Heredity

npg

ORIGINAL ARTICLE

A coalescent model of background selection with recombination, demography and variation in selection coefficients

K Zeng

There is increasing evidence that background selection, the effects of the elimination of recurring deleterious mutations by natural selection on variability at linked sites, may be a major factor shaping genome-wide patterns of genetic diversity. To accurately quantify the importance of background selection, it is vital to have computationally efficient models that include essential biological features. To this end, a structured coalescent procedure is used to construct a model of background selection that takes into account the effects of recombination, recent changes in population size and variation in selection coefficients against deleterious mutations across sites. Furthermore, this model allows a flexible organization of selected and neutral sites in the region concerned, and has the ability to generate sequence variability at both selected and neutral sites, allowing the correlation between these two types of sites to be studied. The accuracy of the model is verified by checking against the results of forward simulations. These simulations also reveal several patterns of diversity that are in qualitative agreement with observations reported in recent studies of DNA sequence polymorphisms. These results suggest that the model should be useful for data analysis.
Heredity (2013) 110, 363–371; doi:10.1038/hdy.2012.102; published online 28 November 2012

INTRODUCTION

A classic observation in the study of genetic diversity is that the level of variability at putatively neutral sites is positively correlated with the local recombination rate for the locus in question. Such a relationship has been found in many organisms such as Drosophila (Begun and Aquadro, 1992), humans (Cai et al., 2009) and Caenorhabditis (Cutter and Choi, 2010). Two population genetic models, selective sweeps and background selection, have been widely perceived as potential explanations for this correlation (Sella et al., 2009; Charlesworth, 2012a). The selective sweeps model postulates that neutral mutations linked to a nearby favourable mutation, which is on its way to fixation, will 'hitchhike' to high frequencies or even fixation, causing a substantial loss in diversity (Maynard Smith and Haigh, 1974). In contrast, the background selection model focuses on the effects of the continual removal of recurring deleterious mutations—neutral variants that occur on a haplotype carrying deleterious mutations that are doomed to be eliminated by selection will also be removed from the population, leading to a reduction in variability (Charlesworth et al., 1993). Intuitively, the association between selected and neutral variants loosens with increasing local recombination rates. Thus, both models predict a positive correlation between nucleotide diversity and recombination rates (reviewed by Sella et al., 2009; Stephan, 2010; Charlesworth, 2012a).

Although the relative importance of selective sweeps and background selection in evolution is still an open question (Sella et al., 2009; Stephan, 2010; Charlesworth, 2012a), it has been generally accepted that most newly arising mutations have deleterious effects on fitness, whereas advantageous mutations occur rather rarely (Eyre-Walker and Keightley, 2007). Therefore, it is essential to study background selection and understand how its interplay with other evolutionary forces such as demographic changes can shape genome-wide patterns of diversity. The importance of this has been highlighted by several recent examinations of DNA sequence polymorphisms in humans (McVicker et al., 2009; Hernandez et al., 2011; Lohmueller et al., 2011), Caenorhabditis (Cutter and Choi, 2010) and rice (Flowers et al., 2012) whereby broad-scale patterns of variability have been found to be compatible with predictions of a background selection model. Furthermore, there is evidence that background selection may explain the observation that the ratio of silent DNA sequence diversity for X-linked loci to that for autosomal loci is approximately one in East African populations of Drosophila melanogaster, instead of the expected value of three quarters (Charlesworth, 2012b).

To quantify the importance of background selection, having theoretical models that can efficiently generate detailed predictions about its effects on patterns of diversity is essential. Unfortunately, background selection models are typically difficult to analyse, so that existing models are often limited in scope. For instance, the standard analytic result on which most of the empirical studies cited above rely only makes predictions about the mean coalescent time for two randomly sampled alleles at a neutral site under the influence of background selection, although it takes into account recombination and allows selection coefficients against deleterious mutations to vary across selected sites (Hudson and Kaplan, 1995; Nordborg et al., 1996).

Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK
Correspondence: Dr K Zeng, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.
E-mail: k.zeng@sheffield.ac.uk

On the other hand, the best-studied model for which predictions for larger sample sizes are available is based on the simplifying assumption that there is no recombination and every deleterious mutation has the same selective effect (Charlesworth *et al.*, 1995; Gordo *et al.*, 2002; Wakeley, 2008; Walczak *et al.*, 2012). Recently, a structured coalescent procedure has been used to construct a model capable of making accurate predictions about the joint effects of background selection and recombination on patterns of diversity for an arbitrary sample size (Zeng and Charlesworth, 2011). However, the assumption that all sites in the focal region are subject to selection with a common selection coefficient severely limits the applicability of this model.

Additionally, all the above models of background selection are derived under the assumption that statistical equilibrium has been achieved in the population under consideration. However, evidence for recent changes in population size has been reported in many species. A lack of theoretical models that take into account demographic changes has caused analyses of empirical data to resort to brute-force forward simulations (Lohmueller *et al.*, 2011). But this approach is too time consuming for a thorough examination of the parameter space.

Motivated by these problems, this study attempts to construct a coalescent model of background selection that allows a flexible organization of selected and neutral sites in the region concerned, and that takes into account recombination, changes in population size and variation in selection coefficients. In addtion, a method for generating sequence variability at both selected and neutral sites is developed, allowing the correlation between these two types of sites to be studied. The accuracy of the new model is verified by checking against forward simulations. Several readily testable predictions obtained from these simulations are discussed in relation to observations reported in recent studies of DNA sequence polymorphisms.

Theory

A background selection model with constant population size and variation in selection coefficients. Consider a haploid Wright-Fisher population with constant size N, which is at mutation–selection equilibrium. Focus on a genomic region with L functionally important nucleotide sites. New mutations at these sites are unconditionally deleterious, and back mutations are ignored (unidirectional mutation). The number of deleterious mutations arising per individual per generation has a Poisson distribution with mean U, referred to as Poisson(U). The mutation rate is assumed to be uniform across sites, such that the rate per site is $u = U/L$. The mutation rate is further assumed to be sufficiently low that the infinite sites model (Kimura, 1969) can be applied. The L sites can be divided into K types, with L_k sites of type k ($k \in \{1, 2, ..., K\}$; $\sum_{k=1}^{K} L_k = L$). Therefore, the total mutation rate for the kth type of site is $U_k = uL_k$; these new mutations are referred to as mutations of type k, and their selective effect is denoted by s_k ($0 < s_k < 1$). Assuming that deleterious effects combine multiplicatively across sites, the fitness of a chromosome with d mutations in the region, where $d = (d_1, d_2, ..., d_K)$ and d_k is the number of mutations of type k, can be represented by

$$w(\boldsymbol{d}) = \prod_{k=1}^{K} (1 - s_k)^{d_k} \tag{1}$$

The life cycle is the same as that used previously by Zeng and Charlesworth (2011). Briefly, generations are assumed to be non-overlapping. The N adults in the current generation produce an

infinite pool of gametes, without fertility differences. Mutation, recombination and selection then operate on the resulting infinite population. During recombination, two randomly chosen chromosomes are paired, and genetic materials are exchanged, via crossing over, with probability R. As with mutation, the recombination rate is assumed to be uniform, such that the per-site recombination rate can be calculated as $r = R/(L-1)$. Furthermore, R is assumed to be sufficiently small that at most one crossover takes place between each pair of chromosomes. The life cycle is completed by population size regulation, which reduces the population size to a finite number by randomly sampling N chromosomes with replacement from the infinite population. All evolutionary forces are assumed to be sufficiently weak that their effects on allele frequencies are approximately additive, in which case the ordering of events becomes unimportant (Ewens, 2004).

Under the above model, the genetic composition of the population is often described by the distribution of $\boldsymbol{d}$, denoted by $f(\boldsymbol{d})$. In an infinite population at equilibrium, it has been shown that

$$f(\boldsymbol{d}) = \prod_{k=1}^{K} e^{-\lambda_k} \frac{\lambda_k^{d_k}}{d_k!} \tag{2}$$

where $\lambda_k = U_k/s_k$ (Johnson, 1999). For $K = 1$, Equation (2) holds regardless of the presence or absence of recombination (Zeng and Charlesworth, 2011). Here it is assumed that this result can be generalized to the case with $K > 1$; this is reasonable because the selected sites are expected to be at linkage equilibrium in an infinite population (Kimura and Maruyama, 1966; Charlesworth, 1990).

To construct the coalescent process, consider a sample of n chromosomes taken at random from the adult population of the current generation. Equation (2) suggests that the numbers of mutations carried by the ith ($i \in \{1, 2,..., n\}$) sampled chromosome, denoted by $\boldsymbol{d_i} = (d_{i1}, d_{i2},..., d_{iK})$, can be determined by drawing d_{ik} randomly from Poisson(λ_k) for all $k \in \{1, 2,..., K\}$. Going backwards in time, three types of events can occur—mutation, recombination and coalescent. If N is large and the evolutionary forces are weak, the likelihood for more than one event occurring in any one generation is expected to be vanishingly small (Hudson, 1990; Zeng and Charlesworth, 2011). If time is measured in units of N generations, denoted by τ, these events can be approximated as three independent, competing Poisson processes with rates σ_m, σ_r and σ_c, respectively (Zeng and Charlesworth, 2011). Hence the waiting time to the next event in the coalescent process follows an exponential distribution with rate $\sigma = \sigma_m + \sigma_r + \sigma_c$. The type of the next event is drawn in proportion to its contribution to σ. This process is repeated until only one ancestral lineage remains. Defining $\gamma_k = Ns_k$ and $\rho = NR$, the rest of this subsection sets out to describe how to calculate σ_m, σ_r and σ_c, and how each type of event alters the ancestral lineages in the coalescent process. For ease of presentation, it is assumed that $K = 2$ and $k \in \{1, 2\}$.

Consider a chromosome with $\boldsymbol{d} = (d_1, d_2)$ mutations. Using Equation (A5) of Zeng and Charlesworth (2011), it is straightforward to show that the probability that its parent in the previous generation carried $d_k^{(p)}$ mutations of type k is

$$\prod_{k=1}^{K} \frac{d_k!}{d_k^{(p)}! (d_k - d_k^{(p)})!} (1 - s_k)^{d_k^{(p)}} s_k^{d_k - d_k^{(p)}} \tag{3}$$

Expanding Equation (3) and retaining terms linear in s_1 and s_2, the probability that no mutation occurs during reproduction (that is, $d_k^{(p)} = d_k$) is approximately $1 - d_1 s_1 - d_2 s_2$, whereas the probabilities for the occurrence of one mutation of type 1 or one mutation of

type 2 are approximately $d_1 s_1$ and $d_2 s_2$, respectively. Therefore, when measuring time in units of N generations, the rate at which the lineage leading to this chromosome is affected by mutation is $d_1 \gamma_1 + d_2 \gamma_2$. Since mutations occur independently in different ancestral chromosomes, which are represented by different ancestral lineages in the coalescent process, we have $\sigma_m = \Sigma(d_{i1} \gamma_1 + d_{i2} \gamma_2)$ where d_{i1} and d_{i2} are the numbers of mutations on the ith lineage, and the summation is taken over all ancestral lineages at the time of calculation. Conditional on the occurrence of a mutation, the probability that the ith ancestral chromosome loses one of the deleterious mutations of type k is $d_{ik} \gamma_k / \sigma_m$.

The rate of the occurrence of the next recombination event, σ_r, is equal to the product of ρ and the number of lineages at the time of calculation (Hudson, 1990). Conditional on the occurrence of a recombination event, one of the lineages is taken at random. Assume that the chosen lineage represents an ancestral chromosome with $\boldsymbol{d} = (d_1, d_2)$ mutations. Recombination splits the lineage into two ancestral chromosomes as illustrated in Figure 1. Since the recombination rate is uniform across the region, the breakpoint, denoted by b, can be determined by sampling from a uniform distribution on $\{1, 2, ..., L-1\}$ (that is, the breakpoint falls between the bth and $(b+1)$th sites). For the special case with only one type of sites (that is, $K = 1$ and $\boldsymbol{d} = (d_1)$), it can be shown, by using the spatial homogeneity of the model (uniform mutation rate and equal selective effects), that the number of mutations $d_1^{(a1)}$ inherited from the first ancestral chromosome (that is, those located to the left of the breakpoint in the descendant) follows a binomial distribution with index d_1 and parameter b/L, referred to as Binomial(d_1, b/L), whereas the remaining $d_1^{(a2)} = d_1 - d_1^{(a1)}$ are inherited from the second ancestral chromosome (see Figure 2 of Zeng and Charlesworth, 2011). For $K > 1$, we note that the spatial homogeneous property holds among sites of the same type. Thus, it can be assumed that the above result for $K = 1$ is applicable to each type of site individually. Under this assumption, the number of mutations of type k inherited from the first ancestor, denoted by $d_k^{(a1)}$, can be determined by sampling from Binomial(d_k, b_k/L_k), where b_k is the number of sites of type k that fall to the left of the breakpoint b. Correspondingly,

$d_k^{(a2)} = d_k - d_k^{(a1)}$ is the number of mutations from the second ancestor (Figure 1).

To complete the reconstruction of the two ancestral chromosomes, we also need to determine the numbers of mutations in the segments non-ancestral to the descendant lineage (shown by the open boxes in Figure 1). Based on the assumption made in the previous paragraph and using the result of Zeng and Charlesworth (2011), it can be assumed that the numbers of mutations of type k in the non-ancestral segments in the first and second ancestral lineages, denoted by $q_k^{(a1)}$ and $q_k^{(a2)}$, follows Poisson distributions with means $(1-b_k/L_k)\lambda_k$ and $(b_k/L_k)\lambda_k$, respectively.

Consider the first ancestral chromosome in Figure 1. Its reconstruction has confined the $d_k^{(a1)}$ and $q_k^{(a1)}$ mutations of type k to the left and right of the breakpoint b_k, respectively; a similar restriction applies to the mutations on the second ancestral chromosome. This is in contrast to the lack of prior knowledge about the locations of the mutations on the descendant chromosome. To record this location information, which is required for calculating the coalescent probability (see below), a *genetic background* is defined for each chromosome in the coalescent process. For example, for the first ancestral chromosome in Figure 1, we can define $\mathcal{G}_k^{(a1)} = \{(b_k, L_k); (d_k^{(a1)}, q_k^{(a1)})\}$ whereby the first list in $\mathcal{G}_k^{(a1)}$ defines two non-overlapping intervals, $[1, b_k]$ and $[b_k + 1, L_k]$, and the second list contains the numbers of mutations of type k located in these two regions. Note that the two intervals in $\mathcal{G}_k^{(a1)}$ are defined respect to sites of type k alone. The overall genetic background of this chromosome can be defined as $\boldsymbol{\mathcal{G}}^{(a1)} = (\mathcal{G}_1^{(a1)}, \mathcal{G}_2^{(a1)})$.

To illustrate the calculation of the coalescent probability, first consider the special case where $K = 1$. The coalescent probability for the ith and jth lineages with genetic backgrounds $\boldsymbol{\mathcal{G}}^{(i)} = (\mathcal{G}_1^{(i)})$ and $\boldsymbol{\mathcal{G}}^{(j)} = (\mathcal{G}_1^{(j)})$ is non-zero only if $\mathcal{G}_1^{(i)}$ and $\mathcal{G}_1^{(j)}$ are compatible (see Figure 3 of Zeng and Charlesworth, 2011), in which case $\mathcal{G}_1^{(i)}$ and $\mathcal{G}_1^{(j)}$ jointly define the genetic background of the potential ancestor, denoted by $\boldsymbol{\mathcal{G}}^{(a)} = (\mathcal{G}_1^{(a)})$. Since chromosomes with compatible genetic backgrounds and their potential ancestor necessarily have identical numbers of mutations, letting $\boldsymbol{d} = (d_1)$ where d_1 is the number of mutations in $\mathcal{G}_1^{(i)}$, $\mathcal{G}_1^{(j)}$ and $\mathcal{G}_1^{(a)}$, the coalescent probability can be calculated as

$$p_c = \frac{1}{N} \frac{P(\mathcal{G}_1^{(a)})}{f(\boldsymbol{d}) P(\mathcal{G}_1^{(i)}) P(\mathcal{G}_1^{(j)})} \qquad (4)$$

where $P()$ is given by Equation (10) of Zeng and Charlesworth (2011). $P()$ describes the proportion of individuals with $\boldsymbol{d}$ mutations whose spatial distribution of mutations is specified by a given genetic background (for example, $\mathcal{G}_1^{(a)}$).

The arguments leading to Equation (4) can be generalized for $K > 1$ as follows. Given two lineages with $\boldsymbol{\mathcal{G}}^{(i)} = (\mathcal{G}_1^{(i)}, \mathcal{G}_2^{(i)})$ and $\boldsymbol{\mathcal{G}}^{(j)} = (\mathcal{G}_1^{(j)}, \mathcal{G}_2^{(j)})$, they have non-zero coalescent probability only if $\mathcal{G}_k^{(i)}$ and $\mathcal{G}_k^{(j)}$ are compatible for $1 \leqslant k \leqslant K$. The overall genetic background of the potential ancestor of the two descendant lineages is $\boldsymbol{\mathcal{G}}^{(a)} = (\mathcal{G}_1^{(a)}, \mathcal{G}_2^{(a)})$, where $\mathcal{G}_k^{(a)}$ can be determined by $\mathcal{G}_k^{(i)}$ and $\mathcal{G}_k^{(j)}$ using the algorithm of Zeng and Charlesworth (2011). As above, the numbers of mutations in $\boldsymbol{\mathcal{G}}^{(i)}$, $\boldsymbol{\mathcal{G}}^{(j)}$ and $\boldsymbol{\mathcal{G}}^{(a)}$ are the same and can be denoted by $\boldsymbol{d} = (d_1, d_2)$. The key assumption is that the function $P()$ can be generalized as

$$P(\boldsymbol{\mathcal{G}}^{(i)}) = \prod_{k=1}^{K} P(\mathcal{G}_k^{(i)}) \qquad (5)$$

which should be reasonable because the derivation of the original $P()$ relies on the spatial homogeneity property of the model, which is
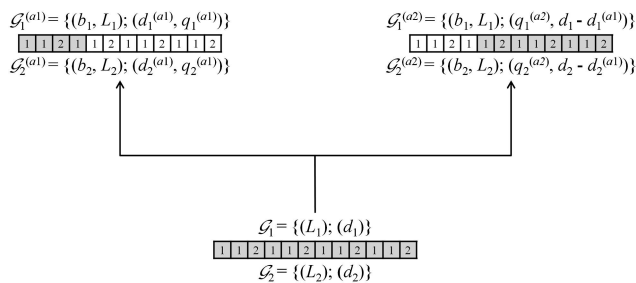

$\mathcal{G}_1^{(a1)} = \{(b_1, L_1); (d_1^{(a1)}, q_1^{(a1)})\}$

$\mathcal{G}_2^{(a1)} = \{(b_2, L_2); (d_2^{(a1)}, q_2^{(a1)})\}$

$\mathcal{G}_1^{(a2)} = \{(b_1, L_1); (q_1^{(a2)}, d_1 - d_1^{(a1)})\}$

$\mathcal{G}_2^{(a2)} = \{(b_2, L_2); (q_2^{(a2)}, d_2 - d_2^{(a1)})\}$

$\mathcal{G}_1 = \{(L_1); (d_1)\}$

$\mathcal{G}_2 = \{(L_2); (d_2)\}$

Figure 1 Reconstructing the two ancestor of a recombinant. The region concerned is composed of 12 selected sites ($L = 12$). A site is represented by a box and the number in a box shows the type of the corresponding site ($K = 2$, $L_1 = 8$ and $L_2 = 4$). The overall genetic background of the descendant chromosome is $\boldsymbol{\mathcal{G}} = (\mathcal{G}_1, \mathcal{G}_2)$, where $\mathcal{G}_k$ is defined with respect to sites of type k, and d_k is the number of mutations of type k. A breakpoint, b, is chosen at random from $\{1, 2,..., L-1\}$. Assuming $b = 4$, the figure shows how the descendant is split to two, with the grey and open boxes indicating sites that are ancestral and non-ancestral to the corresponding sites in the descendant, respectively. Note that there are three and one sites of types 1 and 2 located to the left of the breakpoint, respectively ($b_1 = 3$ and $b_2 = 1$). Methods for determining $d_k^{(a1)}$, $q_k^{(a1)}$ and $q_k^{(a2)}$ are described in the main text.

expected to be valid with respect to sites of type k. Thus, replacing $P()$ in Equation (4) by that defined by Equation (5), the coalescent probability between a given pair of lineages can be calculated. Under the scaled time, the rate for a pair of lineages to coalescent is Np_c. Assuming that it is extremely unlikely for multiple coalescent events or coalescent events involving more than two lineages to take place in one generation, the coalescent rate, σ_c, can be obtained by enumerating all possible pairs of lineages in the coalescent process at the time of calculation, and summing up values of Np_c calculated for those pairs with compatible genetic backgrounds. Conditional on the occurrence of a coalescent event, a pair of lineages is randomly chosen with probability Np_c/σ_c and is merged into an ancestral lineage with genetic background $\boldsymbol{\mathcal{G}}^{(a)}$.

Neutral sites. Suppose that, in addition to the K types of selected sites, there are also neutral sites in the focal region (that is, $\sum_{k=1}^{K} L_k < L$, where L is the total length of the region). The theory described in the previous subsection suggests that the calculations of σ_m, σ_r and σ_c are unaffected by the addition of neutral sites because the calculations concerning σ_m and σ_c, as well as the reconstruction of ancestral chromosomes (Figure 1), are carried out with respect to the selected sites alone, whereas, for σ_r, ρ is already defined with respect to the entire region. The challenge lies in determining how a recombination breakpoint will alter the genetic background of the affected lineage. This can be done by recording the relative positions of the various types of sites in the simulation algorithm.

Changes in population size. Equation (3) implies that the probability for an ancestral lineage to lose one of the d mutations it carries one generation back is approximately $\Sigma d_k s_k$, independent of the population size. This stems from the fact that strongly deleterious mutations (that is, $\gamma_k \gg 1$) are under the control of selection rather than drift, so that frequencies of deleterious variants become largely independent of the population size and approach the deterministic limit of u/s_k at each selected site (Charlesworth and Charlesworth, 2010, Sects. 6.3 and 6.4). Let $N(t)$ be the population size of the tth generation before the present ($t \geqslant 0$). If we rescale time in units of the extant population size $N(0)$, and redefine γ_k as $N(0)s_k$, the rate at which mutations carried by an ancestral lineage is being lost is $\Sigma d_k \gamma_k$; thus σ_m can be determined. Similarly, the probability that the lineage of interest is a recombinant of two ancestral chromosomes in the previous generation is R, also independent of the population size. Redefining ρ as $N(0)R$, σ_r can be calculated as described previously. In contrast, the coalescent probability is a function of the population size, and Equation (4) should be redefined as $p_c(t)$ ($t \geqslant 1$) by replacing N with $N(t)$. We can obtain σ_c by applying $p_c(t)$ to standard formulae for calculating time-dependent coalescence rates (for example, Equation (4) of Slatkin and Hudson, 1991; see Supplementary Text). With these amendments, the coalescent process can be constructed.

Generating sequence variability at both selected and neutral sites. Sequence variability at selected sites can be generated by making use of the location information on the deleterious mutations that is recorded when reconstructing the history of the sampled chromosomes. This is illustrated in Figure 2 for a case where two chromosomes are sampled, and the focal region consists of 10 selected sites of the same type (that is, $K = 1$ and $L = L_1 = 10$). The two sampled chromosomes carry one and two deleterious mutations, respectively, as indicated by their genetic backgrounds $\mathcal{G}_1^{(1)}$ and $\mathcal{G}_1^{(2)}$. However, we do not have any prior knowledge of the locations of these mutations; nor do we know whether they are specific to one of
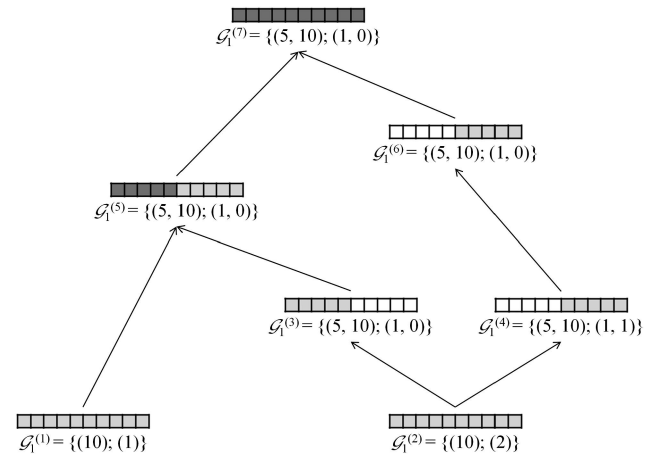

Figure 2 An example genealogy for a sample of two chromosomes. The region concerned is composed of 10 selected sites of the same type ($K = 1$, $L = L_1 = 10$). The genetic background of the ith individual in the genealogy is denoted by $\mathcal{G}_1^{(i)}$ ($i \in \{1, 2, ..., 7\}$). A light grey box represents a site ancestral to the corresponding site in one of the two sampled chromosomes, a dark grey box represents a site ancestral to the corresponding site in both sampled chromosomes and an open box represents a site non-ancestral to the sample. See the main text for a description of how sequence variability at the selected sites can be determined at the same time as the genealogy is reconstructed.

the sampled chromosomes or are shared. Going backwards, the first event is a crossover with a breakpoint between the 5th and 6th sites. This splits $\mathcal{G}_1^{(2)}$ into two ancestral lineages, with genetic backgrounds $\mathcal{G}_1^{(3)}$ and $\mathcal{G}_1^{(4)}$, respectively. In this example, interval [1, 5] of $\mathcal{G}_1^{(3)}$ (indicated with the light grey boxes) is ancestral to the corresponding region of $\mathcal{G}_1^{(2)}$, whereas the other half of $\mathcal{G}_1^{(2)}$ is inherited from interval [6, 10] of $\mathcal{G}_1^{(4)}$. Each of these two ancestral segments contributes one deleterious mutation to $\mathcal{G}_1^{(2)}$. Note that there is no deleterious mutation in the second half of $\mathcal{G}_1^{(3)}$, which is non-ancestral to $\mathcal{G}_1^{(2)}$ (indicated with the open boxes), whereas one mutation exists in the non-ancestral segment in $\mathcal{G}_1^{(4)}$.

Next, $\mathcal{G}_1^{(1)}$ and $\mathcal{G}_1^{(3)}$ coalesce into a common ancestor $\mathcal{G}_1^{(5)}$. Since the mutation in $\mathcal{G}_1^{(3)}$ is confined to [1, 5], this restricts the mutation in $\mathcal{G}_1^{(5)}$ to [1, 5] (see also Figure 3 of Zeng and Charlesworth, 2011). Note that the most recent common ancestor of the first five sites in the focal region has been reached in this event (indicated with the dark grey boxes). Combining these two factors, we can deduce that a deleterious mutation located somewhere in [1, 5] is shared by both chromosomes in the sample. Going further backwards, $\mathcal{G}_1^{(4)}$ is changed to $\mathcal{G}_1^{(6)}$ by a mutation event that removes the mutation located in [6, 10] of $\mathcal{G}_1^{(4)}$. Since [6, 10] of $\mathcal{G}_1^{(4)}$ is ancestral to the corresponding segment of $\mathcal{G}_1^{(2)}$, this mutation therefore only exists in the second sampled chromosome. Finally, a common ancestor event occurs between $\mathcal{G}_1^{(5)}$ and $\mathcal{G}_1^{(6)}$, thereby completing the reconstruction of the ancestry of the sample. If there are neutral sites in the focal region, sequence variability can be generated using standard methods conditional on the reconstructed genealogy (Hudson, 1990).

MATERIALS AND METHODS
Simulations
The validity of the coalescent model was checked against forward simulations. The forward simulation algorithm is similar to the one described previously (Zeng and Charlesworth, 2011), but with the added ability of simulating models with variation in selection coefficients and changes in population size. Unless stated otherwise, a haploid population size of $N = 5000$ was used. For

each combination of parameter values, the population was first allowed to evolve for $10N$ generations for statistical equilibrium to be achieved; then random samples of size n were taken every $4N$ generations, until 2000 samples were obtained. The coalescent model was implemented by extending the algorithm of Zeng and Charlesworth (2011), which is in turn a modified version of the *ms* simulator (Hudson, 1990). Since the coalescent model is computationally much more efficient, 10^5 samples were simulated for each parameter combination. In both types of simulations, gene genealogies at a number of sites across the focal region were recorded. All the computer programs are available upon request.

Statistics of interest
Two tree-based statistics were calculated using the genealogies obtained from the simulations—T_n, the total branch length, and ξ_n, the ratio of the total length of external branches (that is, branches leading to chromosomes in the sample) to T_n, where n is the sample size. In the rest of this study, both statistics are expressed relative to their neutral expectations, such that, when calculated using data simulated under strict neutrality, their expected values, denoted by $E(T_n)$ and $E(\xi_n)$, are one. For $n=2$, the effect of background selection on $E(T_2)$ at a linked neutral site can be approximated by

$$E(T_2) \approx \exp\left\{ - \sum_i \frac{u_i}{s_i(1+(1-s_i)r_i/s_i)^2} \right\} \qquad (6)$$

where r_i is the recombination rate between the ith selected site and the neutral site concerned, and u_i and s_i are the mutation rate and the selection coefficient at the ith selected site (Hudson and Kaplan, 1995; Nordborg *et al.*, 1996).

The accuracy of the coalescent model was further assessed by several sequence-based statistics. Sequence variability at both selected and neutral sites was used to calculate: π, the nucleotide site diversity; Watterson's θ_W (Watterson, 1975); Tajima's D (Tajima, 1989); Fu and Li's D (Fu and Li, 1993). Note that T_2 is proportional to π, and T_n ($n>2$) is proportional to θ_W. On the other hand, ξ_n, which measures the relative prevalence of singletons (that is, polymorphic sites with the derived allele present in one copy in the sample), is closely related to Tajima's D and Fu and Li's D. For all statistics (both tree- and sequence-based), mean and s.d. were estimated. In some cases, the correlation coefficient (Cor) between values of a statistic calculated for the selected sites and those calculated for the linked neutral sites was also estimated.

RESULTS
Approximating the effects of variation in selection coefficients in an equilibrium population
The accuracy of the coalescent model was first assessed by comparing its predictions about the mean and s.d. of the two tree-based statistics, T_n and ξ_n, where n is the sample size, to those obtained from forward simulations. For $n=2$, the coalescent model was also checked against the analytic approximation (that is, Equation (6)). In Table 1, the case

with complete linkage between sites is considered. It can be seen that estimates of mean and s.d. obtained from the coalescent model are rather close to those obtained from forward simulations. Compared with the coalescent model, predictions of $E(T_2)$ based on Equation (6) tend to be more different from the forward simulation results, and are consistently smaller. The extent of this underestimation is greatest when the deleterious mutation rate is high (compare cases 4 and 5 in Table 1). In contrast, the coalescent model appears to be much more robust to the presence of a high mutation rate.

Figure 3 examines the case where recombination is incorporated. The simulated region is composed of 2001 sites, close to the average size of the coding region of a gene in the *D. melanogaster* genome (Loewe and Charlesworth, 2007). For each triplet in the region, the first two sites are subject to selection with $\gamma_1 = Ns_1 = 20$ and $\gamma_2 = Ns_2 = 50$, respectively, whereas the third site is assumed to be neutral. Thus, the focal region has two types of selected sites ($K=2$ and $L_1 = L_2 = 667$). This is to imitate the codon structure whereby the selected sites correspond to first and second codon positions, at which most mutations are non-synonymous, and the neutral sites correspond to third codon positions, at which most mutations are synonymous. The values of γ_1 and γ_2 concerned are broadly in line with recent estimates of the intensity of selection against non-synonymous mutations in *Drosophila* (Loewe and Charlesworth, 2007). The scaled mutation rate per site is $\theta = Nu = 0.005$, which corresponds to an average nucleotide site diversity of $\pi = 0.01$, comparable to the value observed in African populations of *D. melanogaster*.

Simulations have been conducted to generate random samples of size 20 under three different levels of recombination with $r/u = 0.25$, 1 and 5, respectively, where r is the recombination rate per site (Figure 3). The coalescent model is able to provide fairly accurate approximations to both $E(T_{20})$ and $E(\xi_{20})$, although a slight tendency to overestimate $E(T_{20})$ can be seen (see Discussion). From Figure 3, it is clear that background selection reduces diversity at linked neutral sites (as measured by T_n, which is proportional to Watterson's θ_W), and increases the prevalence of low-frequency variants (as measured by ξ_n). This effect cannot be neglected even in regions where recombination is relatively frequent (for example, $r/u = 5$).

Approximating the joint effects of background selection and changes in population size
A population bottleneck model is considered (Figure 4). Time and the mutation and selection parameters are scaled by $N = 10\,000$ generations (the population size prior to the bottleneck). Let τ denote the

Table 1 Patterns of diversity under a background selection model with complete linkage and variation in selection coefficients

Case	$E(T_2)$			$E(T_{20})$		$E(\xi_{20})$	
	Equation (6)	Forward	Coalescent	Forward	Coalescent	Forward	Coalescent
1	0.472	0.515 (0.418)	0.550 (0.487)	0.587 (0.162)	0.623 (0.178)	1.300 (0.385)	1.295 (0.386)
2	0.449	0.485 (0.411)	0.521 (0.461)	0.552 (0.156)	0.593 (0.169)	1.306 (0.379)	1.297 (0.385)
3	0.311	0.391 (0.314)	0.414 (0.330)	0.481 (0.122)	0.510 (0.126)	1.454 (0.374)	1.459 (0.371)
4	0.069	0.238 (0.173)	0.224 (0.136)	0.327 (0.074)	0.327 (0.059)	1.688 (0.355)	1.783 (0.322)
5	0.263	0.351 (0.275)	0.370 (0.286)	0.436 (0.109)	0.463 (0.110)	1.490 (0.367)	1.502 (0.367)

T_n is the total branch length of the genealogy for a sample of size n, and ξ_n is the proportion of T_n accounted for by external branches. Both statistics are expressed relative to their neutral expectations, such that $E(T_n) = E(\xi_n) = 1$ under neutrality. The numbers shown in parentheses are s.d. There are K types of sites in the focal region, with mutations of type k having selective effect s_k ($k \in \{1, 2, ..., K\}$). The number of sites of type k is denoted by L_k. In all the simulations, the per-site mutation rate u is uniform across the region. The mutation and selection parameters are scaled by N, the population size, such that $\theta = Nu$ and $\gamma_k = Ns_k$. In the absence of recombination, the properties of the statistics are unaffected by the physical organization of the selected sites and the presence/absence of neutral sites. For case 1, $\theta = 0.005$, $K = 2$, $L_1 = L_2 = 1000$, $\gamma_1 = 10$ and $\gamma_2 = 20$. For case 2, $\theta = 0.005$, $K = 3$, $L_1 = L_2 = L_3 = 1000$, $\gamma_1 = 10$, $\gamma_2 = 25$ and $\gamma_3 = 50$. For case 3, $\theta = 0.005$, $K = 2$, $L_1 = 1000$, $L_2 = 2000$, $\gamma_1 = 10$ and $\gamma_2 = 15$. Cases 4 and 5 are similar, both with $K = 2$, $L_1 = 1334$, $L_2 = 2668$, $\gamma_1 = 10$ and $\gamma_2 = 20$; however, the scaled mutation for case 4 is $\theta = 0.01$, which is twofold higher than the value of 0.005 for case 5.
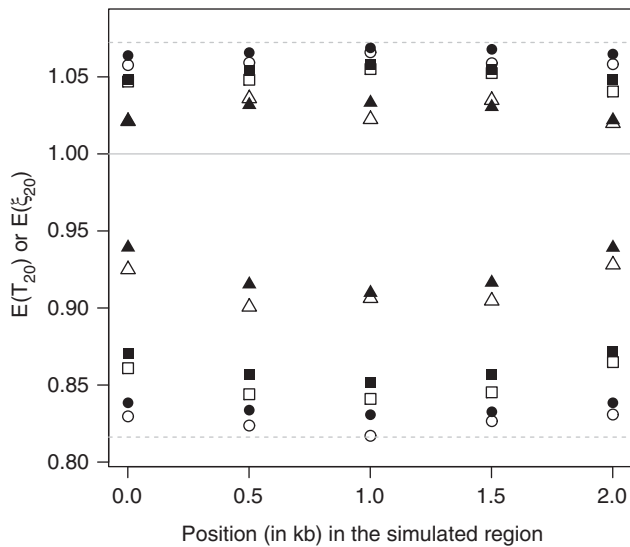
Figure 3 $E(T_{20})$ and $E(\xi_{20})$ across a region where the selective effects vary across sites. The population is assumed to be at statistical equilibrium. The simulated region is composed of 2001 sites. For each triplet of sites, the first two sites are subject to selection with $\gamma_1 = Ns_1 = 20$ and $\gamma_2 = Ns_2 = 50$, respectively, whereas the third site is assumed to be neutral. The scaled mutation is $\theta = Nu = 0.005$ per site. The level of recombination is measured by r/u, where r is the per-site recombination rate. Simulations have been conducted with $r/u = 5$, 1 and 0.25, with the corresponding results represented by triangles, squares and circles, respectively. Among the results represented by symbols of the same shape, those placed above one are for $E(\xi_{20})$, whereas those placed below are for $E(T_{20})$. Filled and open symbols show estimates obtained from coalescent and forward simulations, respectively. The dashed lines at the top and bottom of the figure show values of $E(\xi_{20})$ and $E(T_{20})$ when $r/u = 0$, respectively.

scaled time. The first change in population size is assumed to take place at $\tau = 0$. The tree-based statistics are calculated using random samples of size 20 taken randomly at various time points during the bottleneck (coloured arrows in Figure 4a), and are expressed relative to the equilibrium values expected under a neutral model with a constant population size of 10 000. The simulated region is composed of a centrally located selected region (indicated with the horizontal red lines in Figures 4b and c) and two flanking neutral regions. The selected region is the same as that used in Figure 3 (that is, 2001 sites, $K = 2$, $L_1 = L_2 = 667$, $\gamma_1 = 20$, $\gamma_2 = 50$ and $\theta = 0.005$ per site).

Figure 4 suggests that the coalescent model can provide very accurate predictions about patterns of diversity during the course of the bottleneck. Interestingly, across the focal region, the spatial distribution of $E(T_{20})$ remains U-shaped throughout the event. A similar observation applies to $E(\xi_{20})$ whose spatial distribution is always n-shaped, although the spatial pattern for $E(\xi_{20})$ is less visible than that for $E(T_{20})$. The quality of approximation to the s.d. of the statistics is also good (Supplementary Table S1). In Supplementary Figures S1 to S3, results obtained from simulations concerning a different bottleneck model, a population expansion model and a population contraction model have been presented. In all cases, the coalescent model is able to accurately capture the joint effects of background selection and demographic changes.

Studying sequence variability at both selected and neutral sites
So far only the tree-based statistics have been considered. Nonetheless, these variables are not directly observable, but have to be inferred

from sequence variability. Therefore, the coalescent model has been further examined by checking its ability to predict the statistical properties of four widely used statistics: π, Watterson's θ_W, Tajima's D and Fu and Li's D. First the bottleneck model used to produce Figure 4 is considered. Neutral variants located in the two flanking regions and neutral variants located within the selected region have been analysed separately (Table 2). For all statistics, the mean and s.d. produced by the coalescent model are quantitatively close to those obtained from the forward simulations, suggesting that the coalescent model is also able to accurately capture aspects of the frequency spectrum that are not represented by the two tree-based statistics.

The data presented in Table 2 contain information useful for empirical studies. First, the level of neutral diversity within the selected region (as measured by π and θ_W) is expected to be lower than that in the flanking regions. This appears to be true for equilibrium populations ($\tau < 0$) and populations that have experienced recent changes in population size ($\tau > 0$; see also Supplementary Figures S1 to S3). In contrast, the mean values of the two D-test statistics calculated from neutral variants embedded in the selected region can be either smaller ($\tau < 0$) or larger ($\tau = 0.325$) than those calculated from neutral variants in the flanking regions, despite the observation of a higher prevalence of singletons for the former set of neutral sites (Figure 4c; see also Supplementary Figures S1 to S3). The mean values of Fu and Li's D deviate more from the neutral expectation of zero compared with Tajima's D, suggesting that Fu and Li's D may be more informative about background selection (Charlesworth et al., 1995), although its statistical power may be compromised by its high variance.

The last experiment assesses whether the coalescent model is also able to predict patterns of diversity at selected sites. Results obtained from four parameter combinations are presented in Table 3. Several observations can be made. First, the mean values of the four sequence-based statistics predicted by the coalescent model match closely with those obtained from forward simulations. Second, for all statistics, there is a positive correlation (that is, $\text{Cor} > 0$) between values obtained from the selected variants and values obtained from the neutral variants (see also Supplementary Figure S4). The correlation for Fu and Li's D appears to be somewhat stronger, although this may be hard to detect in data due to statistical noise. As expected, increasing the recombination rate reduces both s.d. and Cor. On the other hand, everything else being equal, the extent of correlation decreases with increasing γ (see also Supplementary Figure S4).

The coalescent model is able to capture these general trends regarding s.d. and Cor. However, it tends to underestimate s.d. and Cor for π and θ_W, especially when selection is relatively weak and recombination rate is low (for example, $\gamma = 7.5$ and $r/u = 0$). This is probably caused by a combination of more variable allele frequencies at weakly selected sites in a finite population due to drift and the extra variance induced by the interference between linked selected variants (Hill and Robertson, 1966). Interestingly, the estimates of these higher-order moments for both Tajima's D and Fu and Li's D remain fairly accurate for all the parameter combinations considered, suggesting that these statistics may be more suitable to be used with the coalescent model in data analysis.

DISCUSSION
The results presented above suggest that the coalescent model can make accurate predictions about the effects of background selection on patterns of diversity in the presence of recombination, recent changes in population size and variation in selection coefficients across selected sites. The level of accuracy seems to be higher for $E(\xi_n)$
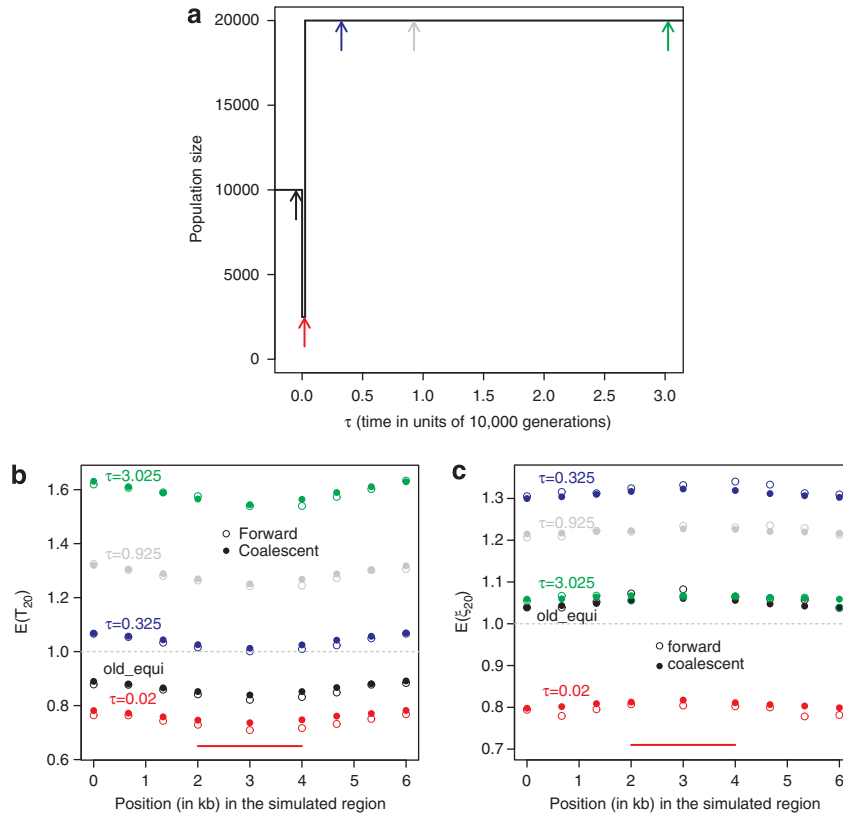
Figure 4 Patterns of diversity in a region under background selection during the course of a population bottleneck event. Initially, the population is at equilibrium with $N = 10\,000$. Let τ denote the time measured in units of $10\,000$ generations. As shown in (**a**), the population size decreases instantly to 2500 at $\tau = 0$ and increases instantly to $20\,000$ at $\tau = 0.025$. Random samples of size 20 were simulated at the five time points indicated with the coloured arrows in (**a**). The simulated region is composed of a centrally located selected region with 2001 sites, as indicated with the horizontal red lines in (**b**) and (**c**), and two flanking, neutrally evolving regions with 2000 sites each. The organization of the selected and neutral sites within the selected region is the same as that described in Figure 3 (that is, $K = 2$, $L_1 = L_2 = 667$). Scaling by $N = 10\,000$, $\gamma_1 = 20$, $\gamma_2 = 50$ and $\theta = 0.005$ per site. The per-site recombination rate r is equal to $0.5u$. Estimates of $E(T_{20})$ and $E(\xi_{20})$ are shown in (**b**, **c**), respectively. In these two plots, open and filled symbols indicate results obtained from forward and coalescent simulations, respectively.

Table 2 Mean and s.d. of four summary statistics calculated using neutral variants at three time points during the bottleneck

τ	Region	$\pi\ (\times 10^{-3})$		$\theta_W\ (\times 10^{-3})$		Tajima's D		Fu and Li's D	
		Forward	Coalescent	Forward	Coalescent	Forward	Coalescent	Forward	Coalescent
<0	Flanking	8.4 (2.1)	8.6 (2.1)	8.5 (1.6)	8.7 (1.7)	−0.066 (0.50)	−0.092 (0.49)	−0.106 (0.61)	−0.143 (0.61)
	Embedded	8.1 (3.2)	8.2 (3.3)	8.2 (2.6)	8.4 (2.7)	−0.095 (0.72)	−0.115 (0.73)	−0.125 (0.85)	−0.154 (0.87)
0.02	Flanking	7.9 (2.1)	8.1 (2.1)	7.4 (1.6)	7.6 (1.7)	0.272 (0.55)	0.227 (0.55)	0.467 (0.64)	0.402 (0.66)
	Embedded	7.7 (3.2)	7.8 (3.3)	7.2 (2.5)	7.4 (2.6)	0.216 (0.78)	0.181 (0.77)	0.339 (0.85)	0.287 (0.85)
0.325	Flanking	9.2 (2.1)	9.3 (2.1)	10.3 (1.7)	10.5 (1.7)	−0.457 (0.44)	−0.470 (0.43)	−0.708 (0.59)	−0.739 (0.59)
	Embedded	8.8 (3.2)	9.0 (3.3)	9.9 (2.7)	10.1 (2.8)	−0.449 (0.65)	−0.478 (0.65)	−0.621 (0.85)	−0.663 (0.85)

Standard deviations are shown in parentheses. The model used to produce Figure 4 was simulated. Random samples of size 20 were generated at three time points (indicated with black, red and blue arrows in Figure 4a). The statistics were calculated using sequence variability at two types of neutral sites—those in the two flanking regions (Flanking) and those embedded in the centrally located selected region (Embedded).

than for $E(T_n)$ in many cases (Supplementary Table S2). Due to computational constraints, relatively small population sizes were used in the forward simulations (typically 5000). As discussed in the Supplementary Text, this is likely to have contributed to the observed differences between the coalescent and forward simulations (especially with respect to $E(T_n)$). This discussion has led to the development of two diagnostic criteria that can be used to predict the reliability of the

coalescent model (Supplementary Equations (S1) and (S2)). These diagnostic criteria have been tested using simulation results obtained from 203 combinations of parameter values. The results suggest that the diagnostic criteria, which make predictions based solely on data generated by the coalescent model, have high degrees of accuracy (80–85%) and low error rates ($\sim$5%; see Supplementary Text for details). These estimates are likely to be conservative because the

Table 3 Patterns of diversity at selected sites in the focal region

γ	r/u	Model	π			θ_W			Tajima's D			Fu and Li's D		
			Mean	s.d. ($\times 10^{-4}$)	Cor	Mean	s.d. ($\times 10^{-4}$)	Cor	Mean	s.d.	Cor	Mean	s.d.	Cor
15	0	Forw	6.7×10^{-4}	2.88	0.028	0.0015	5.88	0.082	−1.741	0.342	0.083	−2.312	0.918	0.144
		Coal	6.5×10^{-4}	2.12	0.021	0.0015	4.96	0.064	−1.793	0.324	0.079	−2.417	0.910	0.157
15	3	Forw	6.7×10^{-4}	2.81	0.026	0.0015	5.79	0.071	−1.755	0.329	0.072	−2.336	0.896	0.112
		Coal	6.5×10^{-4}	2.12	0.019	0.0015	4.88	0.057	−1.811	0.305	0.076	−2.449	0.883	0.114
7.5	0	Forw	0.0014	5.44	0.242	0.0025	8.21	0.233	−1.542	0.430	0.274	−1.977	0.971	0.279
		Coal	0.0013	2.91	0.063	0.0025	6.47	0.139	−1.712	0.341	0.191	−2.228	0.922	0.288
7.5	3	Forw	0.0014	4.67	0.084	0.0025	7.70	0.126	−1.625	0.367	0.115	−2.064	0.901	0.158
		Coal	0.0013	2.91	0.031	0.0026	6.30	0.088	−1.777	0.306	0.108	−2.311	0.863	0.159

In all cases, the focal region is composed of 2001 sites. Every third site is neutrally evolving, and all the other sites are subject to selection with the same γ. The scaled mutation is $\theta = 0.005$ per site and the sample size is 20. The mean and s.d. of the four statistics were estimated using sequence variability at the selected sites alone. In contrast, the correlation coefficient (Cor) was estimated using data from both the selected and neutral sites. Note that the expected value of π in an infinite population is 6.7×10^{-4} for $\gamma = 15$ and 0.0013 for $\gamma = 7.5$.

parameters were chosen from regions of the parameter space that are likely to be challenging for the coalescent model (that is, when selection is weak with $\gamma \leqslant 7.5$; Supplementary Tables S2 and S3).

The simulations reveal several patterns that are readily testable using DNA sequence polymorphisms: (1) a positive correlation between neutral diversity and recombination rates (Figure 3; Shapiro et al., 2007; Cai et al., 2009; Cutter and Choi, 2010); (2) a positive correlation between neutral diversity and the distance to regions subject to background selection (Figure 4b; McVicker et al., 2009); (3) a deficit of diversity in the middle of a selected region (Figures 3 and 4b; Comeron and Guthrie, 2005); (4) a negative correlation between the prevalence of singletons (or low-frequency variants) and recombination rates (Figure 3; Shapiro et al., 2007; Lohmueller et al., 2011); (5) a negative correlation between the prevalence of singletons and the distance to regions subject to background selection (Figure 4c; Lohmueller et al., 2011); (6) an excess of singletons in the middle of a selected region (Figures 3 and 4c). Encouragingly, these patterns exist in the presence of recent demographic changes (Figure 4; Supplementary Figures S1 to S3), suggesting that previous analyses based on these correlations are likely to be robust.

Table 3 suggests that, under background selection, there should be a positive correlation between values of a summary statistic calculated separately using selected and neutral variants; the extent of this correlation decreases as γ increases (see also Supplementary Figure S4). A significant positive correlation between π_A (π at non-synonymous sites) and π_S (π at synonymous sites) has been reported in D. melanogaster (Haddrill et al., 2011). These data were reanalysed by focusing on the 112 loci with $K_A < 7.5\%$; this should remove fast-evolving loci that may have been subject to recurrent positive selection (see Figure 4 of Haddrill et al., 2011). These loci were further divided into two equal-sized groups according to their K_A values, so that the group with lower (or higher) K_A should mainly consist of genes under substantial (or weaker) selective constraints. Consistent with higher levels of constraint, the first group has a significantly lower mean π_A (0.0022 versus 0.0034; Mann–Whitney test, $P = 0.001$) and a nominally more negative mean TD_A (Tajima's D calculated using non-synonymous variants; −0.34 versus −0.16; Mann–Whitney test, $P = 0.15$). The correlation coefficient between π_A and π_S is 0.29 in the more constrained group, lower than the value of 0.49 in the other group (quantitatively similar results were obtained after controlling for K_A and K_S using partial correlation methods).
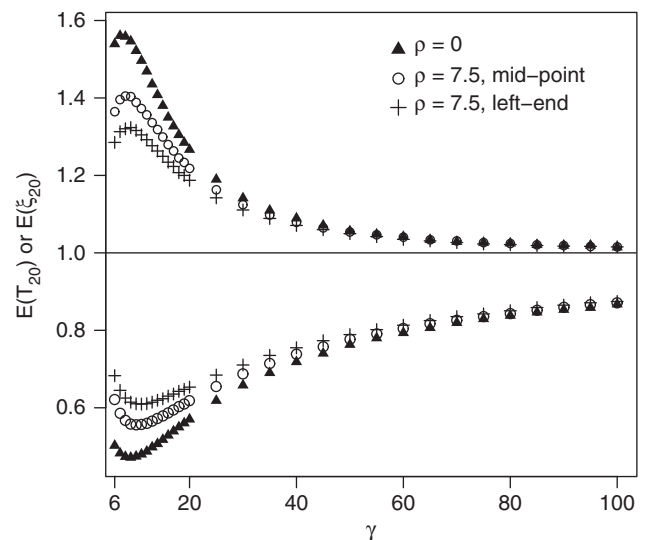

Figure 5 Patterns of diversity as a function of γ. An equilibrium background selection model with a single selected region is considered. The focal region has 3000 sites, with $\theta = 0.005$ per site. Coalescent simulations were used to generate random samples of size 20 under the two levels of recombination ($\rho = NR = 0$ or 7.5). $E(T_{20})$ and $E(\xi_{20})$ were estimated at the mid-point and the left-end of the focal region, and are shown below and above the solid line, respectively.

For Tajima's D, the correlations between values calculated for the non-synonymous and synonymous variants are 0.12 and 0.48, respectively, in the two groups of genes. These increases in correlation in the less constrained group are consistent with the model prediction. The correlation coefficients for Tajima's D are, however, not higher than those for π (cf. Supplementary Figure S4); this is probably due to statistical noise (Tajima's D is statistically more variable, and was calculated using a subset of sites where data were available from all individuals).

Another potential use of the coalescent model is to assist the estimation of the distribution of fitness effects (DFE; Eyre-Walker and Keightley, 2007). Existing methods for estimating the DFE from DNA sequence polymorphisms rely on the assumption that all sites under consideration are at linkage equilibrium (Keightley and Eyre-Walker,

2007; Boyko *et al.*, 2008), although simulations suggest that linkage between sites may have a relatively small impact on their reliability (Eyre-Walker and Keightley, 2009). Nevertheless, these methods have limited ability to quantify the proportion of highly deleterious mutations, as well as their effects on fitness, since these mutations contribute little, if any, to observed polymorphism, and as a result, inferences of their properties are based on extrapolations of the information gathered from the more weakly selected mutations. However, valuable information may be obtained by making use of variability at linked neutral sites.

To illustrate the above point, consider a model where mutations at the selected sites in the focal region share a common γ (Figure 5). The probability that a selected site with $\gamma = 50$ is polymorphic in a sample of size 20 is 0.18%, about 19 times lower than that for a neutrally evolving site. With $\gamma = 100$, the probability is $<0.10\%$. On the other hand, conditioning on a site being polymorphic, the chance that the site is a singleton is 91.6% for $\gamma = 50$ and 95.6% for $\gamma = 100$. Thus, the frequency spectrum at selected sites is dominated by singletons, and is expected to have a very similar appearance for very different values of γ, making it difficult to make reliable inferences. Nonetheless, with $\gamma = 50$, we expect to observe a 22.3% reduction in the level of diversity and a 5.4% excess of singletons at neutral sites in the vicinity of the selected region; with $\gamma = 100$, these values become 12.9 and 1.6%, respectively (Figure 5). Hence, patterns of diversity at linked neutral sites can potentially help to refine estimates of this part of the DFE. However, existing methods often assume that the DFE follows a continuous distribution (for example, a gamma distribution), whereas only discrete distributions can be considered by the coalescent model. Further research is needed to reconcile these two approaches.

DATA ARCHIVING
There were no data to deposit.

CONFLICT OF INTEREST
The author declares no conflict of interest.

ACKNOWLEDGEMENTS

Begun DJ, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**: 519–520.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE *et al.* (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**: e1000083.

Cai JJ, Macpherson JM, Sella G, Petrov DA (2009). Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet* **5**: e1000336.

Charlesworth B (1990). Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet Res* **55**: 199–221.

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

Charlesworth B, Morgan MT, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Charlesworth D, Charlesworth B, Morgan MT (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619–1632.

Comeron JM, Guthrie TB (2005). Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in Drosophila. *Mol Biol Evol* **22**: 2519–2530.

Cutter AD, Choi JY (2010). Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. *Genome Res* **20**: 1103–1111.

Ewens WJ (2004). *Mathematical Population Genetics*, 2nd edn. Springer-Verlag: Berlin.

Eyre-Walker A, Keightley PD (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**: 610–618.

Eyre-Walker A, Keightley PD (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* **26**: 2097–2108.

Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD (2012). Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol Biol Evol* **29**: 675–687.

Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Gordo I, Navarro A, Charlesworth B (2002). Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**: 835–848.

Haddrill PR, Zeng K, Charlesworth B (2011). Determinants of synonymous and nonsynonymous variability in three species of Drosophila. *Mol Biol Evol* **28**: 1731–1743.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G *et al.* (2011). Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.

Hill WG, Robertson A (1966). The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269–294.

Hudson RR (1990). Gene genealogies and the coalescent process. *Oxford Surv Evol Biol* **7**: 1–44.

Hudson RR, Kaplan NL (1995). Deleterious background selection with recombination. *Genetics* **141**: 1605–1617.

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.

Keightley PD, Eyre-Walker A (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.

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

Kimura M, Maruyama T (1966). The mutational load with epistatic gene interactions in fitness. *Genetics* **54**: 1337–1351.

Loewe L, Charlesworth B (2007). Background selection in single genes may explain patterns of codon bias. *Genetics* **175**: 1381–1393.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N *et al.* (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**: e1002326.

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

McVicker G, Gordon D, Davis C, Green P (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**: e1000471.

Nordborg M, Charlesworth B, Charlesworth D (1996). The effect of recombination on background selection. *Genet Res* **67**: 159–174.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009). Pervasive natural selection in the Drosophila genome? *PLoS Genet* **5**: e1000495.

Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA *et al.* (2007). Adaptive genic evolution in the Drosophila genomes. *Proc Natl Acad Sci USA* **104**: 2271–2276.

Slatkin M, Hudson RR (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.

Stephan W (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Phil Trans R Soc B* **366**: 1245–1253.

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

Wakeley J (2008). Conditional gene genealogies under strong purifying selection. *Mol Biol Evol* **25**: 2615–2626.

Walczak AM, Nicolaisen LE, Plotkin JB, Desai MM (2012). The structure of genealogies in the presence of purifying selection: a fitness-class coalescent. *Genetics* **190**: 753–779.

Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.

Zeng K, Charlesworth B (2011). The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics* **189**: 251–266.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)