparing recent balancing selection with the corresponding sweep model, with respect to their effects on diversity and LD levels in surrounding genomic regions. The parameters of the balancing selection model (bls) are $\gamma_1 = 500$ and $\hat{p}_2 = 0.75$ (i.e., the same as in Figure 7). The corresponding sweep model (ssw) has $\gamma_1 = 500$. In (a), the expected total branch length for a sample of $n = 2$ alleles, divided by its neutral value, is presented. In (b), we consider the level of LD between the selected locus and a linked neutral site, as measured by σ^2. Fixation (taken as the time when the mutant allele frequency reaches $1 - \epsilon$) occurs at $t = 0.025$ under the sweep model. The reduction in diversity reaches its maximum at $t \approx 0.04$ under the balancing selection model.

A comparison between the two selection models with respect to their effects on LD patterns in the surrounding neutral region is shown in Figure 8b. Both models result in elevated LD. As expected, the corresponding sweep model leads to a more pronounced build-up of LD (red vs black dotted lines). This suggests that recent balancing selection is harder to detect than a comparable beneficial mutation. Under both models, LD starts to decay before the reduction in diversity is maximal (pink vs grey dashed lines). The decay appears to be much faster under the sweep model. This is because, under the balancing selection model, A_2 approaches an equilibrium frequency, instead of fixation. Therefore, a sizeable genomic region remains at elevated levels of LD with the selected locus for a longer period. Recall that diversity levels also take much longer to reach equilibrium under balancing selection (Figure 8a). Thus, there may well be a bigger window of opportunity for detecting targets of recent balancing selection, despite the fact that the signals they produce tend to be less dramatic than those produced by the corresponding sweep model.

The site frequency spectrum

The SFS can also be obtained using the time discretisation procedure. Here the state space is the same as that detailed for the equilibrium balancing selection model. As above, we obtain the sub-intensity matrix for epoch h by replacing $\hat{p}_1$ and $\hat{p}_2$ in the sub-intensity matrix for the equilibrium model (e.g., Supplementary Text S.4) with $q_{1,h}$ and $q_{2,h}$, respectively. We then use Theorem 1 to calculate $X_i(n_1, n_2)$. It is more instructive to consider the SFS for a sample of n randomly collected alleles, defined as:

$$X_i = \sum_{j=0}^{n} \binom{n}{j} p_1^j p_2^{n-j} X_i(j, n-j) \tag{26}$$

where p_1 and p_2 are the frequencies of A_1 and A_2 at the time of sampling. The effects selection has on the shape of the SFS are visualised using the ratio $X_i / X_i(\text{neutral})$, where $X_i(\text{neutral}) = 2\theta / i$.

In Figure 9, we present the SFS at different time points since the arrival of the mutant allele, for both the balancing selection model and the corresponding sweep model considered in Figure 8. When the frequency of the selected variant is rapidly increasing in the population, both types of selection produce a U-shaped SFS, with an excess of both low and high frequency derived variants. The extent of distortion is maximised around the time when the reduction in neutral diversity is also the most pronounced (see plots in the second row). The corresponding sweep model has a much bigger effect on the shape of the SFS. For example, under the sweep model, at the time of fixation ($t = 0.025$), $X_9/X_8 = 4.91$ and $X_1/X_2 = 8.05$. In contrast, when the SFS is most distorted under the balancing selection model ($t = 0.04$), $X_9/X_8 = 1.34$ and $X_1/X_2 = 3.29$. The excess of high frequency derived variants quickly disappears after the selected allele has stopped its rapid increase in frequency (plots in the third row), although the SFS remains U-shaped for longer under balancing selection. The plots in the last row shows the transition from a situation with reduced diversity and an excess of low frequency variants to a situation that resembles the pattern expected under long-term balancing selection, with an elevated diversity level and an excess of intermediate frequency variants. Qualitatively similar dynamics have been observed for the balancing selection models with $\hat{p}_2 = 0.5$ and 0.25, respectively (Figure S13). Again, the SFS-distorting effect is weaker when $\hat{p}_2$ is smaller, with the case with $\hat{p}_2 = 0.25$ producing hardly any excess of low and high frequency variants even when A_2 is increasing in frequency.

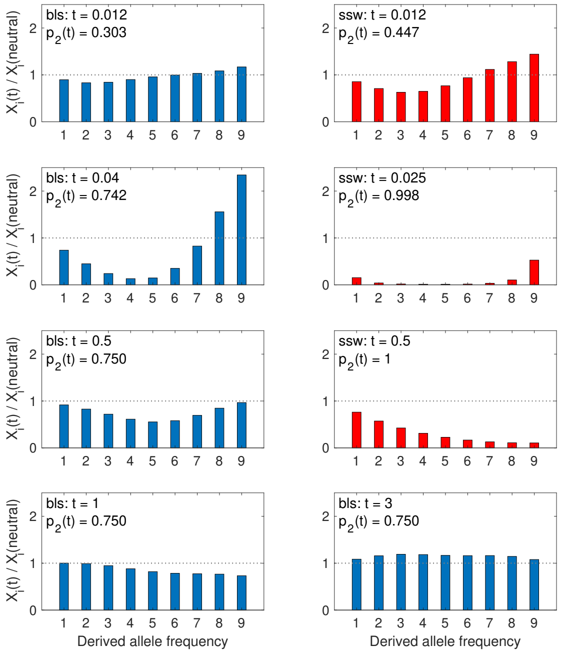


Figure 9 The SFS at various time points after the arrival of the selected variant for a random sample of 10 alleles. The balancing selection (bls) and selective sweep (ssw) models are the same as those shown in Figure 8. The scaled recombination frequency between the focal neutral site and the selected site is $\rho = 2$. The reduction in diversity reaches its maximum at $t \approx 0.04$ and 0.025 (fixation) under the balancing selection and selective sweep models, respectively. The SFS under selection is expressed relative to its neutral expectation.

To investigate the SFS further, we consider π (the nucleotide site diversity) and Watterson's θ_W. Recall that, under the infinite sites model, $\pi = 2\theta T$, where T is defined by (9). Let $\$$ be the expected number of segregating sites in a sample of size n. We have $\$ = \theta L$. Because $\theta_W = \$/a_n$ where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$, we have $\theta_W = \theta L / a_n$. Following Becher *et al.* (2020), we define

$$\Delta\theta_W = 1 - \frac{\pi}{\theta_W} = 1 - \frac{2\theta T}{\theta L / a_n} = 1 - \frac{2 a_n T}{L}. \tag{27}$$

$\Delta\theta_W = 0$ under neutrality, > 0 when there is an excess of rare variants, and < 0 when there is an excess of intermediate frequency variants.

Figure 10 shows $\Delta\theta_W$ for the balancing selection model with $\gamma_1 = 500$ and $\hat{p}_2 = 0.75$ (as in Figures 6 - 9); the corresponding sweep model is also included for comparison. At $t = 0.012$, the balancing selection model produces no obvious deviation from neutrality (black dotted line), whereas the sweep model has already started to cause a significant excess of rare variants (red dotted line). This is consistent with the much slower increase in the frequency of A_2 under balancing selection ($p_2(0.012) = 0.303$ vs $p_2^*(0.012) = 0.447$). The extent of deviation caused by the sweep is maximal around the time when A_2 becomes fixed ($t \approx 0.025$; pink dashed line). Under the balancing selection model, the maximum deviation appears when the frequency of A_2 becomes close to its equilibrium value ($t \approx 0.04$; grey dashed line), but is less pronounced than under the sweep model. After the maximum is achieved, diversity patterns gradually return to neutrality over $4N_e$ generations under the sweep model. For the balancing selection model, there is a much longer period of non-stationary dynamics as shown by the light blue and blue

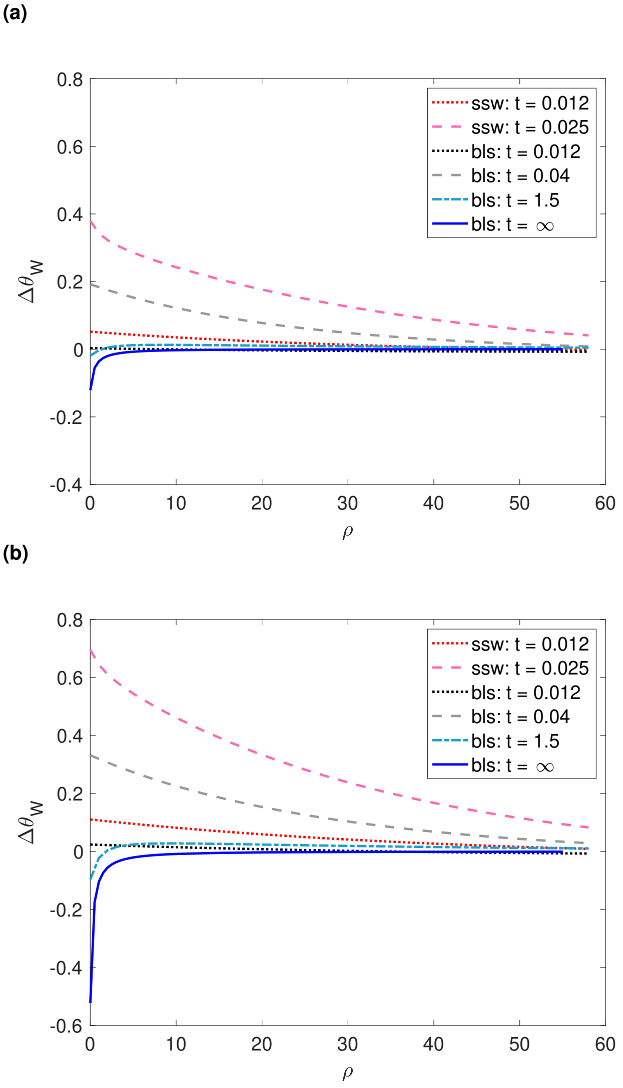


Figure 10 $\Delta\theta_W$ as a function of ρ and t. The two selection models are the same as those considered in Figure 9. "bls: $t = \infty$" corresponds to the equilibrium under balancing selection. The sample size is 10 in (a) and 35 in (b).

lines. The observations are qualitatively similar for the two sample sizes considered ($n = 10$ vs $n = 35$). Nonetheless, the extent of deviation in the SFS is more conspicuous when $n = 35$, suggesting an increase in statistical power.

It is informative to compare the three balancing selection models with $\gamma_1 = 500$, but different equilibrium allele frequencies (Figure 6). The model with $\hat{p}_2 = 0.75$ produces the strongest sweep-like signals (Figure 10 vs Figure S14). At the other extreme, the model with $\hat{p}_2 = 0.25$ effectively emits no such signal (Figure S14). Thus, targets of recent balancing selection with larger $\hat{p}_2$ are easier to detect. However, for older targets of selection, the excess of intermediate frequency variant (i.e., negative $\Delta\theta_W$) is most noticeable for selection targets with $\hat{p}_2 \approx 0.5$ (Figure S14), making them the most amenable to detection. Altogether, it seems that balancing selection targets with low equilibrium allele frequencies (e.g., $\hat{p}_2 \approx 0.25$) are difficult to identify regardless of their age.

Simulations

We performed simulations with stochastic allele frequency trajectories at the selected site using `mbs`. The simulation method is similar to that described earlier (see also Supplement Text S.6). In Figure S15, $\gamma_1 = 500$ and the equilibrium frequency of A_2 is 0.75 (i.e., the same as Figure 9). The theoretical predictions for both the balancing selection and selective sweep models are highly accurate. In an additional experiment, we reduced γ_1 to 20, but kept the equilibrium frequency of A_2 at 0.75. This is to examine the robustness of our predictions against increased stochasticity induced by weaker selection. The results in Figure S16 suggest that our theory remains accurate for both models.

Discussion

In this study, we have used the power and flexibility afforded by phase-type theory to study the effects of balancing selection on patterns of genetic variability and LD in nearby genomic regions. Our results go beyond previous attempts in that they provide a unifying framework for calculating important statistics for both equilibrium and nonequilibrium cases. In what follows, we discuss how our results can be used in data analyses and future method developments. We will also discuss the usefulness of phase-type theory in general.

Accommodating other biological factors

Here we have only considered selection on an autosomal locus in a randomly mating population. However, our results can be readily extended to accommodate other important biological factors. Take self-fertilization as an example. Let f be the selfing rate and $F = f/(2-f)$ be the corresponding inbreeding coefficient. For this model, $N_e = N/(1+F)$, where N is the number of breeding individuals (Charlesworth 2009). Because selfing increases the frequency of homozygotes in the population, it reduces the effective frequency of recombination to $r_e = (1-F)r$, where r is the autosomal recombination rate in a random-mating population (Nordborg 1997; see Hartfield and Bataillon 2020 for a more accurate expression for r_e). Finally, for the model of recent balancing selection, we also need to consider the effects of selfing on the frequency trajectory of A_2. This can be achieved by replacing (22) with:

$$\frac{\mathrm{d}p_2}{\mathrm{d}t} = p_1 p_2 \left[(1-F)(p_1\gamma_1 - p_2\gamma_2) + F(\gamma_1 - \gamma_2)\right]. \qquad (28)$$

Other factors, including separate sexes, mode of inheritance (e.g., X-linkage vs autosomal), and background selection, can also be modelled (Charlesworth 2009; Vicoso and Charlesworth 2009; Glémin 2012; Charlesworth 2020a; Hartfield and Bataillon 2020).

Detecting long-term balancing selection

We have examined two models of long-term balancing selection, one with a constant population size and the other with recent demographic changes. We confirm the well-known result that long-term balancing selection leads to elevated diversity, increased LD, and an excess of intermediate frequency variants in the SFS (Figures 2 - 4, 10; Charlesworth 2006; Fijarczyk and Babik 2015). Because the strength of these signals is weak except at sites very close to the locus under selection, they could be useful in pinpointing targets of balancing selection. On the other hand, we find that, under our two-allele model, these signals are strongest when the equilibrium frequencies of the selected variants are close to 50% (Figures 2 - 4, 10, and S14).

This implies that genome scan methods are likely to be biased towards detecting selection targets where the selected variants are more common, which appears to be the case for some detection methods (Bitarello *et al.* 2018; Siewert and Voight 2020).

Our results can be used to improve existing methods for detecting balancing selection. For example, the T_1 test by DeGiorgio *et al.* (2014), which has been shown to be among the most powerful, is based on L, the expected total branch length. The recursion equations DeGiorgio *et al.* (2014) used to obtain L assumes a constant population size. We can now relax this assumption by incorporating changes in population size. The increase in the strength of signals of long-term balancing selection after population size reduction (Figure 5b) points to the importance of incorporating non-equilibrium demographic dynamics, which may help to increase statistical power and reduce false positive rates. Nonetheless, the results presented in Figures 4 and 10 show that L does not capture all of the information about balancing selection. Instead, statistical power can be gained by making use of the SFS. This explains why the T_1 test (based on L) is often less powerful than the T_2 test (based on the SFS) (DeGiorgio *et al.* 2014). However, DeGiorgio *et al.* (2014) obtained the SFS via stochastic simulations, due to a lack of analytical methods. Here we have filled this gap. As above, it is of importance to extend the T_2 test, so that it includes both the equilibrium and non-equilibrium models.

Detecting recent balancing selection

It has long been suggested that signals generated by recent balancing selection should be similar to those generated by incomplete sweeps (Charlesworth 2006; Fijarczyk and Babik 2015). However, the allele frequency trajectories under these two models are similar only when the mutant allele is rather rare in the population (Figure 6). This period accounts for a small fraction of the time it takes to fix a positively selected mutation subject to a comparable level of selection. In addition, the rate of allele frequency change in this period is slower than when the mutant allele is more common. Combining these two factors, it is unsurprising that, at the time when the allele frequency trajectories under the two models start to diverge, neither model produces a noticeable effect on diversity patterns in nearby genomic regions (data not shown). Thus, this initial period of identity contributes very little signal.

After the initial period, the frequency of the positively selected mutation increases rapidly. In contrast, the rate of growth under the balancing selection model is much slower, especially when the equilibrium frequency of the mutant allele is low (Figure 6). Nonetheless, the increase in frequency of a recent balanced polymorphism does produce sweep-like diversity patterns. These include reductions in genetic variability, a skew towards high and low frequency derived variants in the SFS, and a build-up of LD between the selected and linked neutral sites (Figures 7 - 10). In addition, the maximum build-up of LD appears before the reduction in diversity levels and the distortion of the SFS peak, suggesting that these signals complement each other. Although these patterns are not as pronounced as those produced by sweeps of a comparable strength, we expect them to be detectable by methods designed for identifying sweeps (Booker *et al.* 2017; Pavlidis and Alachiotis 2017), as has been shown previously (Zeng *et al.* 2006). An open question is whether it is possible to distinguish between these two types of selection. On the other hand, because recent balancing selection causes diversity and LD patterns to be in a non-equilibrium state for a long period (Figures 10 and S14), it is unclear whether these patterns can be exploited for detecting selection targets.

Comparing the three balancing selection models with equilibrium allele frequencies $\hat{p}_2 = 0.25$, 0.5, and 0.75, respectively (Figure 6), mutations with $\hat{p}_2 = 0.75$ produce the strongest sweep-like patterns (e.g., Figure 9 vs Figure S13). They are probably the easiest to detect, although they may also be the most difficult to be distinguished from sweeps. On the other hand, although selection targets with $\hat{p}_2 = 0.5$ are not as easy to detect when they are young, they produce the strongest deviation from neutrality if they have been maintained for a sufficiently long period of time (Figures 2, 3, and S14), suggesting that they are most likely to be identified by methods for detecting long-term selection targets. Finally, it seems that selection targets with $\hat{p}_2 = 0.25$ are the most difficult to detect regardless of the age of the mutant allele.

Using phase-type theory to assess the accuracy of simpler approximations

We have shown the ease for which phase-type theory can be used to analyse complex models. In some cases, this can lead to simple analytic solutions (e.g., (7) and (8)). When explicit analytic solutions are difficult to obtain, phase-type theory can be useful in searching for simpler approximations. Take the model of recent balancing selection as an example. By using a large number of bins in the discretisation scheme (Figure S8), we can obtain results that are effectively exact. It is, however, impossible to write them as simple equations. Nonetheless, if we make an additional assumption that the recombination frequency between the selected locus and the neutral locus is not too high relative to the strength of selection, we can adopt the methods developed in Charlesworth (2020b) for selective sweeps, such that they can be used to obtain the expected pairwise coalescence time (see Supplementary Text S.8 for details).

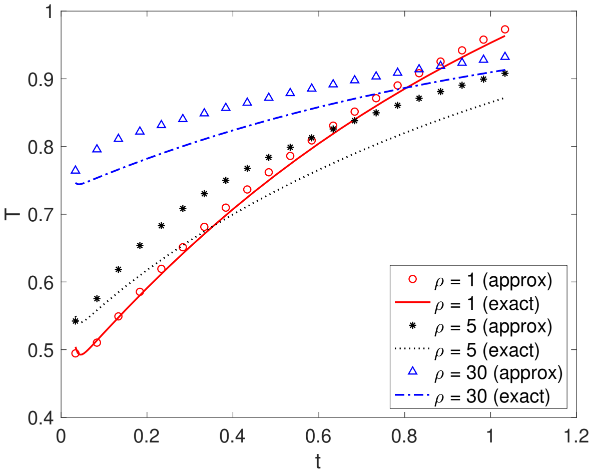


Figure 11 Comparing expected pairwise coalescence times obtained by phase-type theory (exact) and an approximation assuming low recombination rates. The model of recent balancing selection model has the following parameters: $\gamma_1 = 500$ and $\hat{p}_2 = 0.75$ (i.e., the same as in Figures 7 - 10). t is the time since the arrival of A_2. The discretisation scheme has $H = 76$ epochs. Details of the approximation are given in Supplementary Text S.8.

We can assess the reliability of this approximation by comparing its results with those obtained using the phase-type method. As expected, the approximate results match the exact results closely when the recombination rate is low (e.g., $\rho = 1$ in Figure 11). For higher recombination rates, the approximation underestimates the diversity-reducing effect of the spread of A_2. The main reason for this discrepancy is that the approximation assumes that the recombination rate is low, and the "sweep phase" is short. When these assumptions hold, once recombination during the sweep phase has moved a lineage from allelic class 2 to allelic class 1, back migration to allelic class 2 can be ignored. Although these assumptions work well for selective sweep models (Charlesworth 2020b), they are less suitable for the model of recent balancing selection, because the increase in allele frequency is much slower, leading to a longer sweep phase, and hence more opportunities for recombination. Thus, by preventing lineages from being moved back into allelic class 2, the approximation artificially slows down the rate of coalescence during the sweep phase, explaining the overestimation of pairwise coalescence time. Using results produced by phase-type theory as the baseline is desirable because, unlike stochastic simulations, these results are analytical, making comparisons straightforward and small differences easier to detect.

Differences from previous studies and limitations

The equilibrium model of balancing selection has been analysed previously using coalescent theory (Hudson and Kaplan 1988; Nordborg 1997). Phase-type theory has allowed us to reproduce well known results (e.g., (8)). Additionally, it has made it feasible to obtain other important summary statistics (e.g., total branch length, LD and SFS) and introduce non-equilibrium scenarios (changes in population size or recent selection). Recently, Kern and Hey (2017) analysed a coalescent model with isolation and migration. Although the authors did not consider selection, the approach they used is related in that it involves performing calculations directly using the underlying continuous time Markov process. However, the results derived using our formulation is more compact (e.g., Theorem 1), which facilitates the accommodation of more complex situations (e.g., recent selection). Furthermore, we are able to obtain other useful results such as the second moment of the mean time to MRCA (Theorem 2 in Supplementary Text S.7).

A limitation of the phase-type approach is that the size of the state space increases quickly with the sample size, meaning that the computational cost will become too high for large samples. However, there is evidence that samples with as few as 20 alleles, which is computationally feasible using our approach, offer sufficient statistical power for detecting balancing selection (Siewert and Voight 2017; Bitarello *et al.* 2018). More importantly, our method provides a way of analysing complex models, which will help us to understand their properties. This may in turn enable us to obtain computationally more efficient approximations, as shown in the previous section. Finally, although the speed of forward simulators has improved significantly (Haller and Messer 2019), the phase-type approach is still much faster for moderate sample sizes. This is because, for a given set of parameters, we only need to perform the calculation once to obtain, for instance, the expected total branch length. In contrast, obtaining this quantity accurately using simulations requires at least tens of thousands of replicates. Simulations are, however, highly flexible and can be used to study models that are too difficult to analyse mathematically. Thus, both mathematical modelling and simulations are important.

Applying phase-type theory to other population genetic models

Phase-type theory can be applied to many different models in population genetics. For example, Hobolth *et al.* (2019) used a time-homogeneous version of the theory to study the standard Kingman's coalescent with and without recombination, coalescent models with multiple mergers, and coalescent models with seed banks. They showed the ease for which useful results can be obtained (e.g., all the moments of the pairwise coalescence time, the covariance in coalescence times between two linked loci, or the SFS). By extending the framework to non-equilibrium cases (see Theorem 1, Corollary 1 in Supplementary Text S.5, and Theorem 2 in Supplementary Text S.7), we make this approach applicable to a yet larger class of models. For instance, we can introduce population size fluctuations into the models considered by Hobolth *et al.* (2019). Even for models that have been analysed before using other approaches (e.g., Matuszewski *et al.* 2017), it is worth exploring whether the new theory provides a better alternative, both in terms of ease of analysis and numerical stability of the resulting method, which may be beneficial for parameter estimation purposes (e.g., Kern and Hey 2017).

The phase-type approach may be particularly useful for models that involve selection on a single locus at which the frequencies of the selected variants change deterministically (Maynard Smith and Haigh 1974; Kaplan *et al.* 1988; Coop and Ralph 2012). These include the balancing selection models considered here, selective sweep models (Barton 1998; Kim and Stephan 2002; Kim and Nielsen 2004; Ewing *et al.* 2010; Charlesworth 2020a; Hartfield and Bataillon 2020), soft sweeps caused by recurrent mutation or migration (Pennings and Hermisson 2006), incomplete sweeps (Vy and Kim 2015), and recurrent sweeps (Kaplan *et al.* 1989; Kim 2006; Campos and Charlesworth 2019).

Here, we have briefly considered selective sweep models with semi-dominance and compared it to the corresponding balancing selection model (see (24) and Figures 6, 8 - 10). In a related study, we will use the phase-type approach to investigate some of the sweep models listed above more systematically (K. Zeng and B. Charlesworth, *in preparation*). Because we can use phase-type theory to obtain exact solutions, it provides a convenient way to determine the accuracy of existing approximations. For instance, for the sweep model with semi-dominance, a widely-used approximation assumes that there is no coalescence during the sweep phase, such that the gene tree for a set of alleles sampled immediately after a sweep has a simple "star shape" (Maynard Smith and Haigh 1974; Barton 2000; Durrett and Schweinsberg 2004). However, a recent study of the pairwise coalescence time suggests that this approximation can be rather inaccurate when the ratio of the recombination rate to the selection coefficient is high (Charlesworth 2020b). It is important to also assess the effect of this simplifying assumption on the SFS, given that both nucleotide site diversity and the SFS are informative when it comes to estimating the strength and prevalence of (recurrent) sweeps (Corbett-Detig *et al.* 2015; Elyashiv *et al.* 2016; Booker *et al.* 2017; Comeron 2017). In addition, we can also explore the joint effects of recurrent sweeps and recent population size changes. These are not well understood, but are important for estimating the relative importance of background selection and recurrent sweeps in shaping genome-wide patterns of variability (e.g., Johri *et al.* 2020).

Data availability

The methods presented in this paper have been implemented in an R package named bls, which is available from http://zeng-lab.group.shef.ac.uk. In addition to the models considered here, the package can also obtain the total branch length and the SFS for (1) neutral models with changes in population size, (2) neutral models with two demes and changes in migration rates and/or deme sizes, and (3) isolation with migration models. Supplementary Material available at https://doi.org/10.25386/genetics.14186819.

Acknowledgements

We thank two anonymous reviewers for their comments on an earlier version of this paper, which led to considerable improvements.

Literature Cited

Al-Mohy, A. H. and N. J. Higham, 2010 A new scaling and squaring algorithm for the matrix exponential. SIAM J. Matrix Anal. Appl. **31**: 970–989.

Andersen, L. N., T. Mailund, and A. Hobolth, 2014 Efficient computation in the IM model. J. Math. Biol. **68**: 1423–51.

Andres, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, *et al.*, 2009 Targets of balancing selection in the human genome. Mol. Biol. Evol. **26**: 2755–2764.

Bakker, E. G., C. Toomajian, M. Kreitman, and J. Bergelson, 2006 A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell **18**: 1803–1818.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72**: 123–133.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. B **355**: 1553–1562.

Becher, H., B. C. Jackson, and B. Charlesworth, 2020 Patterns of genetic variability in genomic regions with low rates of recombination. Curr. Biol. **30**: 94–100 e3.

Bitarello, B. D., C. de Filippo, J. C. Teixeira, J. M. Schmidt, P. Kleinert, *et al.*, 2018 Signatures of long-term balancing selection in human genomes. Genome Biol. Evol. **10**: 939–955.

Bladt, M. and B. F. Nielsen, 2017 *Matrix-exponential distributions in applied probability*. Springer, New York.

Booker, T. R., B. C. Jackson, and P. D. Keightley, 2017 Detecting positive selection in the genome. BMC Biol. **15**: 98.

Campos, J. L. and B. Charlesworth, 2019 The effects on neutral variability of recurrent selective sweeps and background selection. Genetics **212**: 287–303.

Castric, V. and X. Vekemans, 2004 Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Mol. Ecol. **13**: 2873–2889.

Charlesworth, B., 2009 Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. **10**: 195–205.

Charlesworth, B., 2020a How long does it take to fix a favorable mutation, and why should we care? Am. Nat. **195**: 753–771.

Charlesworth, B., 2020b How good are predictions of the effects of selective sweeps on levels of neutral diversity? Genetics **216**: 1217–1238.

Charlesworth, B. and D. Charlesworth, 2010 *Elements of Evolutionary Genetics*. Roberts and Company Publishers, Greenwood Village (Colorado).

Charlesworth, B., M. Nordborg, and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70**: 155–174.

Charlesworth, D., 2006 Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. **2**: 379–384.

Cheng, X. H. and M. DeGiorgio, 2019 Detection of shared balancing selection in the absence of trans-species polymorphism. Mol. Biol. Evol. **36**: 177–199.

Comeron, J. M., 2017 Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Philos. Trans. R. Soc. B **372**: 20160471.

Connallon, T. and A. G. Clark, 2014 Balancing selection in species with separate sexes: Insights from Fisher's geometric model. Genetics **197**: 991–1006.

Coop, G. and P. Ralph, 2012 Patterns of neutral diversity under general models of selective sweeps. Genetics **192**: 205–224.

Corbett-Detig, R. B. and D. L. Hartl, 2012 Population genomics of inversion polymorphisms in *Drosophila melanogaster*. PLoS Genet **8**: e1003056.

Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton, 2015 Natural selection constrains neutral diversity across a wide range of species. PLoS Biol **13**: e1002112.

Crow, J. F. and M. Kimura, 1970 *An introduction to population genetics theory*. Harper & Row Publishers, New York.

DeGiorgio, M., K. E. Lohmueller, and R. Nielsen, 2014 A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. **10**: e1004561.

Desai, M. M. and D. S. Fisher, 2007 Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics **176**: 1759–98.

Dobzhansky, T., 1970 *Genetics of the evolutionary process*. Columbia University Press, New York.

Durrett, R. and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. **66**: 129–138.

Eanes, W. F., 1999 Analysis of selection on enzyme polymorphisms. Annu. Rev. Ecol. Evol. Syst. **30**: 301–326.

Elyashiv, E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, *et al.*, 2016 A genomic map of the effects of linked selection in *Drosophila*. PLoS Genet. **12**: e1006130.

Ewing, G., J. Hermisson, P. Pfaffelhuber, and J. Rudolf, 2010 Selective sweeps for recessive alleles and for other modes of dominance. J. Math. Biol. **63**: 399–431.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. **31**: 1275–1291.

Fijarczyk, A. and W. Babik, 2015 Detecting balancing selection in genomes: limits and prospects. Mol. Ecol. **24**: 3529–3545.

Fisher, R. A., 1922 On the dominance ratio. Proc. R. Soc. Edinb. **42**: 321–341.

Gilbert, K. J., F. Pouyet, L. Excoffier, and S. Peischl, 2020 Transition from background selection to associative overdominance promotes diversity in regions of low recombination. Curr. Biol. **30**: 101–107 e3.

Glémin, S., 2012 Extinction and fixation times with dominance and inbreeding. Theor. Popul. Biol. **81**: 310–316.

Haller, B. C. and P. W. Messer, 2019 SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol. Biol. Evol. **36**: 632–637.

Hartfield, M. and T. Bataillon, 2020 Selective sweeps under dominance and inbreeding. G3 **10**: 1063–1075.

Hedrick, P. W., 2011 Population genetics of malaria resistance in humans. Heredity **107**: 283–304.

Hobolth, A., A. Siri-Jegousse, and M. Bladt, 2019 Phase-type distributions in population genetics. Theor. Popul. Biol. **127**: 16–32.

Hudson, R. R. and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120**: 831–840.

Hudson, R. R., M. Kreitman, and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116**: 153–159.

Innan, H. and M. Nordborg, 2003 The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. Genetics **165**: 437–444.

Johnston, S. E., J. Gratten, C. Berenos, J. G. Pilkington, T. H. Clutton-Brock, *et al.*, 2013 Life history trade-offs at a single locus maintain sexually selected genetic variation. Nature **502**: 93–95.

Johri, P., B. Charlesworth, and J. D. Jensen, 2020 Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics **215**: 173–192.

Kaplan, N. L., T. Darden, and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120**: 819–829.

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123**: 887–899.

Karlin, S. and H. E. Taylor, 1981 *A second course in stochastic processes*. Academic Press, New York.

Kern, A. D. and J. Hey, 2017 Exact calculation of the joint allele frequency spectrum for isolation with migration models. Genetics **207**: 241–253.

Kim, K. W., B. C. Jackson, H. Zhang, D. P. L. Toews, S. A. Taylor, *et al.*, 2019 Genetics and evidence for balancing selection of a sex-linked colour polymorphism in a songbird. Nat. Commun. **10**: 1852.

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics **172**: 1967–1978.

Kim, Y. and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167**: 1513–1524.

Kim, Y. and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160**: 765–777.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.

Küpper, C., M. Stocks, J. E. Risse, N. Dos Remedios, L. L. Farrell, *et al.*, 2016 A supergene determines highly divergent male reproductive morphs in the ruff. Nat. Genet. **48**: 79–83.

Kwiatkowski, D. P., 2005 How malaria has affected the human genome and what hum. genet. can teach us about malaria. Am. J. Hum. Genet. **77**: 171–192.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Segurel, A. Auton, *et al.*, 2013 Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science **339**: 1578–1582.

Matuszewski, S., M. E. Hildebrandt, G. Achaz, and J. D. Jensen, 2017 Coalescent processes with skewed offspring distributions and nonequilibrium demography. Genetics **208**: 323–338.

Maynard Smith, J., 1976 What determines the rate of evolution? Am. Nat. **110**: 331–338.

Maynard Smith, J. and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23**: 23–35.

McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162**: 987–991.

Moler, C. and C. Van Loan, 2003 Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. **45**: 3–49.

Nagylaki, T., 1980 The strong-migration limit in geographically structured populations. J. Math. Biol. **9**: 101–114.

Nicolaisen, L. E. and M. M. Desai, 2013 Distortions in genealogies due to purifying selection and recombination. Genetics **195**: 221–230.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146**: 1501–1514.

Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. B **263**: 1033–1039.

Ohta, T. and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68**: 571.

Pavlidis, P. and N. Alachiotis, 2017 A survey of methods and tools to detect recent and strong positive selection. J. Biol. Res. Thessaloniki **24**.

Pennings, P. S. and J. Hermisson, 2006 Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. **23**: 1076–1084.

Polanski, A. and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165**: 427–436.

Ross, S. M., 1996 *Stochastic processes*. John Wiley & Sons, New York, second edition.

Sellis, D., B. J. Callahan, D. A. Petrov, and P. W. Messer, 2011 Heterozygote advantage as a natural consequence of adaptation in diploids. Proc. Natl. Acad. Sci. U.S.A. **108**: 20666–20671.

Siewert, K. M. and B. F. Voight, 2017 Detecting long-term balancing selection using allele frequency correlation. Mol. Biol. Evol. **34**: 2996–3005.

Siewert, K. M. and B. F. Voight, 2020 Betascan2: Standardized statistics to detect balancing selection utilizing substitution data. Genome Biol. Evol. **12**: 3873–3877.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58**: 167–175.

Spurgin, L. G. and D. S. Richardson, 2010 How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. R. Soc. B **277**: 979–988.

Stephan, W., T. H. E. Wiehe, and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism - analytical results based on diffusion-theory. Theor. Popul. Biol. **41**: 237–254.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103**: 545–555.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–595.

Takahata, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. U.S.A. **87**: 2419–2423.

Takahata, N. and M. Nei, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124**: 967–978.

Takahata, N. and Y. Satta, 1998 Footprints of intragenic recombination at HLA loci. Immunogenetics **47**: 430–441.

Teshima, K. M. and H. Innan, 2009 mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. BMC Bioinf. **10**: 166.

van Diepen, L. T., Å. Olson, K. Ihrmark, J. Stenlid, and T. Y. James, 2013 Extensive trans-specific polymorphism at the mating type locus of the root decay fungus heterobasidion. Mol. Biol. Evol. **30**: 2286–2301.

Vekemans, X. and M. Slatkin, 1994 Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137**: 1157–65.

Vicoso, B. and B. Charlesworth, 2009 Effective population size and the faster-X effect: An extended model. Evolution **63**: 2413–2426.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4**.

Vy, H. M. T. and Y. Kim, 2015 A composite-likelihood method for detecting incomplete selective sweep from population genomic data. Genetics **200**: 633–649.

Waltoft, B. L. and A. Hobolth, 2018 Non-parametric estimation of population size changes from the site frequency spectrum. Stat. Appl. Genet. Mol. Biol. **17**.

Zeng, K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity **110**: 363–371.

Zeng, K. and P. Corcoran, 2015 The effects of background and interference selection on patterns of genetic variation in subdivided populations. Genetics **201**: 1539–54.

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174**: 1431–1439.