The Effects of Background and Interference Selection on Patterns of Genetic Variation in Subdivided Populations

Kai Zeng¹ and Pádraic Corcoran

Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

ABSTRACT It is well known that most new mutations that affect fitness exert deleterious effects and that natural populations are often composed of subpopulations (demes) connected by gene flow. To gain a better understanding of the joint effects of purifying selection and population structure, we focus on a scenario where an ancestral population splits into multiple demes and study neutral diversity patterns in regions linked to selected sites. In the background selection regime of strong selection, we first derive analytic equations for pairwise coalescent times and F_{ST} as a function of time after the ancestral population splits into two demes and then construct a flexible coalescent simulator that can generate samples under complex models such as those involving multiple demes or nonconservative migration. We have carried out extensive forward simulations to show that the new methods can accurately predict diversity patterns both in the nonequilibrium phase following the split of the ancestral population and in the equilibrium between mutation, migration, drift, and selection. In the interference selection regime of many tightly linked selected sites, forward simulations provide evidence that neutral diversity patterns obtained from both the nonequilibrium and equilibrium phases may be virtually indistinguishable for models that have identical variance in fitness, but are nonetheless different with respect to the number of selected sites and the strength of purifying selection. This equivalence in neutral diversity patterns suggests that data collected from subdivided populations may have limited power for differentiating among the selective pressures to which closely linked selected sites are subject.

KEYWORDS background selection; interference selection; population subdivision; coalescent; F_{ST}

It was pointed out a long time ago that most mutations that affect the phenotype must be deleterious, because the biological system has been under selection for billions of years to improve its performance (e.g., Muller 1950). Evidence collected from both mutation accumulation experiments and analyses of large-scale DNA sequence variation is consistent with the prevalence of deleterious mutations and the relative rarity of advantageous ones (reviewed by Eyre-Walker and Keightley 2007; Halligan and Keightley 2009; Charlesworth 2012a). It is therefore unsurprising that the effects of selection against deleterious mutations are an integral part in the study of many fundamental questions in evolutionary genetics, such as the evolution of sex and recombination (Otto and

Lenormand 2002), the consequences of inbreeding and self-fertilization (Charlesworth and Charlesworth 1987), and the origin and nature of genetic variation underlying diseases (Wright *et al.* 2003).

An important line of research is to understand the effects of selection against recurrent deleterious mutations on patterns of genetic diversity at linked neutral sites. This has been extensively studied in the *background selection regime* (Charlesworth *et al.* 1993) in which deleterious mutations are assumed to be sufficiently strongly selected against that their frequencies stay very close to the deterministic limit expected in an infinitely large population (Kimura and Maruyama 1966; Haigh 1978). It is well established that, in a panmictic population, background selection causes a reduction in diversity and an excess of low-frequency variants in linked neutral regions relative to neutral expectations (Charlesworth *et al.* 1995; Hudson and Kaplan 1995; Nordborg *et al.* 1996; Zeng and Charlesworth 2011; Nicolaisen and Desai 2012, 2013), regardless of whether the population is in equilibrium or has experienced recent changes in population

Copyright © 2015 by the Genetics Society of America

doi: 10.1534/genetics.115.178558

Manuscript received June 4, 2015; accepted for publication September 24, 2015; published Early Online October 2, 2015.

Supporting information is available online at www.genetics.org/content/early/2015/09/30/genetics.115.178558/suppl/DC3.

¹Corresponding author: Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom. E-mail: k.zeng@sheffield.ac.uk

size (Zeng 2013). Given the ubiquity of deleterious mutations and the fact that diversity patterns under background selection can be similar to those expected in the presence of recurrent selective sweeps, it has been suggested that a model incorporating selection against recurrent deleterious mutations may be a better null hypothesis than the standard neutral model in the study of genome-wide polymorphism patterns (Cutter and Payseur 2013; Comeron 2014).

The effects of background selection in a subdivided population are less well understood. Existing knowledge is mainly based on the theoretical work of Nordborg (1997) and Charlesworth *et al.* (1997). Their model assumes that the population is subdivided into two demes (or subpopulations; these two terms are used interchangeably) and is in statistical equilibrium between mutation, selection, drift, and migration. Analytical equations were derived for $\overline{T_B}$ and $\overline{T_S}$, the expected coalescent time for a pair of alleles sampled either from two different demes or from the same deme, respectively. These are then used to examine how background selection affects F_{ST} , the standard measure of population differentiation, which can be defined as

$$F_{ST} = \frac{\overline{T_B} - \overline{T_S}}{\overline{T_B}} \quad (1)$$

(cf. Slatkin 1991) (for other definitions, see Charlesworth 1998 or Holsinger and Weir 2009). Under this model, the effects of background selection on pairwise coalescent times and F_{ST} can be approximated as a reduction in the effective population size, denoted by N_e , such that the effective number of migrants is reduced relative to neutrality, leading to an increase in F_{ST} (Charlesworth *et al.* 1997; Nordborg 1997). Because, all else being equal, the F_{ST} -elevating effect of background selection is expected to increase as local recombination rates decrease, the fact that most species studied to date show substantial variation in recombination rate across their genomes (Smukowski and Noor 2011) implies that background selection alone can generate large “genomic islands” that have significantly higher differentiation levels compared to the rest of the genome (Charlesworth 1998). Indeed, the importance of excluding background selection as a confounding factor in the search for genes underlying local adaptation and speciation, which can also lead to genomic islands (Wu 2001; Beaumont 2005), has been highlighted in recent reviews (Noor and Bennett 2009; Cruickshank and Hahn 2014).

Although the model of Nordborg (1997) and Charlesworth *et al.* (1997) has provided important insights into the joint effects of background selection and population subdivision, it has several limitations. First, the equilibrium assumption means that it is not applicable to the study of populations that consist of demes derived recently from a common ancestral population (*e.g.*, *Drosophila melanogaster* and humans). Second, the effect of background selection is modeled as a reduction in N_e . However, analyses of models with panmixia have shown that this approximation does not capture the fact

that background selection causes an excess of low-frequency variants in the site-frequency spectrum (SFS) (Charlesworth *et al.* 1995; Zeng and Charlesworth 2011; Nicolaisen and Desai 2012, 2013). It is unclear how the SFS is affected under the joint influence of population structure and background selection. Finally, the model provides predictions about expected diversity levels for individual neutral sites. But with the ever-increasing availability of large-scale data sets, it is desirable to be able to generate predictions for statistics that make use of multiple linked sites (*e.g.*, measures of linkage disequilibrium, the number of haplotypes), so that information in the data can be used more effectively.

In contrast to the background selection regime, the *interference selection regime* considers models with a large number of tightly linked selected sites (McVean and Charlesworth 2000; Comeron and Kreitman 2002). Interference selection is therefore an important factor for the study of asexual or highly self-fertilizing organisms as well as large genomic regions in sexually reproducing organisms that lack recombination (*e.g.*, the fourth chromosome of *D. melanogaster*). Models in this regime generally behave very differently from those in the background selection regime (reviewed by Neher 2013), and analytical results derived under background selection are typically not applicable (Kaiser and Charlesworth 2009; O’Fallon *et al.* 2010; Seger *et al.* 2010; Zeng and Charlesworth 2010). An interesting result obtained recently by Good *et al.* (2014) is that, in an equilibrium panmictic population, diversity patterns in linked neutral regions are determined by the variance in fitness at the selected sites, such that models that have different parameter values (*i.e.*, different numbers of selected sites and selection coefficients) but nevertheless identical variance in fitness will produce indistinguishable neutral diversity patterns. This *equivalence* between models implies that there is a limit to how much neutral diversity patterns can inform us about selection in their vicinity. It is of considerable interest to examine whether such a limit also exists when the population is subdivided.

In this study, we analyze models in both the background selection and interference selection regimes to address the questions raised above. In both cases, we focus on a demographic scenario where a panmictic ancestral population splits into multiple demes and examine diversity patterns for a period starting from the nonequilibrium phase immediately after the split, until the establishment of a new equilibrium. For the background selection regime, we first derive analytical results for both $\overline{T_B}$ and $\overline{T_S}$ as a function of time since the split under a two-deme model. We then construct a flexible coalescent simulator that can accommodate biologically important features including multiple demes, (nonconservative) migration, changes in population size, variation in selection coefficients against deleterious mutations across sites, and recombination. The reliability of these approximations is evaluated by checking them against extensive forward simulations. For the interference selection regime, because of a lack of a known solution of the Good *et al.* (2014) model

in subdivided populations, we present forward simulation results. Specifically, we consider models that have different parameter values but identical variance in fitness in the panmictic ancestral population and ask whether they also have similar neutral diversity patterns following the emergence of population structure.

The Model

Our analysis in both the background selection and interference selection regimes is concerned with a haploid population with a diploid phase in the life cycle where recombination may take place (e.g., Zeng and Charlesworth 2011). Forward in time, the model initially contains a single panmictic ancestral population, which later splits instantly into K demes ($K \geq 2$). Define t as the number of generations before the present and let T represent the time when the ancestral population gave rise to the demes ($T \geq 1$). The size of the ancestral population is denoted by $N_0(t)$ ($t > T$), whereas the sizes of the demes are denoted by $N_k(t)$ ($k = 1, 2, \dots, K$). In each generation, a proportion $m_{kl}(t)$ of individuals in deme k are immigrants from deme l ($k \neq l$). We focus on a genomic region with $\mathbb{S}$ linked selected sites. New mutations arise at the i th selected at rate u_i per generation. Mutation is unidirectional, from ancestral (the wild type) to derived (deleterious). The fitness of an individual with a mutation at the i th site is reduced by a fraction $1 - s_i$ (multiplicative fitness), where s_i is the selection coefficient. Mutations are assumed to accumulate in a similar manner at the neutral sites in the region.

We can describe the life cycle, using the parameters defined above. An individual in the next generation in deme k is a migrant from deme l ($l \neq k$) with probability $m_{kl}(t)$. A migrant is created from a parent or a pair of parents in the deme of origin, depending on whether recombination occurs; otherwise, it is created from parent(s) in the current deme. Parents are chosen according to their relative fitnesses within the deme of their origin. The number of new deleterious mutations arising in each chromosome is Poisson distributed with mean $U = \sum u_i$; neutral mutations are added similarly. The number of crossovers between two parental chromosomes is Poisson distributed with mean R , where R is the total genetic distance between the two ends of the genomic region of interest. This life cycle corresponds to fertility selection among adults, followed by reproduction and juvenile migration.

Approximations in the Background Selection Regime

In this section, we derive analytical results and construct a coalescent simulator, assuming that purifying selection is sufficiently strong that the frequencies of deleterious mutations are maintained close to their deterministic limit (i.e., u_i/s_i at the i th selected site), which requires $N_k(t)s_i \gg 1$ and $u_i \ll s_i$ ($i = 1, 2, \dots, \mathbb{S}$ and $k = 0, 1, \dots, K$) (see *Results* for more discussion on the condition under which the approximations are expected to be valid).

Recursion equations for a two-deme model

Assume that the ancestral population split into $K = 2$ demes at time T in the past. We focus on a single neutral site embedded in a genomic region with $\mathbb{S}$ linked selected sites. We seek to find the effects of selection at the selected sites on the distribution of coalescent times at the neutral site. The restrictive assumptions set out here are later relaxed to allow multiple neutral sites and more than two demes when the coalescent simulator is constructed. We refer to the single neutral site as the focal site. The genetic distance between the focal site and the i th selected site is denoted by r_i .

First, consider the panmictic ancestral population. With strong selection, it has been shown that the selected sites behave approximately independently (Hudson and Kaplan 1995; Johnson 1999; Nicolaisen and Desai 2013). This property can be used to derive the effective population size in the presence of background selection backward in time. Specifically, if we follow the ancestry of a chromosome sampled from the present-day population, the number of deleterious mutations linked to the focal site changes because of mutation and recombination (e.g., figure 2 of Zeng 2013). Let $P(D_i, t)$ be the probability that, at time t in the past, the focal site in an ancestral lineage (chromosome) had variant D_i at selected site i , where $D_i = 0$ or 1 depending on whether the wild type or the mutant type occurs at the site. Nicolaisen and Desai (2013) have shown that

$$P(D_i = 1, t) = 1 - P(D_i = 0, t) = \lambda_i \left(\frac{r_i}{r_i + s_i} + \frac{s_i}{r_i + s_i} e^{-(r_i + s_i)t} \right), \quad (2)$$

where $\lambda_i = u_i/s_i$. The statistical independence between selected sites suggests that the joint probability that an ancestral lineage carried variants D_i ($i = 1, 2, \dots, \mathbb{S}$) t generations ago, denoted by $P(D_1, D_2, \dots, D_{\mathbb{S}}, t)$, can be calculated as

$$P(D_1, D_2, \dots, D_{\mathbb{S}}, t) = \prod_{i=1}^{\mathbb{S}} P(D_i, t). \quad (3)$$

Note that, when $u_i \ll s_i$, the random variable D_i is approximately Poisson distributed with mean λ_i at time $t = 0$. Thus, $\prod_i P(D_i, 0)$ is equivalent to the well-known result that the number of deleterious mutations that a randomly sampled chromosome carries follows a Poisson distribution with mean $\lambda = \sum \lambda_i$ (Johnson 1999).

Two ancestral lineages in the ancestry of a sample can coalesce only when they are in the same *genetic background* (i.e., having exactly the same set of variants at the selected sites) (Hudson and Kaplan 1995; Zeng and Charlesworth 2011), which occurs with probability $P(D_1, D_2, \dots, D_{\mathbb{S}}, t)^2$. When this happens, the two lineages coalesce with probability $1/N_{\text{gb}}(t)$, where $N_{\text{gb}}(t)$ is the total number of individuals in that genetic background. To obtain $N_{\text{gb}}(t)$, we note

that, with strong selection, the selected sites remain approximately independent and the frequencies of mutants remain close to u_i/s_i , even in the presence of changes in the population size (Zeng 2013), suggesting that $N_{gb}(t) = N_0(t)\Pi_i P(D_i, 0) = N_0(t)P(D_1, D_2, \dots, D_S, 0)$. Finally, the probability that a pair of ancestral chromosomes coalesce at time t , $P_c(t)$, can be interpreted as $1/N_{e0}(t)$, where $N_{e0}(t)$ is the effective population size of the ancestral population. Put together, we arrive at the result of Nicolaisen and Desai (2013):

$$\begin{aligned} P_c(t) &= \frac{1}{N_{e0}(t)} \\ &= \sum_{\{\text{All possible combinations of values of } D_1, D_2, \dots, \text{ and } D_S\}} \frac{P(D_1, D_2, \dots, D_S, t)^2}{N_0(t)P(D_1, D_2, \dots, D_S, 0)} \\ &= \frac{1}{N_0(t)} \prod_{i=1}^S \left[\frac{P(D_i = 1, t)^2}{P(D_i = 1, 0)} + \frac{P(D_i = 0, t)^2}{P(D_i = 0, 0)} \right] \\ &\approx \frac{1}{N_0(t)} \exp \left\{ \sum_{i=1}^S \lambda_i \left(\frac{s_i}{r_i + s_i} \left(1 - e^{-(r_i + s_i)t} \right) \right)^2 \right\}. \end{aligned} \quad (4)$$

Rearranging Equation 4, the effective population size at time t is approximately

$$\begin{aligned} N_{e0}(t) &= \exp \left\{ - \sum_{i=1}^S \lambda_i \left(\frac{s_i}{r_i + s_i} \left(1 - e^{-(r_i + s_i)t} \right) \right)^2 \right\} N_0(t) \\ &= B(t)N_0(t). \end{aligned} \quad (5)$$

Letting t approach infinity, we obtain

$$N_{e0}(\infty) = \exp \left\{ - \sum_{i=1}^S \lambda_i \left(\frac{s_i}{r_i + s_i} \right)^2 \right\} N_0(t) = CN_0(t), \quad (6)$$

which is the classic result derived by Hudson and Kaplan (1995) and Nordborg *et al.* (1996) ($C = \lim_{t \rightarrow \infty} B(t)$). In other words, the classic result approximates the effects of background selection by using a neutral model with a reduced N_e and therefore does not predict that the shape of the SFS should deviate from neutral expectations. In contrast, Equation 5 shows that $N_{e0}(0) = N_0(0)$ and $N_{e0}(t)$ declines as time proceeds backward (see [Supporting Information, Figure S1](#)). Put differently, forward in time, Equation 5 approximates the effects of background selection by using a neutral model with a growing population size, which explains the characteristic excess of low-frequency variants (Charlesworth *et al.* 1995; Zeng and Charlesworth 2011; Nicolaisen and Desai 2013). Hence, we expect Equation 5 to provide better approximations when there is a significant chance for coalescent events to occur at a stage where $N_{e0}(t)$ differs substantially from $N_{e0}(\infty)$ (e.g., when the sample size is large; [Figure S1](#)).

When the ancestral population splits, because the juvenile populations of the two demes that exist before migration are propagated in the same way as a new generation is created in the ancestral population, the dynamics (backward in time) of

a chromosome sampled randomly from either deme are determined by Equation 3. Therefore, we can use Equation 5 to model the coalescent process for a sample of chromosomes. For a chromosome sampled from deme k after migration, it originates from the juvenile population of either deme k or deme l with probability $1 - m_{kl}$ and m_{kl} , respectively (assuming, for simplicity, that the migration rate is constant over time). In this case, Equation 3 can still be employed to determine the distribution of variants at the linked selected sites because $P(D_1, D_2, \dots, D_S, t) = [1 - m_{kl}] \Pi_i P(D_i, t) + m_{kl} \Pi_i P(D_i, t) = \Pi_i P(D_i, t)$. The same heuristic argument can be used to show that Equation 3 is approximately true for any generation after the split. In other words, the population structure introduced by the split does not change the frequency of the mutant variant at each selected site (see also Maruyama 1972), nor does it affect the statistical independence between selected sites. Therefore, the probability that two ancestral chromosomes coalesce in deme k at time t before the present is given by Equation 4 with $N_0(t)$ replaced by $N_k(t)$. Thus, the effective population size of deme k ($k = 1, 2$) under the influence of background selection can be modeled as

$$N_{ek}(t) = B(t)N_k(t). \quad (7)$$

Equations 5 and 7 suggest that the effects of background selection on diversity patterns at the focal site can be approximated by neutral models with time-dependent population (deme) sizes. Thus, we can derive recursion equations to describe the coalescent process, using the neutral framework developed by Teshima and Tajima (2002). As is the case for most coalescent models (Hudson 1990), we assume the population (deme) sizes [*i.e.*, $N_{ek}(t)$] to be sufficiently large and the migration rates [*i.e.*, $m_{kl}(t)$] to be sufficiently small that the chances of the following events occurring are negligible: (1) a coalescent event involving more than two ancestral lineages, (2) more than one coalescent event happening in a generation, and (3) an individual migrating multiple times in a generation.

First, consider the ancestral population. The probability that there are n ancestral lineages at time t , $Q_n(t)$, satisfies

$$Q_n(t) = Q_n(t-1) \left(1 - \frac{\binom{n}{2}}{N_{e0}(t)} \right) + Q_{n+1}(t-1) \frac{\binom{n+1}{2}}{N_{e0}(t)} \quad (8)$$

for $t > T$, with the initial condition $Q_n(T) = 1$, if n is equal to the number of ancestral lineages at time T , and 0 otherwise. The first term on the right-hand side describes a scenario where the n lineages at time $t-1$ persist to time t due to a lack of coalescence, whereas the second term considers the possibility of a coalescent event at time t , reducing the number of lineages from $n+1$ to n .

To obtain recursions for $t \in [1, T]$, during which the population is composed of two demes, we define (1) $Q_{a,n-a}(t)$,

the probability that there are a and $n - a$ ancestral lineages in demes 1 and 2, respectively, at time t [$(a, n - a)$ is referred to as a sampling configuration], and (2) $M_{a,b}(t|n)$, the probability that migration changes the sampling configuration from $(a, n - a)$ to $(b, n - b)$.

Using Equation 7, we can modify the result of Teshima and Tajima (2002), so that it takes into account the effects of background selection on the coalescent process at the focal site,

$$\begin{aligned}
Q_{a,n-a}(t) &= \sum_{b=0}^n Q_{b,n-b}(t-1) M_{b,a}(t-1|n) \left(1 - \frac{\binom{a}{2}}{N_{e1}(t)} - \frac{\binom{n-a}{2}}{N_{e2}(t)} \right) \\
&+ \sum_{b=0}^{n+1} Q_{b,n+1-b}(t-1) M_{b,a+1}(t-1|n+1) \frac{\binom{a+1}{2}}{N_{e1}(t)} \\
&+ \sum_{b=0}^{n+1} Q_{b,n+1-b}(t-1) M_{b,a}(t-1|n+1) \frac{\binom{n+1-a}{2}}{N_{e2}(t)},
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
M_{a,b}(t|n) &= \sum_{c=\max(0,a-b)}^{\min(a,n-b)} \binom{a}{c} \binom{n-a}{b-a+c} m_{12}(t)^c m_{21}(t)^{b-a+c} \\
&\times [1 - m_{12}(t)]^{a-c} [1 - m_{21}(t)]^{n-b-c}.
\end{aligned} \tag{10}$$

The three terms on the right-hand side of Equation 9 describe the following events, respectively: (1) There is no coalescence in either deme and migration changes the sampling configuration from $(b, n - b)$ at time $t - 1$ to $(a, n - a)$ at time t ; (2) migration changes the sampling configuration from $(b, n + 1 - b)$ to $(a + 1, n - a)$, followed by a coalescent event in deme 1, leading to $(a, n - a)$; (3) migration changes the sampling configuration from $(b, n + 1 - b)$ to $(a, n + 1 - a)$, followed by a coalescent event in deme 2, leading to $(a, n - a)$. Equation 10 is the same as the corresponding equation on p. 83 of Teshima and Tajima (2002) and can be derived by noting that the number of migrants among the a chromosomes in deme 1 follows a binomial distribution with parameters a and $m_{12}(t)$, and that in deme 2 is also binomially distributed with parameters $n - a$ and $m_{21}(t)$.

Assume that a_0 and b_0 chromosomes are sampled randomly at $t = 0$ (the present) from demes 1 and 2, respectively. Appealing to Equations 8 and 9, we can calculate $f(n, t | a_0, b_0)$, the probability that the number of ancestral lineages reduces from n to $n - 1$ at time t given the initial sampling configuration (a_0, b_0) , as

$$f(n, t | a_0, b_0) = \begin{cases} \sum_{a=0}^n Q_{a,n-a}(t-1) \left[\frac{\binom{a}{2}}{N_{e1}(t)} + \frac{\binom{n-a}{2}}{N_{e2}(t)} \right], & 1 \leq t \leq T \\ Q_n(t-1) \frac{\binom{n}{2}}{N_{e0}(t)}, & t > T, \end{cases} \tag{11}$$

with the initial condition that $Q_{a,b}(0) = 1$, if $a = a_0$ and $b = b_0$, and 0 otherwise. Following Teshima and Tajima (2002), we have dropped terms of order $1/[N_{ek}(t)]^2$ or higher in Equation 11, which include events such as migration followed by coalescence in the same generation or a coalescent event involving more than two ancestral lineages. Equation 11 can be solved numerically to obtain properties of the coalescent process at the focal site under the joint influence of background selection and population subdivision. For a two-deme model ($K = 2$), the complexity of the recursion equation, measured by the number of $Q_{a,b}(t)$ for a given t , is $(n + 4)(n - 1)/2$ for a sample of size n . Moreover, $f(a, t | a_0, b_0)$, where $a < n$, depends on all $f(b, t | a_0, b_0)$, where $a < b \leq n$. Thus, evaluating the recursion equation becomes complex even for a modest n . In this case, conducting coalescent simulations is a more efficient alternative (see below).

The theory developed above can be used to calculate $\overline{T_B}$ and $\overline{T_S}$, and hence F_{ST} (Equation 1), as a function of T , the time since the split of the ancestral population. We need only to consider a sample of size 2, for which there are three possible sampling configurations at $t = 0$: $(a_0, b_0) \in \{(0, 2), (1, 1), (2, 0)\}$. Equation 11 becomes

$$\begin{aligned}
f(2, t | a_0, b_0) &= \begin{cases} Q_{0,2}(t-1) \frac{1}{B(t)N_2(t)} + Q_{2,0}(t-1) \frac{1}{B(t)N_1(t)}, & 1 \leq t \leq T \\ Q_2(t-1) \frac{1}{B(t)N_0(t)}, & t > T. \end{cases}
\end{aligned} \tag{12}$$

$\overline{T_B}$ and $\overline{T_S}$ can be expressed in terms of $f(2, t | a_0, b_0)$ as

$$\begin{cases} \overline{T_B} = \sum_{t=1}^{\infty} t f(2, t | 1, 1) \\ \overline{T_S} = \frac{1}{2} \left[\sum_{t=1}^{\infty} t f(2, t | 0, 2) + \sum_{t=1}^{\infty} t f(2, t | 2, 0) \right]. \end{cases} \tag{13}$$

It is possible to weigh the contribution of the demes differently in the definition of $\overline{T_S}$ (e.g., Hudson *et al.* 1992). Our choice is partly for simplicity and partly because we used models with two equal-sized demes in most of our simulations.

We illustrate the calculation of $\overline{T_B}$ and $\overline{T_S}$, using a model where $N_0(t) \equiv aN$, $N_1(t) \equiv c_1N$, $N_2(t) \equiv c_2N$, $m_{12}(t) \equiv m_{12}$, and $m_{21}(t) \equiv m_{21}$. The migration probabilities $M_{a,b}(t | 2)$ become time independent and are denoted by $M_{a,b}$.

A matrix $\mathbb{M}$ with elements $M_{a,b}$ ($0 \leq a, b \leq 2$) can be defined as

$$\mathbb{M} = \begin{pmatrix} (1-m_{21})^2 & 2m_{21}(1-m_{21}) & m_{21}^2 \\ m_{12}(1-m_{21}) & (1-m_{12})(1-m_{21}) + m_{12}m_{21} & (1-m_{12})m_{21} \\ m_{12}^2 & 2m_{12}(1-m_{12}) & (1-m_{12})^2 \end{pmatrix}. \quad (14)$$

Equation 9 reduces to

$$\begin{cases} Q_{0,2}(t) = [Q_{0,2}(t-1)M_{0,0} + Q_{1,1}(t-1)M_{1,0} + Q_{2,0}(t-1)M_{2,0}] \\ \quad \times \left(1 - \frac{1}{c_2 NB(t)}\right) \\ Q_{1,1}(t) = Q_{0,2}(t-1)M_{0,1} + Q_{1,1}(t-1)M_{1,1} + Q_{2,0}(t-1)M_{2,1} \\ Q_{2,0}(t) = [Q_{0,2}(t-1)M_{0,2} + Q_{1,1}(t-1)M_{1,2} + Q_{2,0}(t-1)M_{2,2}] \\ \quad \times \left(1 - \frac{1}{c_1 NB(t)}\right), \end{cases} \quad (15)$$

which can be iterated numerically by using the initial condition $Q_{a,b}(0) = 1$, if $a = a_0$ and $b = b_0$, and 0 otherwise. The values of $Q_{a,b}(t)$ can then be fed into Equation 12 to obtain $f(2, t | a_0, b_0)$, which can in turn be used to evaluate Equation 13.

The fast approximation for a two-deme model

It is possible to obtain explicit analytical formulas for predicting $\bar{T}_B$ and $\bar{T}_S$ at the focal site by making an additional assumption: If t has become sufficiently large when the two ancestral lineages coalesce that $B(t) \approx C$ (see Figure S1), then the effects of background selection on the coalescent process can be well approximated by a neutral model with appropriately reduced population (deme) sizes. This is based on the *fast approximation* proposed by Nordborg (1997). It requires that the rate at which the two chromosomes of interest coalesce is much slower than changes in the distribution of variants at the linked selected sites (governed by Equation 3).

Consider, as an example, a special case of the model that leads to Equations 14 and 15 in which $N_0(t) = aN$, $N_1(t) = N_2(t) = N$, and $m_{12}(t) = m_{21}(t) = m$. Under the fast approximation, the effects of background selection can be approximated by a neutral model in which an ancestral population of size aCN gave rise to two demes of size CN in generation T before the present, with migration rates $m_{12} = m_{21} = m$. Thus, we can apply theoretical results derived for this neutral “isolation with migration” model (Nath and Griffiths 1993; Wakeley 1996; Teshima and Tajima 2002; Wilkinson-Herbots 2008). Following Wilkinson-Herbots (2008; see section 3.1 therein), we scale time in units of CN generations and define $\tau_e = T/(CN)$, $\bar{\tau}_B = \bar{T}_B/(CN)$, $\bar{\tau}_S = \bar{T}_S/(CN)$, $M_e = CNm$, $D = (16M_e)^2 + 1$, and

$$\lambda_{\pm} = \frac{4M_e + 1 \pm \sqrt{D}}{2}. \quad (16)$$

The background selection-aware versions of equations 22 and 23 of Wilkinson-Herbots (2008) can be written as

$$\begin{cases} \bar{\tau}_B = 2 + \frac{1}{2M_e} + \frac{1}{\sqrt{D}} \left[\lambda_+ \left(a - \frac{1}{\lambda_-} \right) e^{-\lambda_- \tau_e} - \lambda_- \left(a - \frac{1}{\lambda_+} \right) e^{-\lambda_+ \tau_e} \right] \\ \bar{\tau}_S = 2 + \frac{1}{\sqrt{D}} \left[(\lambda_+ - 1) \left(a - \frac{1}{\lambda_-} \right) e^{-\lambda_- \tau_e} + (1 - \lambda_-) \left(a - \frac{1}{\lambda_+} \right) e^{-\lambda_+ \tau_e} \right]. \end{cases} \quad (17)$$

Letting τ_e go to infinity, we recover the well-known result for a symmetric two-deme model (Nordborg 1997):

$$\bar{\tau}_B = 2 + \frac{1}{2M_e}, \quad \bar{\tau}_S = 2, \quad F_{ST} = \frac{1}{4M_e + 1}. \quad (18)$$

That is, the effects of background selection on F_{ST} are equivalent to a reduction in the scaled migration rate from $M = Nm$ in the absence of recurrent deleterious mutations at linked sites to M_e in the present model. As a result, F_{ST} is increased under background selection relative to neutrality (Charlesworth *et al.* 1997).

The coalescent simulator

The two approximations developed above consider a single neutral site embedded in a genomic region with S selected sites. Here we relax this restriction by extending the coalescent model of Zeng and Charlesworth (2011; Zeng 2013). This extension relies on a similar set of assumptions used to obtain the recursion equations. We demonstrate the key features of the coalescent process by focusing on the following case: (1) a genomic region with L sites, all of which are subject to purifying selection (*i.e.*, $L = S$); (2) the selection coefficient s , the mutation rate u , and the recombination rate r are all constant across sites; and (3) the population is in equilibrium and consists of two demes of size N , connected by symmetric migration occurring at rate m per generation. Thus, the total mutation and recombination rates for the region concerned are $U = uL$ and $R = rL$, respectively. The number of deleterious variants a randomly sampled chromosome carries is Poisson distributed with mean $\lambda = U/s$. Time, denoted by τ , is scaled in units of N generations, with $\tau = 0$ representing the present. The dynamics of the coalescent are determined by the following scaled parameters: $\gamma = Ns$, $\rho = Nr$, and $M = Nm$. Specifically, to generate a random genealogy for a sample of n chromosomes, we first determine the number of deleterious mutations that each chromosome carries at $\tau = 0$, which can be done by sampling from $\text{Poisson}(\lambda)$. Backward in time, four types of events can occur: mutation (loss of deleterious mutation), recombination, migration, and coalescent. Under the assumptions of the model, these events constitute four independent, competing Poisson processes with rates μ_m , μ_b , μ_{mig} , and μ_c , respectively. The time to the next event can be determined by sampling from an exponential distribution with rate $\mu = \mu_m + \mu_r + \mu_{\text{mig}} + \mu_c$. Then, the probability that this event is a mutation event is equal to μ_m/μ , proportional to its contribution to μ ; the same applies to the other events.

The rate at which a deleterious mutation is lost is γ . Therefore, $\mu_m = i\gamma$ when there are in total i mutations among all the n ancestral lineages at that time. If the next event is a

Figure 1 Pairwise coalescent times and F_{ST} as a function of time since an ancestral population splits into multiple demes. The forward simulation began with an ancestral population of size 1000 with no mutation. It was then allowed to evolve for 10^4 generations when it gave rise to 2 demes of size 5000 (A) or 10 demes of size 1000 (B). The size of the simulated region was $L = 2500$ (A) and $L = 750$ (B). In both cases, the sites in the simulated region were completely linked and all mutations were deleterious. The mutation rate per site per generation u was 2×10^{-6} (A) and 10^{-5} (B). The selection coefficient s was set to 0.01 (A) and 0.015 (B). In both models, migration was symmetrical with rate $m = m_{kl} = 5 \times 10^{-5}$ per generation ($k \neq l$). With these parameters, $n_0s = 6.065$ and 9.098 for A and B, respectively. For each time point after the split, $\overline{T}_B$ (in generations), $\overline{T}_S$ (in generations), and F_{ST} were obtained using the following methods: the recursion (Equation 13, for A only), the fast approximation (Equation 17 for A), the coalescent simulator (10^4 replicates), and the forward simulator (500 replicates). In A, the recursion and the fast approximation produce very similar results for $\overline{T}_B$, and therefore only one of the two curves is visible. In B, the fast approximation results were obtained by conducting neutral simulations with the sizes of the ancestral population and the demes reduced to $CN = 606.531$. For comparison, predictions based on neutral models with no linked selected sites were also included.

mutation event, one of the i mutations is chosen at random for removal. The rates to the next migration or recombination events are $\mu_{\text{mig}} = nM$ and $\mu_r = n\rho L$, respectively. When migration happens, a randomly chosen chromosome is moved to the other deme. For a recombination event, we pick one chromosome at random and split it into two ancestral chromosomes. In this process, the deleterious mutations originally carried by the affected chromosome are randomly distributed to the two ancestral chromosomes, using the algorithm of Zeng and Charlesworth (2011). This redistribution of mutations during recombination affects the coalescent probability between lineages in the coalescent process and is

a major feature that distinguishes the current method from that of Hudson and Kaplan (1995).

The calculation of the total coalescent rate is the most time-consuming, because it requires conducting pairwise comparisons for all chromosomes existing at the time, which in turn involves examining whether the physical distributions of the deleterious mutations carried by each pair of chromosomes (determined by the mutation and recombination events that have affected them thus far) are compatible (*i.e.*, whether it is possible for them to have been born to the same parent) and, if so, the physical distribution of these mutations in the ancestral chromosome into which these two chromosomes may

Figure 2 Pairwise coalescent times and F_{ST} under a model with recombination, two types of selected sites, and nonconservative migration. The forward simulation began with an ancestral population of size 1000 with no mutation. After 10^4 generations it split into two subpopulations, with deme 1 of size $N_1 = 1000$ and deme 2 of size $N_2 = 3000$. Shown are the pairwise coalescent times and F_{ST} as a function of time after the split. The migration rate from deme 1 to deme 2 was $m_{12} = 10^{-4}$ per generation and that in the other direction was $m_{21} = 5 \times 10^{-4}$ per generation. A 13-kb region was simulated and was composed of two flanking neutral regions of 5 kb and a central selected region of 3 kb. The selected region was composed of a sequence of triplets. Mutations arising at the first two sites of a triplet were deleterious with selection coefficients $s_1 = 0.01$ and $s_2 = 0.03$, respectively, whereas mutations arising at the third site were neutral. The per-site per-generation rates of mutation and recombination were $u = 2 \times 10^{-6}$ and $r = 10^{-6}$, respectively. Because the simulated region is symmetrical, only data from the left end (position 1), the first quartile point (position 3251), and the midpoint (position 6501) are shown. We conducted 500 replicates of forward simulation and 10^4 replicates of coalescent simulation. To obtain results using the fast approximation, simulations were conducted for the three sites, using neutral models with the population (deme) sizes reduced by $C = 0.89, 0.84$, and 0.79 , respectively. Note that the results obtained from the recursion, the fast approximation, and the coalescent simulator overlap substantially.

merge (similar to the arguments leading to Equation 4; for details, see Zeng and Charlesworth 2011 and Zeng 2013). Let μ_{cij} be the coalescent rate between chromosomes i and j (see equation 11 of Zeng and Charlesworth 2011). Then $\mu_c = \sum \mu_{cij}$ for all $i \neq j$ ($\mu_{cij} = 0$ if the two chromosomes in a pair are from different demes and/or the physical locations of deleterious mutations mean that they could not have descended from the same parent in the previous generation). Given that the next event is a coalescent, we sample i and j from a distribution with probability density μ_{cij}/μ_c and merge them into their corresponding ancestral chromosome.

Introducing neutral regions into the coalescent model is straightforward, because their effects on the coalescent process are equivalent to increasing the recombination rates between selected sites. It is also possible to accommodate multiple selection coefficients, variation in recombination rate, changes in population size, multiple demes, asymmetrical migration rates, and merging and splitting of populations. We present data obtained from some of these more complex models in *Results*.

Simulation Methods

We used forward simulations to (1) test the validity of the three approximations that we have developed in the background selection regime and (2) study neutral diversity patterns in the interference selection regime. In addition to $\overline{T}_B$

and $\overline{T}_S$, we also considered the following summary statistics: (1) the site-frequency spectrum; (2) $\overline{T}$, the total branch length; and (3) E , the ratio of the sum of all external branches (i.e., those leading to the sampled chromosomes) to $\overline{T}$. $E(\overline{T})$ is proportional to the expected number of segregating sites in the sample, and $E(E)$ is the expected proportion of singletons (i.e., polymorphic sites where the derived allele appears only once). All statistics were calculated using data from neutral sites in the simulated region.

Forward simulation

The algorithm, implemented in a Java package named *Forwards*, is an extended version of that used by Zeng and Charlesworth (2010, 2011). It is based on the life cycle described above. In all cases, we initialized the simulation with a panmictic population of size N_0 that was free of deleterious mutations. This population was allowed to evolve for at least $10N_0$ generations to reach statistical equilibrium before splitting into multiple demes. Because forward simulations are computationally expensive, we used $N_0 = 1000$ in most cases and kept track of the local genealogies at only a handful of sites across the simulated region. The results reported below were typically based on 100 replicates of simulations.

Coalescent simulation

The coalescent simulator has been implemented in a computer program called *msbgs*. This program is based on *ms* (Hudson

Figure 3 Site-frequency spectrum in an equilibrium model with background selection and population subdivision. A finite-island model with two equal-sized demes and symmetrical migration rates was simulated. The region had the same structure as the selected region used in Figure 2 (i.e., 3000 sites composed of triplets). Let N be the deme size. We used the following scaled parameters in the coalescent simulations with background selection (10^5 replicates): $\gamma_1 = Ns_1 = 10$, $\gamma_2 = Ns_2 = 30$, $\theta = Nu = 0.01$, $\rho = Nr = 0.005$, and $M = Nm = 0.75$, where the definitions of s_1 , s_2 , u , and r are the same as in Figure 2, and $m_{kl} = m$ for $\forall k \neq l$. Under neutrality (i.e., without any selection at linked sites), the value of M leads to $F_{ST} = 0.25$. In the forward simulations, the ancestral population and the demes all had

$N = 1000$ individuals (the use of this relatively small N in the forward simulations may partly explain the slight excess of low-frequency variants relative to those predicted by the coalescent model with background selection; data not shown). To sample from equilibrium, we let the population evolve for $14N$ generations after the split and then sampled every $4N$ generations until 10^4 samples were taken. A total of 30 alleles were sampled, with each deme contributing 15. The spectra at site 1500, the midpoint, were estimated. Given these parameters, at the midpoint, $C = 0.43$ and $n_0s_1 = 4.33$. The fast approximation, therefore, is equivalent to a neutral model with two equal-sized demes and an effective migration rate $M_e = CM$. Each spectrum was normalized so that the heights of its bars summed to 1.

2002). To improve efficiency, efforts have been spent on enhancing the memory management algorithm and eliminating unnecessary comparisons between ancestral chromosomes when calculating μ_c . A small-scale comparison suggests that *msbgs*, written in C, is up to two orders of magnitude more efficient than a previous nonoptimized implementation and is often comparable to *ms* in speed (data not shown). The data presented below were typically based on 10^4 replicates of simulations.

Data availability

User-friendly versions of the computer programs used in this study have been made available at <http://zeng-lab.group.shef.ac.uk>.

Result

The accuracy of the approximation methods in predicting pairwise coalescent times and F_{ST} under background selection

We assess the accuracy of the three new methods [the recursion (Equation 13), the fast approximation (Equation 17), and the coalescent simulator] by comparing their predictions to results obtained from forward simulations. In addition, we are interested in determining the condition under which the approximations are expected to perform well. Previous analysis of the behavior of Muller's ratchet in randomly mating populations of constant size N without recombination has found that the ratchet is unlikely to click when $n_0s \gg 1$, where n_0 is the number of individuals in the mutation-free class [i.e., $\text{Pr}(D_i = 0, 0)$] (Gordo *et al.* 2002; Jain 2008;

Neher and Shraiman 2012). More generally, n_0 can be expressed as CN , so that it can accommodate recombination and variation in selection coefficients. In the special case where the total deleterious mutation rate is U , the selection coefficient is s , and the total recombination rate is R , CN reduces to $Ne^{-U/(s + R/2)}$ (Nicolaisen and Desai 2013). Thus, we examine whether $n_0s \gg 1$ is also a good indicator of the validity of the new approximations. Specifically, for the rest of this study, we define n_0 as $CN_{\min}$, where $N_{\min}$ is the smallest population size among all the demes and the ancestral population in a particular model.

We begin with cases without recombination. Figure 1, A and B, is based on models where the ancestral population splits into 2 and 10 equal-sized demes, respectively. For the 10-deme case, we conducted neutral simulations in which the population (deme) sizes were reduced by a fraction of C , equivalent to the fast approximation. The analytical equations and the coalescent simulator capture the effects of background selection on the two pairwise coalescent times ($\overline{T}_B$ and $\overline{T}_S$) and F_{ST} well. There is little difference in performance between the three approximations during both the nonequilibrium phase after the split of the ancestral population and the equilibrium phase. This suggests that the effects of background selection on these three measures of diversity can be well summarized by a reduction in N_e , as assumed by the fast approximation. This is not surprising because the pairwise coalescent times are typically long compared to the time it takes for $B(t)$ (Equation 5) to converge to C (Nicolaisen and Desai 2012, 2013). It therefore explains the observation that $\overline{T}_B$ obtained under the fast approximation and the recursion is always very similar, whereas $\overline{T}_S$, which is typically much

Figure 4 The effects of the “scattered” sampling scheme on the site-frequency spectrum in a model with background selection and population subdivision. An equilibrium finite-island model with 50 equal-sized demes and symmetrical migration rates was simulated using the coalescent simulator. The structure of the simulated region and the parameters γ_1 , γ_2 , θ , ρ , and C were the same as those used in Figure 3. The scaled migration rate M was set to 3×10^{-2} . Under neutrality (*i.e.*, without any selection at linked sites), this value of M leads to $F_{ST} = 0.25$. The sample size was 10, with each allele coming from a different deme. The spectra at site 1500 were obtained using three models: (1) the full background selection model (solid bars), (2) a neutral model with deme size CN (the fast approximation, shaded bars), and (3) a neutral model with a randomly mating population of constant size (open bars). Each spectrum was normalized so that the heights of its bars summed to 1. No forward simulation was attempted because the rate of coalescence (and hence the rate of convergence to equilibrium after the split) is very slow in the presence of many demes.

smaller, can be substantially different in some cases (Figure S2, Figure S3, and Figure S4). Compared to $\overline{T}_B$ and $\overline{T}_S$, F_{ST} converges to its new equilibrium value at a much faster rate (Figure 1 and Figure 2; Figure S2, Figure S3, and Figure S4). Additional comparisons also suggest that F_{ST} patterns obtained from models with different parameter values are often more similar to each other than when $\overline{T}_B$ and $\overline{T}_S$ are compared (data not shown). These results imply that F_{ST} may be less informative about ancestral events than other aspects of the data [*e.g.*, π_B (between-deme diversity) and π_S (within-deme diversity), which are proportional to $\overline{T}_B$ and $\overline{T}_S$, respectively].

Based on a large number of parameter combinations (*e.g.*, Figure S2, Figure S3, and Figure S4), we find that the three methods generally provide fairly accurate predictions when $n_{OS} \geq 5$. When $n_{OS} < 1$, all three deviate significantly from the forward simulations, with $\overline{T}_B$ always being overestimated, but there is no consistent pattern regarding their abilities in predicting $\overline{T}_S$. With $n_{OS} < 1$, the recursion and the coalescent simulator tend to provide F_{ST} predictions that are closer to the forward simulation results.

In Figure 2, we consider a model with recombination, two types of selected sites that are subject to different intensities of purifying selection, and nonconservative migration between the two demes (*i.e.*, $N_1 m_{12} \neq N_2 m_{21}$). Because Equation 17, which requires symmetric migration, no longer applies, fast approximation results were obtained by simulating a neutral model with reduced deme sizes. Background selection depresses both $\overline{T}_B$ and $\overline{T}_S$, with the effects on the latter being more pronounced. The net effect is an elevation of F_{ST} . This phenomenon is strongest in the center of the selected region (position 6501 in Figure 2) and weakens as we move farther into the flanking neutral regions (*e.g.*, position 1 in Figure 2). Overall, the new methods agree with the forward simulations very well.

The joint effects of background selection and population subdivision on the site-frequency spectrum

When the sample size increases, so that more coalescent events took place in the recent past when $B(t)$ deviates significantly from C (Nicolaisen and Desai 2012, 2013), we expect the coalescent simulator to perform better than the fast approximation. We illustrate this point, in Figure 3, using a two-deme model, by comparing the SFS obtained from forward simulations with those obtained from conducting coalescent simulations using either the full background selection model or a neutral model with the deme size reduced from N to CN (the fast approximation). A total of 30 alleles were sampled, with each deme contributing 15. As expected, the SFS shifts toward low-frequency variants under the influence of background selection. However, the effects on intermediate- and high-frequency variants seem complex. In this particular example, background selection makes the characteristic uptick in the middle of the SFS less visible and even causes a slight excess of high-frequency variants. Although the fast approximation fails to capture these features, predictions generated by the coalescent simulator are fairly accurate (Figure 3).

Not surprisingly, when coalescence within demes is negligible (which tends to occur in the more recent past), as is the case when there are many demes and only one allele is taken from each deme, background selection has little effect on the shape of the SFS (Figure 4; see Figure S5 for another example). In fact, as the number of demes K increases, the SFS converges to that expected under a panmictic population of constant size $N_{e_large_K} = CNK[1 + 1/(2CNmK)]$, where N is the deme size and $m \equiv m_{kl}$ (*cf.* Chap. 6 of Wakeley 2008). This “scattered” sampling scheme (reviewed in Chap. 6 of Wakeley 2008) can potentially be exploited in the analysis

Table 1 Diversity patterns under different combinations of parameters

Case	N^a	u	s	r	m	T^b		E^c	
						Forward	Coalescent	Forward	Coalescent
1	1000	5×10^{-6}	0.015	0	7.5×10^{-4}	12.25 (3.32)	12.77 (3.50)	0.28 (0.081)	0.28 (0.082)
2	5000	5×10^{-6}	0.015	0	1.5×10^{-4}	11.31 (3.48)	11.41 (3.48)	0.24 (0.078)	0.24 (0.079)
3	5000	10^{-6}	0.003	0	1.5×10^{-4}	12.16 (3.25)	12.78 (3.50)	0.28 (0.084)	0.28 (0.082)
4	1000	10^{-5}	0.03	5×10^{-6}	2.5×10^{-4}	15.42 (5.82)	15.61 (5.88)	0.22 (0.082)	0.22 (0.082)
5	5000	10^{-5}	0.03	5×10^{-6}	5×10^{-5}	14.61 (5.62)	14.87 (5.90)	0.19 (0.077)	0.19 (0.077)
6	5000	10^{-6}	0.002	2×10^{-7}	5×10^{-5}	15.29 (5.61)	15.63 (5.90)	0.22 (0.081)	0.22 (0.082)

In all cases, an equilibrium island model with two equal-sized demes was considered. The simulated region was of length $L = 2000$ and new mutations were deleterious. A sample of 50 alleles, with each deme contributing 25, was considered for all parameter combinations. Genealogies at the midpoint of the region were recorded and were used to estimate the mean values of T and E and their standard errors (shown in parentheses), using 500 and 10^5 replicates of forward and coalescent simulations, respectively.

^a Deme size.

^b The total branch length, in units of N generations.

^c The ratio of the sum of all external branches (*i.e.*, those leading to the sampled alleles) to the total branch length; this ratio is equal to the proportion of singletons in the sample.

of data from structured populations. It should, however, be noted that, due to $N_{e_large_K}$'s dependency on C , the scattered sampling scheme does not remove background selection's influence on local diversity levels. Thus, everything else being equal, regions with more frequent recombination are expected to have more polymorphic sites (Figure S6).

In Table 1, we further explore how diversity patterns depend on the parameters. Cases 1 and 2 have identical u , s , r (and thus, identical C), and $M (= Nm)$, meaning that they are equivalent under the fast approximation; the same applies to cases 4 and 5. However, forward and coalescent simulations suggest that these models have different average total branch lengths (proportional to the number of segregating sites) and different average proportions of the tree accounted for by external branches (proportional to the number of singletons). In contrast, models with identical θ , γ , ρ , and M behave in a similar manner (case 1 vs. case 3; case 4 vs. case 6). These data thus provide further evidence of the inaccuracy of the fast approximation. They also suggest that, although Equations 5 and 9 depend on u , s , r , and m , it is the scaled counterparts of these parameters (*i.e.*, θ , γ , ρ , and M) that determine the behavior of the model.

Equivalence of neutral diversity patterns in the interference selection regime

We have thus far focused exclusively on cases where selection is strong (*i.e.*, $n_0s \gg 1$). However, for the study of large non-recombining genomic regions (*e.g.*, in asexual and highly self-fertilizing organisms), it is important to explore the interference selection regime. In this case, U , the total deleterious mutation rate, can become large due to tight linkage, such that $U/s \gg 1$, and n_0s approaches 0. Good *et al.* (2014) have shown that, in an equilibrium panmictic population of size N , neutral diversity patterns are determined by the variance in fitness, denoted by σ^2 . Specifically, for a given N , different combinations of U and s that lead to the same composite parameter Us^2 will have not only the same σ^2 , but also indistinguishable neutral diversity patterns. Here we use

forward simulations to explore models with the same σ^2 in the ancestral population (which can be calculated numerically using the Python package provided by Good *et al.* 2014) and follow their dynamics after the emergence of population structure and gene flow.

There is evidence for the existence of the equivalence between models with respect to neutral diversity patterns, both in the nonequilibrium phase following the divergence from an ancestral population and later in equilibrium. This has been observed for models with $Ns < 1$ (Figure 5) and those with Ns substantially > 1 (Figure 6), as well as in the following cases: (1) models with selection coefficients and sampling schemes different from those shown in Figure 5 (Figure S7 and Figure S8), (2) selection coefficients sampling from a distribution of fitness effects (Figure 6), (3) more than two demes (Figure 5B; Figure S9), and (4) a two-deme model with nonconservative migration (*i.e.*, $N_1m_{12} \neq N_2m_{21}$ and $m_{12} \neq 0$ and $m_{21} \neq 0$; Figure S10). However, we found a few exceptions, for instance, in a two-deme model in which only one of the migration rates is nonzero (Figure S11). In this case, for a fairly long period after the split, neutral diversity patterns produced by the different selection models clearly overlap before they eventually diverge from one another, suggesting that these models could still be very hard to distinguish for a substantial amount of time.

We have also investigated the effects of the scattered sampling strategy on the SFS in the many-deme limit. Since conducting forward simulations in the presence of a large number of demes is very time-consuming, only two pairs of models have been attempted (Figure 7). Interestingly, although the shape of the SFS remains similar for models with the same σ^2 in the ancestral population (*i.e.*, models 1 and 2; models 3 and 4), there is an excess of both low- and high-frequency variants. This U-shaped SFS, which is characteristic for panmictic populations under interference selection (Neher and Hallatschek 2013), is markedly different from those observed in the background selection regime (*cf.* Figure 4), suggesting some fundamental differences between the two regimes.

Figure 5 Equivalence in neutral diversity patterns between interference selection models with the same $N_0\sigma$ and weak selection. In both A and B, the forward simulation started with a single panmictic ancestral population of size $N_0 = 1000$. It was allowed to evolve for $15N_0$ generations (vertical dashed lines) before splitting into demes of size 1000 that were connected by symmetrical migration. In both models, sites in the simulated region were completely linked and all mutations were deleterious. In A, two demes were simulated and $M = 1$ (i.e., $F_{ST} = 0.2$ under neutrality and no selection at linked sites), whereas in B, five demes were included and $M = 0.025$ (i.e., $F_{ST} = 0.8$ under neutrality and no selection at linked sites). Parameters for the two weak selection models are, for model 1, $N_0U = 20$ (a 2-kb region) and $\gamma = N_0s = 0.1$ and, for model 2, $N_0U = 80$ (an 8-kb region) and $\gamma = 0.05$. Both models have $N_0\sigma = 0.396$. In A, the sample size is 25, with 5 from one deme and 20 from the other deme; in B, the sample size is 10, with two demes contributing one allele, one deme contributing two alleles, and two demes contributing three alleles. Genealogies of the samples were recorded, and at each time point we calculated (based on 100 replicates) the mean values of two summary statistics, the total branch length (in generations) and the proportion of singletons. The first statistic is proportional to the total number of segregating sites, and the latter one is commonly used to measure the excess of low-frequency variants.

Discussion

Advantages and limitations of the three approximations in the background selection regime

The above analyses show that F_{ST} values at neutral sites are expected to be higher in genomic regions that either have lower recombination frequencies (see Jackson *et al.* 2015 for a recent analysis in *D. melanogaster*) or are more tightly linked to functionally important elements (Figure 2). This heterogeneity may lead to false positives in genomic scans for outlier loci with unusually high F_{ST} (Noor and Bennett 2009; Cruickshank and Hahn 2014). A possible use of the new approximations is to generate baseline F_{ST} values in the presence of background selection for comparison. An illustrative example is a recent study by Comeron (2014) in which, by applying analytical results developed for panmictic populations (Equation 6) to a *D. melanogaster* data set, it was found that a model with background selection and recombination (induced by both crossing over and gene conversion) provides a much better explanation of patterns of diversity (as measured by π , nucleotide diversity, in putatively neutral regions) than variation in recombination rates alone. Furthermore, Comeron (2014) identified possible targets of recent positive selection (i.e., loci with lower diversity levels than predicted) and balancing selection (i.e., loci with higher than expected diversity levels). From previous studies of single populations (Charlesworth 1996, 2012b; McVicker *et al.* 2009; Cutter and Choi 2010; Rockman *et al.* 2010; Flowers *et al.* 2012; Comeron 2014), it is clear that applying these

methods requires knowledge of the deleterious mutation rate, the distribution of fitness effects of new mutations, the recombination rate (both crossover and gene conversion), and genome annotations that specify the physical locations of various functional elements. Additionally, for subdivided populations, migration rates and the time since the demes derived from the ancestral population are also needed. With the rapid accumulation of large-scale genomic resources, estimates of these parameters are becoming available to an increasingly large array of taxa (see a recent survey by Corbett-Detig *et al.* 2015). On the other hand, by examining the fit between model predictions and observations [e.g., π_B (between-deme diversity), π_S (within-deme diversity), or F_{ST}], the new methods may assist the estimation of some of the parameters (*cf.* Corbett-Detig *et al.* 2015).

For predicting F_{ST} , the three methods behave similarly (Figure 1 and Figure 2). Hence, the computationally more expensive coalescent simulator is probably not the method of choice. Since none of the methods are expected to work well in the interference regime, they should not be applied to genomic regions with severely restricted recombination rates. The fast approximation is the easiest to compute and can make use of known results derived for neutral models. In cases where analytical results for corresponding neutral models are unavailable, neutral simulations can be employed (see Figure 2). However, because C varies across putatively neutral sites in the presence of recombination (Equation 6), each focal site requires its own set of simulations (Figure 2), which may be computationally prohibitive. The recursion can readily

Figure 6 Equivalence in neutral diversity patterns between interference selection models with the same $N_0\sigma$ and strong selection. The forward simulation started with a single panmictic ancestral population of size $N_0 = 5000$. It was allowed to evolve for $10N_0$ generations (vertical dashed lines) before splitting into two demes of size $N = 5000$ that were connected by symmetrical migration with rate $m = 0.0002$ (i.e., $M = 1$ and $F_{ST} = 0.2$ under neutrality and no selection at linked sites). In both models, sites in the simulated region were completely linked and all mutations were deleterious. For model 1, $N_0U = 711.12$ (a region with 71,112 sites) and $\gamma = N_0s = 15$, and for model 2, $N_0U = 800$ (an 80-kb region) and γ follows a uniform distribution on $[0, 24.49]$ such that $\int U s^2 p(s) ds$ of model 2 equals $U s^2$ of model 1, where $p(s)$ is

the density function of the uniform distribution. Both models have $N_0\sigma = 77.68$. The sample size is 20, with 10 from each deme. The summary statistics were calculated in the same way as in Figure 5.

accommodate complicating factors such as time-dependent migration rates and/or varying population sizes. (These factors often defy the derivation of analytical equations under neutrality and hence the applicability of the fast approximation.) For a model with K demes, the complexity of the recursion, measured by the number of $Q_{k=a,l=2-a}(t)$ (i.e., the chance that, at time t in the past, a of the two chromosomes were in deme k and the rest were in deme l), is $K(K + 1)/2$. Hence, it should be applicable to models with a modest K .

The coalescent simulator possesses several advantages over the other two methods. First, it is the only one that can generate large samples with multiple linked sites and can therefore be used to assess the joint impact of population structure and background selection on statistics such as the number of segregating sites, the number of haplotypes, and measures of linkage disequilibrium, which are often included in inference methods (e.g., Duchon *et al.* 2013). Second, the coalescent simulator is the most flexible among the three with respect to incorporating biologically important factors. In addition to those presented above, it should be possible to include variation in recombination rates (Hellenthal and Stephens 2007) and gene conversion (Wiuf and Hein 2000).

An important area for future research is the generation of predictions about genome-wide diversity patterns for samples containing more than two chromosomes. It is known that the standard implementation of the *neutral* ancestral recombination graph suffers from reduced efficiency when the recombination rate is high (e.g., as a result of simulating a very large genomic region) (Marjoram and Wall 2006; Staab *et al.* 2015). The efficiency and reliability of the coalescent simulator with background selection also decline as a function of the recombination rate (see Zeng 2013 for a method for assessing this potential problem), meaning that it is not applicable to genome-level data sets. A solution proposed for the neutral case is the sequential Markov coalescent model (McVean and Cardin 2005). This approach requires that each genomic position has the same distribution of marginal genealogies, which does not apply in the presence of background

selection (e.g., Figure 2) and is therefore probably not applicable (data collected from a small-scale experiment agree with this intuition; K. Zeng, unpublished results). Another possibility is to employ a sliding-window approach that involves dividing the genome into *linkage blocks* whereby sites within a block are in strong linkage disequilibrium and are treated as if they were completely linked, whereas different blocks are treated as though they were unlinked (Good *et al.* 2014; Weissman and Hallatschek 2014). Because the dependence of diversity levels on the population size under models in the background selection regime is much stronger than under the models in the interference selection regime considered by Good *et al.* (2014) and Weissman and Hallatschek (2014), ignoring the diversity-reducing effects of unlinked blocks on the focal block as in those previous studies is probably undesirable. Fortunately, the effects of unlinked sites are known and can be modeled as a reduction in N_e in the focal block (e.g., equation 4 of Charlesworth 2012a). A third possibility is to carry out neutral simulations for each site, assuming that the effective population size at time t is determined by Equations 5 and 7 (cf. Nicolaisen and Desai 2013). However, this is likely to be computationally demanding because Equations 5 and 7 vary across sites, and therefore it is necessary to carry out simulations for each neutral site in turn. Moreover, it cannot make predictions about patterns of linkage disequilibrium between sites, which often contain useful information about demography. In sum, the development and assessment of methods for generating genome-wide predictions warrant systematic investigations in the future.

Equivalence in the interference selection regime

In contrast to the background selection regime, interference selection is induced by a large number of tightly linked selected sites (reviewed by Neher 2013). Some of the parameters we used in the simulations may not be unrealistic for the fourth chromosome of *D. melanogaster* (e.g., models 1 and 2 in Figure 6; see Kaiser and Charlesworth 2009). Our simulations provide evidence that, in subdivided populations,

Figure 7 Shape of the SFS obtained using interference selection models in the many-deme limit with scattered sampling. In all cases, the forward simulation started with an ancestral population of size $N_0 = 1000$, which, after $10N_0$ generations, gave rise to 50 demes of size $N = 1000$ connected by symmetric migration. The migration rate is $Nm = 0.03$ ($F_{ST} = 0.25$ under neutrality with no selection at linked sites) in all models. No recombination was included. Two pairs of models were considered. The values of N_0U and N_0s are 20 and 0.1 for model 1, 80 and 0.05 for model 2, 400 and 20 for model 3, and 711.12 and 15 for model 4. Thus, models 1 and 2 have the same $N_0\sigma = 0.396$ and models 3 and 4 have $N_0\sigma = 77.68$. To obtain samples from equilibrium, the populations were allowed to evolve for an additional $65N$ (for models 1 and 2, which have longer coalescent times than the other two models) and $30N$ (for models 3 and 4) generations after the split before samples were taken every $15N$ (for models 1 and 2) and $5N$ (for models 3 and 4) generations until 1000 samples were obtained. All the bars of an SFS were scaled so that their heights summed to 1.

neutral diversity patterns obtained under models with fewer selected sites under stronger selection can be very similar to those obtained under models with more selected sites under weaker selection (e.g., Figure 5). In addition, similar diversity patterns can also be generated by models in which selection coefficients follow different distributions (e.g., Figure 6). The requirement for this *equivalence* to occur is that, assuming a common demographic scenario (defined by the size of the ancestral population, the sizes of the demes, and the migration rates), the models have identical values of the compound parameter Us^2 , where $s^2 = \int s^2 \rho(s) ds$, with $\rho(s)$ being the density function of the distribution of fitness effects of new mutations. Thus, although not a proof, our simulations suggest that neutral diversity patterns collected from subdivided populations may have limited ability in differentiating different selection models in their nearby regions, as first pointed out by Good *et al.* (2014) in their analysis of panmictic populations in equilibrium.

The above results point to the importance of further theoretical developments to establish the precise condition under which the equivalence between models occurs. By developing a mapping between the strong and weak selection regimes, Good *et al.* (2014) showed that the equivalence can be exploited such that methods developed in the background selection regime can be used to generate predictions for models in the interference selection regime. If such a mapping could be established in the presence of population subdivision, then the background selection theory developed above may have application outside its intended parameter region (see Figure S12 for an exploratory experiment). Nevertheless, the fact that the scattered sampling method does not alleviate the effect of selection on the SFS, as it does in the background selection regime (Figure 4 vs. Figure 7), suggests that there are important differences between the two

regimes, which may imply that developing a mapping for subdivided populations may be more challenging than for panmictic populations. A further difficulty is that the transition from the background selection regime to the interference regime is not infinitely sharp (compare Figure 6 and Figure S13). Therefore, even if a mapping can be developed, there may be a “gray area” between the two regimes that requires a different kind of theoretical treatment. Overall, these considerations indicate that more analyses to further our understanding of the interference selection regime are warranted.

Acknowledgments

We thank Mark Beaumont, Brian Charlesworth, and two anonymous reviewers for helpful comments. We acknowledge support from the Biotechnology and Biological Sciences Research Council (BB/K000209/1) and the Natural Environment Research Council (NE/L005328/1). This project made use of the computational resources provided by the University of Sheffield’s high-performance computer cluster, Iceberg.

Literature Cited

- Beaumont, M. A., 2005 Adaptation and speciation: What can F_{ST} tell us? *Trends Ecol. Evol.* 20: 435–440.
- Charlesworth, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* 68: 131–149.
- Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* 15: 538–543.
- Charlesworth, B., 2012a The effects of deleterious mutations on evolution at linked sites. *Genetics* 190: 5–22.
- Charlesworth, B., 2012b The role of background selection in shaping patterns of molecular evolution and variation: evidence

- from variability on the *Drosophila* X chromosome. *Genetics* 191: 233–246.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.
- 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., and B. Charlesworth, 1987 Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* 18: 237–268.
- Charlesworth, D., B. Charlesworth, and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. *Genetics* 141: 1619–1632.
- Comeron, J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genet.* 10: e1004434.
- Comeron, J. M., and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. *Genetics* 161: 389–410.
- 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.
- Cruickshank, T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23: 3133–3157.
- Cutter, A. D., and J. Y. Choi, 2010 Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res.* 20: 1103–1111.
- Cutter, A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* 14: 262–274.
- Duchen, P., D. Zivkovic, S. Hutter, W. Stephan, and S. Laurent, 2013 Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. *Genetics* 193: 291–301.
- Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8: 610–618.
- Flowers, J. M., J. Molina, S. Rubinstein, P. Huang, B. A. Schaal *et al.*, 2012 Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol. Biol. Evol.* 29: 675–687.
- Good, B. H., A. M. Walczak, R. A. Neher, and M. M. Desai, 2014 Genetic diversity in the interference selection limit. *PLoS Genet.* 10: e1004222.
- Gordo, I., A. Navarro, and B. Charlesworth, 2002 Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* 161: 835–848.
- Haigh, J., 1978 The accumulation of deleterious genes in a population—Muller's ratchet. *Theor. Popul. Biol.* 14: 251–267.
- Halligan, D. L., and P. D. Keightley, 2009 Spontaneous mutation accumulation studies in evolutionary genetics. *Annu. Rev. Ecol. Syst.* 40: 151–172.
- Hellenthal, G., and M. Stephens, 2007 msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* 23: 520–521.
- Holsinger, K. E., and B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting *F* (*ST*). *Nat. Rev. Genet.* 10: 639–650.
- Hudson, R. R., 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7: 1–44.
- Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–338.
- Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. *Genetics* 141: 1605–1617.
- Hudson, R. R., D. D. Boos, and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* 9: 138–151.
- Jackson, B. C., J. L. Campos, and K. Zeng, 2015 The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. *Heredity* 114: 163–174.
- Jain, K., 2008 Loss of least-loaded class in asexual populations due to drift and epistasis. *Genetics* 179: 2125–2134.
- Johnson, T., 1999 The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. Biol. Sci.* 266: 2389–2397.
- Kaiser, V. B., and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* 25: 9–12.
- Kimura, M., and T. Maruyama, 1966 The mutational load with epistatic gene interactions in fitness. *Genetics* 54: 1337–1351.
- Marjoram, P., and J. D. Wall, 2006 Fast “coalescent” simulation. *BMC Genet.* 7: 16.
- Maruyama, T., 1972 Distribution of gene frequencies in a geographically structured population. II. Distribution of deleterious genes and of lethal genes. *Ann. Hum. Genet.* 35: 425–432.
- McVean, G. A., and N. J. Cardin, 2005 Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360: 1387–1393.
- McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155: 929–944.
- McVicker, G., D. Gordon, C. Davis, and P. Green, 2009 Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* 5: e1000471.
- Muller, H. J., 1950 Our load of mutations. *Am. J. Hum. Genet.* 2: 111–176.
- Nath, H. B., and R. C. Griffiths, 1993 The coalescent in two colonies with symmetric migration. *J. Math. Biol.* 31: 841–851.
- Neher, R. A., 2013 Genetic drift, selective interference, and population genetics of rapid adaptation. *Annu. Rev. Ecol. Syst.* 44: 195–215.
- Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA* 110: 437–442.
- Neher, R. A., and B. I. Shraiman, 2012 Fluctuations of fitness distributions and the rate of Muller's ratchet. *Genetics* 191: 1283–1293.
- Nicolaisen, L. E., and M. M. Desai, 2012 Distortions in genealogies due to purifying selection. *Mol. Biol. Evol.* 29: 3589–3600.
- Nicolaisen, L. E., and M. M. Desai, 2013 Distortions in genealogies due to purifying selection and recombination. *Genetics* 195: 221–230.
- Noor, M. A., and S. M. Bennett, 2009 Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103: 439–444.
- Nordborg, M., 1997 Structured coalescent processes on different time scales. *Genetics* 146: 1501–1514.
- Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 The effect of recombination on background selection. *Genet. Res.* 67: 159–174.
- O'Fallon, B. D., J. Seger, and F. R. Adler, 2010 A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol. Biol. Evol.* 27: 1162–1172.
- Otto, S. P., and T. Lenormand, 2002 Resolving the paradox of sex and recombination. *Nat. Rev. Genet.* 3: 252–261.
- Rockman, M. V., S. S. Skrovanek, and L. Kruglyak, 2010 Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* 330: 372–376.
- Seger, J., W. A. Smith, J. J. Perry, J. Hunn, Z. A. Kaliszewska *et al.*, 2010 Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics* 184: 529–545.
- Slatkin, M., 1991 Inbreeding coefficients and coalescence times. *Genet. Res.* 58: 167–175.

- Smukowski, C. S., and M. A. F. Noor, 2011 Recombination rate variation in closely related species. *Heredity* 107: 496–508.
- Staab, P. R., S. Zhu, D. Metzler, and G. Lunter, 2015 scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* 31: 1680–1682.
- Teshima, K. M., and F. Tajima, 2002 The effect of migration during the divergence. *Theor. Popul. Biol.* 62: 81–95.
- Wakeley, J., 1996 Pairwise differences under a general model of population subdivision. *J. Genet.* 75: 81–89.
- Wakeley, J., 2008 *Coalescent Theory: An Introduction*. Roberts & Co. Publishers, Greenwood Village, CO.
- Weissman, D. B., and O. Hallatschek, 2014 The rate of adaptation in large sexual populations with linear chromosomes. *Genetics* 196: 1167–1183.
- Wilkinson-Herbots, H. M., 2008 The distribution of the coalescence time and the number of pairwise nucleotide differences in the “isolation with migration” model. *Theor. Popul. Biol.* 73: 277–288.
- Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. *Genetics* 155: 451–462.
- Wright, A., B. Charlesworth, I. Rudan, A. Carothers, and H. Campbell, 2003 A polygenic basis for late-onset disease. *Trends Genet.* 19: 97–106.
- Wu, C. I., 2001 The genic view of the process of speciation. *J. Evol. Biol.* 14: 851–865.
- Zeng, K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity* 110: 363–371.
- Zeng, K., and B. Charlesworth, 2010 The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics* 186: 1411–1424.
- Zeng, K., and B. Charlesworth, 2011 The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics* 189: 251–266.

Communicating editor: M. A. Beaumont

GENETICS

Supporting Information

www.genetics.org/content/early/2015/09/30/genetics.115.178558/suppl/DC3

The Effects of Background and Interference Selection on Patterns of Genetic Variation in Subdivided Populations

Kai Zeng and Pádraic Corcoran

Figure S1 The convergence of Eq. (5) to Eq. (6) under different parameter combinations in a panmictic population with constant size. The size of the population is constant over time with $N \equiv N_e(0)$. Three sets of parameters were considered, as shown in the figure, where U is the total deleterious mutation rate across the region and s is the selection coefficient. In the calculations, $N = 1,000$ was assumed and the mutation rate per site per generation, u , was assumed to be 10^{-5} , leading to $Nu = 0.01$, which is realistic for *D. melanogaster*. The recombination rate per site per generation, r , was set to 10^{-5} in all cases. All the sites in the region under consideration were assumed to be selected. The different U/s ratios were obtained by adjusting the number of selected sites. Eq. (5) was used to calculate $N_e(t)$ at the site in the middle of the region. In all three cases, the convergence to Eq. (6) is evident, albeit at different rates.

Figure S2 Pairwise coalescent times and F_{ST} as a function of time since an ancestral population splits into two subpopulations of larger sizes. The forward simulation began with an ancestral population of size 1,000 and with no mutation. It was then allowed to evolve for 10^4 generations when it gave rise to two demes of size 5,000. The size of the simulated region is $L = 12,500$ in (A) and $L = 10,000$ in (B). In both cases, the sites in the simulated region were completely linked and all mutations were deleterious. The mutation rate per site per generation u was 2×10^{-6} . The selection coefficient s was set to 0.025 in (A) and 0.01 in (B). Migration was symmetrical with rate $m = m_{kl} = 5 \times 10^{-5}$ /generation ($k \neq l$) in both cases. With these parameters, $n_0s = 9.197$ and 1.353 for (A) and (B), respectively. For comparison, predictions based on neutral models with the population sizes and migration rates given above were also included.

Figure S3 Pairwise coalescent times and F_{ST} as a function of time since an ancestral population splits into two subpopulations of smaller sizes. The forward simulation began with an ancestral population of size 5,000 and with no mutation. It was then allowed to evolve for 5×10^4 generations when it gave rise to two demes of size 1,000. The size of the simulated region is $L = 7,500$ in (A) and $L = 5,000$ in (B). In both cases, the sites in the simulated region were completely linked and all mutations were deleterious. The mutation rate per site per generation u was 2×10^{-6} . The selection coefficient s was set to 0.025 in (A) and 0.01 in (B). Migration was symmetrical with rate $m = m_{kl} = 2.5 \times 10^{-4}$ /generation ($k \neq l$) in both cases. With these parameters, $n_0 s = 18.196$ and 3.679 for (A) and (B), respectively. For comparison, predictions based on neutral models with the population sizes and migration rates given above were also included.

Figure S4 Pairwise coalescent times and F_{ST} as a function of time since an ancestral population splits into 10 subpopulations. The forward simulation began with an ancestral population of size 1,000 and with no mutation. It was then allowed to evolve for 10^4 generations when it gave rise to 10 demes of size 1,000. The size of the simulated region is $L = 2,500$ in (A) and $L = 1,500$ in (B). In both cases, the sites in the simulated region were completely linked, and all mutations were deleterious. The mutation rate per site per generation u was 1×10^{-5} . The selection coefficient s was set to 0.025 in (A) and 0.0075 in (B). Migration was symmetrical with rate $m = m_{kl} = 5 \times 10^{-5}$ /generation ($k \neq l$) in both cases. With these parameters, $n_0s = 9.197$ and 1.015 for (A) and (B), respectively. The fast approximation results were obtained by conducting neutral simulations with the population sizes of the ancestral population and the demes reduced to $CN = 367.879$ in (A) and 135.335 in (B). For comparison, predictions based on neutral models with the population sizes and migration rates given above were also included.

Figure S5 The effects of the “scattered” sampling scheme on the site-frequency spectrum in a model with background selection and population subdivision. An equilibrium finite-island model with 50 equal-sized demes and symmetrical migration rates was simulated using the coalescent simulator. The structure of the simulated region and the parameters γ_1 , γ_2 , θ , ρ , and C were the same as those used in Figure 4. The scaled migration rate M was, however, set to 0.75. Under neutrality (i.e., without any selection at linked sites), this value of M leads to $F_{ST} = 0.013$. The sample size was 10, which each allele coming from a different deme. The spectra at site 1,500 were obtained using three models: (1) the full background selection model (black); (2) a neutral model with deme size CN (fast approximation; grey); (3) a neutral model with a randomly-mating population of constant size (open bars). Each spectrum was normalised so that the heights of its bars summed up to 1.

Figure S6 The effects of the “scattered” sampling scheme on the site-frequency spectrum in a model with background selection, population subdivision, and varying level of recombination. The background selection model considered was the same as the one used in Figure 4, except that two levels of recombination were simulated ($\rho = 0$ or 0.005). “No BGS” corresponds to a neutral model with 50 demes that is free of any effect of selection at linked sites. A branch is of size i if mutations falling on this branch is ultimately inherited by i chromosomes in the sample. The y-axis is the branch length, measured in units of N (deme size) generations.

Figure S7 Equivalence in diversity patterns between interference selection models with the same $N_0\sigma$. The forward simulation started with a single panmictic ancestral population of size $N_0 = 1,000$. It was allowed to evolve for $15N_0$ generations (vertical dashed lines) before splitting into two demes of size 1,000 that were connected by symmetrical migration. In both models, sites in the simulated region were completely linked, and all mutations were deleterious, and $M = 0.375$ (i.e., $F_{ST} = 0.4$ under neutrality and no selection at linked sites). Parameters for the two weak selection models are, for model 1, $N_0U = 100$ and $\gamma = N_0s = 0.5$, and for model 2, $N_0U = 2500$ and $\gamma = 0.1$. Both models have $N_0\sigma = 3.032$. The sample size is 20, with 10 from each deme. Note that branch length is measured in generations.

Figure S8 Equivalence in diversity patterns between interference selection models with the same $N_0\sigma$. The forward simulations started with a single panmictic ancestral population of size $N_0 = 5,000$. It was allowed to evolve for $10N_0$ generations (vertical dashed lines) before splitting into two demes of size $N = 5,000$ that were connected by symmetrical migration with rate $m = 0.0002$ (i.e., $M = 1$ and $F_{ST} = 0.2$ under neutrality and no selection at linked sites). In both models, sites in the simulated region were completely linked and all mutations were deleterious. For Model 1, $N_0U = 400$ (a 40 kb region) and $\gamma = N_0s = 20$, and for Model 2, $N_0U = 711.12$ (a region with 71,112 sites) and $\gamma = 15$. Both models have $N_0\sigma = 77.68$. The sample size is 20, with 10 from each deme. Note that branch length is measured in generations.

Figure S9 Equivalence in diversity patterns between interference selection models with the same $N_0\sigma$ and 30 demes. The simulation procedures and the two models presented were the same as those used to generate Figure 5. The only difference is that the ancestral population gave rise to 30 demes of size 1,000. M was set to 0.025 (i.e., $F_{ST} = 0.4$ under neutrality and no selection at linked sites). The expected coalescent times, $\overline{T}_B$ and $\overline{T}_S$, were estimated.

Figure S10 Equivalence in diversity patterns between interference selection models with the same $N_0\sigma$ and two-way migration with asymmetrical rates. The simulation procedures and the two models presented were the same as those used to generate Figure 5A (i.e., the two-deme case). However, the migration rates were asymmetrical: $m_{12} = 5 \times 10^{-4}$ and $m_{21} = 10^{-4}$.

Figure S11 Equivalence in diversity patterns between interference selection models with the same $N_0\sigma$ and unidirectional migration. The simulation procedures and the two models presented were the same as those used to generate Figure 5A (i.e., the two-deme case). However, the migration rates were asymmetrical: $m_{12} = 5 \times 10^{-4}$ and $m_{21} = 0$.

Figure S12 Comparing diversity patterns in the background selection regime with those in the interference selection regime. The data for the interference selection regime was taken from Figure 5A (i.e., the two-deme case), although only data from the equilibrium phase were shown in this figure. To conduct the coalescent simulations in the background selection regime, we assumed the region was completely linked, and *arbitrarily* set γ to 20 (i.e., Ns , where N is the deme size), but varied U , the total mutation rate.

Figure S13 Evidence of slow transition from the background selection regime to the interference selection regime. The forward simulations started with a single panmictic ancestral population of size $N_0 = 5,000$. It was allowed to evolve for $10N_0$ generations (vertical dashed lines) before splitting into two demes of size $N = 5,000$ that were connected by symmetrical migration with rate $m = 0.0002$ (i.e., $M = 1$ and $F_{ST} = 0.2$ under neutrality and no selection at linked sites). In both models, sites in the simulated region were completely linked and all mutations were deleterious. For Model 1, $N_0U = 100$ (a 10 kb region) and $\gamma = N_0s = 20$, and for Model 2, $N_0U = 177.78$ (a region with 17, 778 sites) and $\gamma = 15$. Both models have the same Us^2 . Note that the models are identical to those in Figure 6, except for the fact that the N_0U values are four times smaller. With these parameters, for Models 1 and 2, respectively, $U/s = 5$ and 11.85, and $n_0s = e^{-U/s}N_0s = 0.13$ and 0.00011. The sample size is 20, with 10 from each deme. Branch lengths are measured in generations.