FUNDAMENTAL CONCEPTS IN GENETICS

Effective population size and patterns of molecular evolution and variation

Brian Charlesworth

Abstract | The effective size of a population, N_e, determines the rate of change in the composition of a population caused by genetic drift, which is the random sampling of genetic variants in a finite population. N_e is crucial in determining the level of variability in a population, and the effectiveness of selection relative to drift. This article reviews the properties of N_e in a variety of different situations of biological interest, and the factors that influence it. In particular, the action of selection means that N_e varies across the genome, and advances in genomic techniques are giving new insights into how selection shapes N_e.

Genetic drift
The process of evolutionary change involving the random sampling of genes from the parental generation to produce the offspring generation, causing the composition of the offspring and parental generations to differ.

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK.
e-mail: Brian.Charlesworth@ed.ac.uk
doi:10.1038/nrg2526
Published online
10 February 2009

The effective size of a population (N_e) is one of several core concepts introduced into population genetics by Sewall Wright, and was initially sketched in his magnum opus, *Evolution in Mendelian Populations*[1]. Its purpose is to provide a way of calculating the rate of evolutionary change caused by the random sampling of allele frequencies in a finite population (that is, genetic drift). The basic theory of N_e was later extended by Wright[2–5], and a further theoretical advance was made by James Crow[6], who pointed out that there is more than one way of defining N_e, depending on the aspect of drift in question. More recently, the theoretical analysis of the effects of demographic, genetic and spatial structuring of populations has been greatly simplified by the use of approximations that resolve drift into processes operating on different timescales[7].

What biological questions does N_e help to answer? First, the product of mutation rate and N_e determines the equilibrium level of neutral or weakly selected genetic variability in a population[8]. Second, the effectiveness of selection in determining whether a favourable mutation spreads, or a deleterious mutation is eliminated, is controlled by the product of N_e and the intensity of selection. The value of N_e therefore greatly affects DNA sequence variability, and the rates of DNA and protein sequence evolution[8].

The importance of N_e as an evolutionary factor is emphasized by findings that N_e values are often far lower than the census numbers of breeding individuals in a species[9,10]. Species with historically low effective population sizes, such as humans, show evidence for reduced variability and reduced effectiveness of selection in comparison with other species[11]. N_e may also vary across different locations in the genome of a species, either as a result of differences in the modes of transmission of different components of the genome (for example, the X chromosome versus the autosomes[12]), or because of the effects of selection at one site in the genome on the behaviour of variants at nearby sites[13]. An important consequence of the latter process is that selection causes reduced N_e in genomic regions with low levels of genetic recombination, with effects that are discernible at the molecular sequence level[14,15]. BOX 1 summarizes the major factors influencing N_e, which will be described in detail below.

In the era of multi-species comparisons of genome sequences and genome-wide surveys of DNA sequence variability, there is more need than ever before to understand the evolutionary role of genetic drift, and its interactions with the deterministic forces of mutation, migration, recombination and selection. N_e therefore plays a central part in modern studies of molecular evolution and variation, as well as in plant and animal breeding and in conservation biology. In this Review, I first describe some basic theoretical tools for obtaining expressions for N_e, and then show how the results of applying these tools can be used to describe the properties of a single population, and how to include the effects of selection. Finally, I describe the effects of structuring of populations by spatial location or by genotype, and discuss the implications of genotypic structuring for patterns of variation and evolution across the genome.

Describing genetic drift and determining N_e

There are three major ways in which genetic drift can be modelled in the simplest type of population, which are outlined below. These theoretical models lead to a general approach that can be applied to situations of greater biological interest, which brings out the utility of the concept of the effective population size.

Poisson distribution
This is the limiting case of the binomial distribution (see next page), valid when the probability of an event is very small. The mean and variance of the number of events are then equal.

Coalescent theory
A method of reconstructing the history of a sample of alleles from a population by tracing their genealogy back to their most recent common ancestral allele.

Coalescence
The convergence of a pair of alleles in a sample to a common ancestral allele, tracing them back in time.

Fast timescale approximation
Used to simplify calculations of effective population size, by assuming that the rate of coalescence is slower than the rate at which alleles switch between different compartments of a structured population as we trace them back in time.

Panmictic
A panmictic population lacks subdivision according to spatial location or genotype, so that all parental genotypes potentially contribute to the same pool of offspring.

The Wright–Fisher population. To see why N_e is so useful, we need to understand how genetic drift can be modelled in the simple case of a Wright–Fisher population[1,16,17]. This is a randomly mating population, consisting of a number of diploid hermaphroditic individuals (N). The population reproduces with discrete generations, each generation being counted at the time of breeding. New individuals are formed each generation by random sampling, with replacement, of gametes produced by the parents, who die immediately after reproduction. Each parent thus has an equal probability of contributing a gamete to an individual that survives to breed in the next generation. If N is reasonably large, this implies a Poisson distribution of offspring number among individuals in the population. A population of hermaphroditic marine organisms, which shed large numbers of eggs and sperm that fuse randomly to make new zygotes, comes closest to such an idealized situation.

With this model, the rate at which genetic drift causes an increase in divergence in selectively neutral allele frequencies between isolated populations, or loss of variability within a population, is given by $1/(2N)$ (BOX 2). An alternative approach, which has a central role in the contemporary modelling and interpretation of data on DNA sequence variation[7,18], is provided by the theory of the coalescent process (the coalescent theory) (BOX 3). Instead of looking at the properties of the population as a whole, we consider a set of alleles at a genetic locus that have been sampled from a population. If we trace their ancestry back in time, they will eventually be derived from the same ancestral allele, that is, they have undergone coalescence (BOX 3). This is obviously closely related to the inbreeding coefficient approach to drift described in BOX 2, and the rate of the coalescent process in a Wright–Fisher population is also inversely related to the population size.

Box 1 | **Factors affecting the effective size of a population**

- Division into two sexes: a small number of individuals of one sex can greatly reduce effective population size (N_e) below the total number of breeding individuals (N).
- Variation in offspring number: a larger variance in offspring number than expected with purely random variation reduces N_e below N.
- Inbreeding: the correlation between the maternal and paternal alleles of an individual caused by inbreeding reduces N_e.
- Mode of inheritance: the N_e experienced by a locus depends on its mode of transmission; for example, autosomal, X-linked, Y-linked or organelle.
- Age- and stage-structure: in age- and stage-structured populations, N_e is much lower than N.
- Changes in population size: episodes of low population size have a disproportionate effect on the overall value of N_e.
- Spatial structure: the N_e determining the mean level of neutral variability within a local population is often independent of the details of the migration process connecting populations. Limited migration between populations greatly increases N_e for the whole population, whereas high levels of local extinction have the opposite effect.
- Genetic structure: the long-term maintenance of two or more alleles by balancing selection results in an elevation in N_e at sites that are closely linked to the target of selection. In contrast, directional selection causes a reduction in N_e at linked sites (the Hill–Robertson effect).

More realistic models of drift. The assumptions of the Wright–Fisher population model do not, however, apply to most populations of biological interest: many species have two sexes, there may be nonrandom variation in reproductive success, mating may not be at random, generations might overlap rather than being discrete, the population size might vary in time, or the species may be subdivided into local populations or distinct genotypes. In addition, we need to analyse the effects of deterministic evolutionary forces, such as selection and recombination, as well as drift.

The effective population size describes the timescale of genetic drift in these more complex situations: we replace $2N$ by $2N_e$, where N_e is given by a formula that takes into account the relevant biological details. Classically, this has been done by calculations based on the variance or inbreeding coefficient approaches[19–23], but more recently coalescent theory has been employed[7]. In general, the use of N_e only gives an approximation to the rate of genetic drift for a sufficiently large population size (such that the square of $1/N$ can be neglected compared with $1/N$), and is often valid only asymptotically, that is, after enough time has elapsed since the start of the process. Exact calculations of changes in variance of allele frequencies or inbreeding coefficient are, therefore, often needed in applications in which the population size is very small or the timescale is short, as in animal and plant improvement or in conservation breeding programmes[19–21,24].

Determining N_e: a general method. Coalescent theory provides a flexible and powerful method for obtaining formulae for N_e, replacing the term involving N in the rate of coalescence in BOX 3 by N_e, which can then be directly inserted in place of N into the results from coalescent theory (BOX 3). A core approach for estimating N_e under different circumstances is outlined briefly below and is discussed in more detail in the following sections of this Review.

This approach involves the structured coalescent process, in which there are several 'compartments' (such as ages or sexes) in the population from which alleles can be sampled[7,25,26]. Alleles are initially sampled from one or more of these compartments, and the probabilities of allele movements to the other compartments, as we go back in time, are determined by the rules of inheritance. A useful simplification is to assume that alleles flow among the different compartments at a much faster rate than the coalescence of alleles: this is termed the fast timescale approximation. This means that we can treat the sampled alleles as coming from the equilibrium state of the process[7,27–31]. This provides a general formula for the rate of coalescence, which is easy to apply to individual cases[7,28–31].

Determining N_e of a single population

The structured coalescent process can be applied to different biologically important scenarios. In this section, I discuss how it can be applied to panmictic populations (BOX 2), with particular reference to the effects on N_e of variation in offspring number among individuals, the

Box 2 | **Using the Wright–Fisher model to describe genetic drift**

Consider the effects of genetic drift on selectively neutral variants, assuming that the population is closed (there is no migration from elsewhere) and panmictic. We also ignore the possibility of mutation. Assume that there are two alternative variants at an autosomal site, A_1 and A_2, with frequencies p_0 and $q_0 = 1 - p_0$ in an initial generation; these might represent two alternative nucleotide pairs at a given site in a DNA sequence, such as GC and AT.

The state of the population in the next generation can then be described by the probability that the new frequency of A_2 is $i/(2N)$, where i can take any value between 0 and $2N$. $2N$ is used because with diploid inheritance there are $2N$ allele copies in N individuals; if the species were haploid, we would use N. The Wright–Fisher model is identical to the classical problem in probability theory of determining the chance of i successes out of a specified number ($2N$) of trials (a success being the choice of A_2 rather than A_1) when the chance of success on a single trial is q. Tossing an unbiased coin $2N$ times corresponds to the case in which $q = 0.5$.

Probability theory tells us that the chances of obtaining i copies of A_2 in the next generation, corresponding to a frequency of $q = i/(2N)$, is given by the binomial distribution[1,16]. The new mean frequency of A_2 is simply q_0, as drift does not affect the mean. But the frequency in any given population will probably change somewhat, becoming $q_0 + \delta q$, where the change δq has variance $V_{\delta q}$, given by:

$$V_{\delta q} = \frac{p_0 q_0}{2N} \qquad (3)$$

After a further generation, the new frequency will be $q_0 + \delta q + \delta q'$, where $\delta q'$ has a mean of zero and a variance of $(p_0 - \delta q)(q_0 + \delta q)/(2N)$, and so on. If we follow a single population, there will be a succession of random changes in q, until eventually A_2 either becomes fixed in the population ($q = 1$) or is lost ($q = 0$).

From equation 3 above, the rate of increase in variance per generation is proportional to $1/(2N)$. This variance can be thought of as measuring the extent of differentiation in allele frequencies between a large set of completely isolated populations, all of which started with the same initial state. Alternatively, it represents the variation in allele frequencies among a set of independent loci within the genome, all with the same initial state.

An alternative way of looking at drift is to use the concept of identity by descent[84,141,142]. Two different allelic copies of a given nucleotide site drawn from a population are identical by descent (IBD) if they trace their ancestry back to a single ancestral copy. The progress of a population towards genetic uniformity is measured by the probability that a pair of randomly sampled alleles are IBD (a value termed the inbreeding coefficient, f), measured relative to an initial generation in which all the alleles in the population are not IBD. Just as for the variance in allele frequency, the inbreeding coefficient increases at a rate that is governed by $1/(2N)$, and the inbreeding coefficient at a given time is equal to the variance divided by $p_0 q_0$ (REFS 1,5). Approach to uniformity thus occurs at the same rate as increase in variance of allele frequencies.

mode of inheritance and the consequence of changes in population size. By looking at real-life data we see that different methods of estimating N_e can give very different answers if the population size has changed greatly.

Outbreeding populations with constant size. First we consider a population with no inbreeding and a Poisson distribution of offspring number. $1/N_e$ for autosomal (A) inheritance and two sexes (m, male; f, female) is given by:

$$\frac{1}{N_{eA}} \approx \frac{1}{4N_m} + \frac{1}{4N_f} \qquad (1)$$

With a 1:1 sex ratio among breeding individuals, the effective size in this case is approximately equal to the total population size ($N = 2N_f = 2N_m$), so that the population then has the same properties as the Wright–Fisher model. But if the numbers of females and males are not the same, the effective size is much less than N. For example, if there is only a small number of breeding males compared with females, the reciprocal of N_m dominates equation 1, and N_e is close to $4N_m$. This reflects the fact that half of the genes in a new generation must come from males, regardless of their numbers relative to females. This situation is approached in populations of farm animals, where artificial insemination is used in selective breeding, causing serious problems with inbreeding[32].

With nonrandom variation in offspring numbers, but with the same variance in offspring number for the two sexes and a 1:1 sex ratio, we have:

$$\frac{1}{N_{eA}} \approx \frac{(2 + \Delta V)}{2N} \qquad (2)$$

An excess variance in offspring numbers compared with random expectation thus reduces N_e below N (REFS 3,4). Conversely, if there is less than random variation, N_e can be greater than N; it equals $2N$ in the extreme case when all individuals have equal reproductive success. This is important for conservation breeding programmes, as it is desirable to maximize N_e in order to slow down the approach to homozygosity[33]. In animals, a major cause of a nonrandom distribution of reproductive success is sexual selection, when males compete with each other for access to mates[34]. Sexual selection is thus likely to have a major effect on N_e, with the magnitude of the effect being dependent on the details of the mating system[35,36].

The effect of inbreeding. An excess of matings between relatives reduces N_e by a factor of $1/(1 + F_{IS})$ (REF. 30), where F_{IS} is the inbreeding coefficient of an individual, caused by an excess frequency over random mating expectation of matings between relatives[37]. N_e is reduced because inbreeding causes faster coalescence of an individual's maternal and paternal alleles compared with random mating[38]. With partial self-fertilization with frequency S in an hermaphrodite population, the equilibrium inbreeding coefficient is $F_{IS} = S/(2 - S)$ (REF. 19). Selfing causes N_e to be multiplied by a factor of $(2 - S)/2$ if there is random variation in offspring number; this approaches 1/2 for 100% selfing[30,38,39].

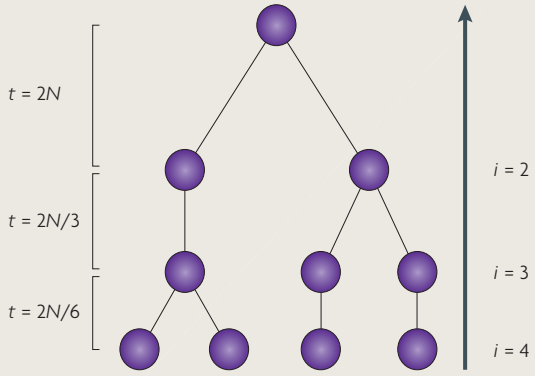
From equation 4 in BOX 3, with N_e replacing N, this result suggests that neutral variability within populations of highly self-fertilizing species, such as *Arabidopsis thaliana* and *Caenorhabditis elegans*, should be reduced to approximately half the value for randomly mating populations of similar size. Indeed, these species do have low levels of genetic variability[40,41] compared with their outcrossing relatives[42,43] (TABLE 1). Additional possible reasons for this low variability are discussed below.

The effects of mode of inheritance. The mode of inheritance can also greatly alter N_e, and hence expected levels of neutral diversity (as shown by the equations in BOX 4). For example, with X-linked inheritance and random mating, a 1:1 sex ratio and Poisson distribution of offspring numbers imply that $N_{eX} = 3N/4$, consistent with the fact that there are only three-quarters as many X chromosomes as autosomes in the population. It is

Binomial distribution
Describes the probability of observing i independent events in a sample of size n, when the probability of an event is p. The mean and variance of the number of events are np and $np(1 - p)$, respectively.

Neutral diversity
Variability arising from mutations that have no effect on fitness.

Box 3 | **The coalescent process**



We consider a sample of alleles at a genetic locus that have been obtained from a population (see the figure; the four bottom circles). For simplicity, assume that no recombination can occur in the locus, as would be true for a mitochondrial genome or Y chromosome, or for a nuclear gene in a region of a chromosome with severely reduced recombination. If we trace the ancestry of the alleles back in time (upward arrow), two of the alleles in the sample will be seen to be derived from the same ancestral allele — they have coalesced at a time in the history of the population when the other two alleles still trace back to two distinct alleles. At this time, there are three distinct alleles from which the sample is descended ($i = 3$). If we continue back in time, the ancestry of the alleles in the sample follows a bifurcating tree, in which the time (t) between successive nodes (points of branching) is dependent on $2N$ and the number of alleles that are present at the later node; with i alleles, the expected time to a coalescent event that generate $i - 1$ alleles is $4N/i(i - 1)$ (REFS 7,18,112,143). This assumes that N is sufficiently large that, at most, one coalescent event can occur in a given generation. The time itself follows an exponential distribution, with a standard deviation equal to the mean. In the figure, t represents the expected times at which the successive coalescent events occur in a Wright–Fisher population, corresponding to the numbers of distinct alleles, i, on the right.

This description of a gene tree is purely theoretical, as gene trees cannot be observed directly. However, the results are relevant to data on population samples, because variation in a sample of allelic sequences reflects mutations that have arisen in different branches of the tree since the most recent common ancestor. To model a sample, we simply allow mutations to occur on the lineages in the gene tree. The simplest model to use is the infinite sites model: the mutation rate probability per generation per site is u, and u is assumed to be low, so that at most one mutation arises per site in the tree[7,18,112,144].

This allows derivations of formulae to predict the values of commonly used measures of variability such as the nucleotide site diversity, that is, the frequency with which a pair of randomly sampled alleles differ at a given nucleotide site. Consider a given pair of alleles taken randomly from the sample. There is a time (t) connecting each of them to their common ancestor. They will be identical at a site if no mutation has arisen over the time separating them from each other, which is $2t$. The probability that a mutation has arisen at that site, and caused them to differ in state, is $2tu$. From the above considerations, t has an expected value of $2N$, so that the net probability of a difference in state at a given nucleotide site is $4Nu$. Averaging over all pairs of alleles in the sample, and over a large number of sites, gives the expected value of the nucleotide site diversity for a sample (π):

$$\pi = 4Nu \qquad (4)$$

In addition to generating simple and useful expressions for the expected level of variability in a sample from a population, coalescent theory allow the computation of the probability distributions of statistics that describe the frequencies of variants in the sample. This permits statistical tests to be applied to data, to test whether the assumptions of the standard model (demographic equilibrium and neutrality) are violated[7,18,112].

therefore common practice to adjust diversity estimates for X-linked loci by multiplying by 4/3 when comparing them with data for autosomal genes; see REF 44 for an example. But the formulae in BOX 4 show that this is an over-simplification. If there is strong sexual selection among males, the effective size for X-linked loci can approach or even exceed that for autosomal loci. N_{eX}/N_{eA} has an upper limit of 1.125; the reason that this ratio can exceed 1 is that autosomes are transmitted through males more often than X chromosomes, and the males' effective population size is small. Surveys of variability in the putatively ancestral African populations of *Drosophila melanogaster* show that the mean silent site nucleotide diversity for X-linked loci is indeed slightly higher than for autosomal loci[45–47], consistent with the operation of very strong sexual selection, although other factors might also be involved[46,48].

For ZW sex determination systems, the predicted difference between males and females is reversed. For Z-linked inheritance, N_{eZ}/N_{eA} with strong sexual selection can be as low as 9/16. Data on DNA sequence variability in introns in domestic chickens gave a ratio of Z-linked to autosomal variability of 0.24, even lower than expected under strong sexual selection[49]. For organelle inheritance, with strictly maternal transmission, N_e is one-quarter of the autosomal value with random variation in offspring number, but is expected to be much larger with sexual selection (BOX 4).

Age- and stage-structure. To calculate N_e for populations in which reproductive individuals have a range of ages or developmental stages, the fast timescale approximation can again be applied. In this case, alleles flow between ages or stages as well as sexes. Expressions can be derived for N_e in an age- or stage-structured population

Table 1 | **Effective population size (N_e) estimates from DNA sequence diversities**

Species	N_e	Genes used	Refs
Species with direct mutation rate estimates			
Humans	10,400	50 nuclear sequences	145
Drosophila melanogaster (African populations)	1,150,000	252 nuclear genes	108
Caenorhabditis elegans (self-fertilizing hermaphrodite)	80,000	6 nuclear genes	41
Escherichia coli	25,000,000	410 genes	146
Species with indirect mutation rate estimates			
Bonobo	12,300	50 nuclear sequences	145
Chimpanzee	21,300	50 nuclear sequences	145
Gorilla	25,200	50 nuclear sequences	145
Gray whale	34,410	9 nuclear gene introns	147
Caenorhabditis remanei (separate sexes)	1,600,000	6 nuclear genes	43
Plasmodium falciparum	210,000–300,000	204 nuclear genes	148

For data from genes, synonymous site diversity for nuclear genes was used as the basis for the calculation, unless otherwise stated.

reproducing at discrete time-intervals, such as annually breeding species of birds or mammals[29–31,50–52]. The results show that N_e is usually considerably less than the number of breeding individuals present at any one time. There is, however, no satisfactory treatment of populations in which individuals reproduce more or less continuously, such as humans and many tropical species[52].

The effect of changes in population size. It is also possible to model changes over time in population size N, while otherwise retaining the Wright–Fisher model[3,4,53]. The expected coalescence time is then similar to that with constant population size, that is, approximately $2N_H$, where N_H is the harmonic mean of N over the set of generations in question (the reciprocal of the mean of the reciprocals of the values of N). This allows the use of N_H instead of N in the equation for expected neutral diversity (BOX 3). For more complex population structures, we can replace the N values for each generation by the corresponding N_e values from BOX 4, provided that the flow between different compartments equilibrates over a short timescale compared with changes in population size.

A population that has recently grown from a much smaller size, such as a population that has recovered from a bottleneck associated with colonization of a new habitat, will thus have a much lower effective size than one that has always remained at its present size, as the harmonic mean is strongly affected by the smallest values in the set[54]. There is increasingly strong evidence for such bottleneck effects in both human[55,56] and *D. melanogaster* populations[46,48,57] that have moved out of Africa.

Estimating N_e for natural and artificial populations. It is obviously of importance to have estimates of N_e, both for practical purposes, such as designing conservation or selective breeding programmes, and for interpreting data on DNA sequence variation and evolution. This can be done simply by using demographic information and substituting into equations of the type shown in BOX 4 (REFS 9,10). More recently, two different approaches that use information on genetic markers have been employed. First, N_e for a large natural population can be estimated from silent nucleotide site diversities, as diversity at equilibrium between drift and mutation depends on the product of mutation rate per nucleotide site, u, and N_e (replacing N by N_e in equation 4 in BOX 3). If the mutation rate is known, either from a direct experimental estimate or from data on DNA sequence divergence between species with known dates of separation, N_e can be estimated as $\pi/(4u)$, where π is nucleotide site diversity. Some examples are shown in TABLE 1. Second, for very small populations, such as those used in animal and plant breeding or in the captive breeding of endangered species, N_e can be estimated from observed changes between generations in the frequencies of putatively neutral variants[9,58–60].

As might be expected from the theoretical results, effective population sizes are often found to be much lower than the observed numbers of breeding individuals in both natural and artificial populations[9,10,61]. The human population, for example, is estimated from DNA sequence variability to have an N_e of 10,000 to 20,000, because of its long past history of small numbers of individuals and relatively recent expansion in size[55,62]. Larger population sizes in the past other than for extant populations have, however, sometimes been inferred from diversity estimates; for example, Atlantic whales, probably reflecting the devastating effects of whaling on their population sizes[63].

The above two genetic methods of estimating N_e can therefore yield very different results if there have been large changes in population size, because the first approach relates to the harmonic mean value of population size over the long period of time required for diversity levels to equilibrate, and the second to the present day population size. A large increase in population size, as in the case of humans, means that the N_e estimated

Box 4 | **Effective population sizes for some common situations**

Using the fast timescale approximation described in the text, formulae for N_e can be derived for various types of discrete generation populations. These provide insights into the effects of different demographic and genetic factors.

Autosomal inheritance:

$$\frac{1}{N_{eA}} \approx \frac{(1+F_{IS})}{4}\left\{\frac{1}{N_f}+\frac{1}{N_m}+\frac{(1-c)^2\,\Delta V_f}{N_f}+\frac{c^2\,\Delta V_m}{N_m}\right\} \quad (5)$$

X-linked inheritance (*Z*-linked inheritance, with female heterogamety, is described by interchanging female and male subscripts, *f* and *m*):

$$\frac{1}{N_{eX}} \approx \frac{1}{9}\left\{\frac{4\,(1+F_{IS})}{N_f}+\frac{2}{N_m}+\frac{4\,(1+F_{IS})\,(1-c)^2\,\Delta V_f}{N_f}+\frac{2c^2\,\Delta V_m}{N_m}\right\} \quad (6)$$

Y-linked inheritance (*W*-linked inheritance, with female heterogamety, is described by replacing the male subscripts, *m*, with the female subscript, *f*):

$$\frac{1}{N_{eY}} \approx \frac{2\,(1+c^2\,\Delta V_m)}{N_m} \quad (7)$$

Maternally transmitted organelles:

$$\frac{1}{N_{eC}} \approx \frac{2\,(1+(1-c)^2\,\Delta V_f)}{N_f} \quad (8)$$

Discrete generations with constant population size are assumed. N_f and N_m are the numbers of breeding females and males, respectively; *c* is the fraction of males among breeding individuals, that is, $c = N_m/(N_f + N_m)$; ΔV_f and ΔV_m are the excesses of the variances in offspring numbers over the Poisson values for females and males, respectively; F_{IS} is the inbreeding coefficient within the population caused by an excess of matings between relatives over random mating expectation[5,19]. Equations are taken from REF. 30.

from diversity data might be irrelevant to estimates of future changes caused by drift. Care must therefore be taken to apply estimates of N_e only to situations in which they are appropriate.

The simultaneous effects of selection and drift

Although the models outlined above indicate how N_e can be used in models of genetic drift in panmictic populations, in order to understand evolutionary processes more fully we need to include the effects of selection into the models. The effects of selection can be most easily studied by using diffusion equations[16,19,23,64].

Diffusion equations. These provide approximation for the rate of change in the probability of allele frequency *q* at time *t*. For diffusion approximations to be valid, the effects of both drift and deterministic forces must both be weak. The evolutionary process is then completely determined by the mean and variance of the change in allele frequency per generation, $M_{\delta q}$ and $V_{\delta q}$, respectively[19,23,64].

The effects of drift in situations can be modelled by $V_{\delta q} = pq/(2N_e)$, where N_e replaces *N* for non-Wright–Fisher populations in equation 3 in BOX 2. In this context, N_e is known as the variance effective size. Intuitively, it might seem that we can just use the expressions for N_e derived for the neutral coalescent process. However, there are situations in which this is not correct[6,65]. If the population size changes between generations, the rate of the coalescent process depends on the population size in the parental generation, whereas the change in variance depends on the size of the offspring generation. In addition, the binomial expression for $V_{\delta q}$ (equation 3 in BOX 2) is only an approximation when there is selection or when the population does not follow the Wright–Fisher model[22,66,67]. The coalescent N_e that we have used should, however, provide a good approximation to the variance N_e when all evolutionary forces are weak and the population size is constant.

Probability of fixation of a new mutation. A major conclusion from the use of diffusion equations is that the effectiveness of a deterministic force is controlled by the product of N_e and the measure of its intensity[19,23,64]. This principle is exemplified by the probability of fixation of a new mutation, denoted here by *Q*[8,16,17,64,68] (BOX 5). This is probably the most useful index of the effectiveness of selection versus genetic drift. For a deleterious mutation (with a selection coefficient (*s*) less than 0), *Q* is not much below the neutral value when $-N_e s \leq 0.25$; a deleterious mutation has almost no possibility of becoming fixed by drift once $-N_e s > 2$. For a favourable mutation, if $N_e s \leq 0.25$, *Q* behaves close to neutrally; once $N_e s > 1$, *Q* is close to that for an infinitely large population, that is, $Q = s(N_e/N)$.

A reduction in N_e below *N* reduces the efficacy of selection compared with a Wright–Fisher population of size *N*. This result applies to a wide variety of causes of reduced N_e, as we shall see in the next section. Given the large values of long-term N_e in TABLE 1, weak selection can therefore be very effective in evolution, as was strongly emphasized by Fisher[68]. Indeed, studies of polymorphisms at the sequence level find selection coefficients of a few multiples of $1/N_e$ for many deleterious polymorphic amino-acid variants in human and *Drosophila* populations[56,69–71]; these are sufficient to prevent them becoming fixed in the population with any significant probability. Variants at synonymous or non-coding sites are generally under much weaker selection, with selection coefficients in the order of $1/N_e$ or less[72–75]; this means that drift and mutation as well as selection have a considerable influence on the states of such sites[8,76,77]. There is increasing evidence that the rate of evolution of protein sequences is affected by differences in N_e in the way predicted by theory[11,14,15,78–82].

Determining N_e of a structured population

Having discussed the issue of how to determine the effective size of a population and considered the effects of selection in panmictic populations, the final section of this Review examines how to do this when the population is divided into geographically or genetically defined subpopulations. This is a field that has experienced rapid development in the past few years. New theoretical approaches that use fast timescale approximations have been applied to both spatial and genetic structuring of populations. There is also a growing appreciation of the fact that the genetic structuring of populations with respect to genotypes with different fitnesses implies the existence of differences in N_e values among different parts of the genome of the same species.

Heterogamety
The presence of two different sex-determining alleles or chromosomes in one of the two sexes.

Selection coefficient
(*s*). The effect of a mutation on fitness, relative to the fitness of wild-type individuals. With diploidy, this is measured on mutant homozygotes.

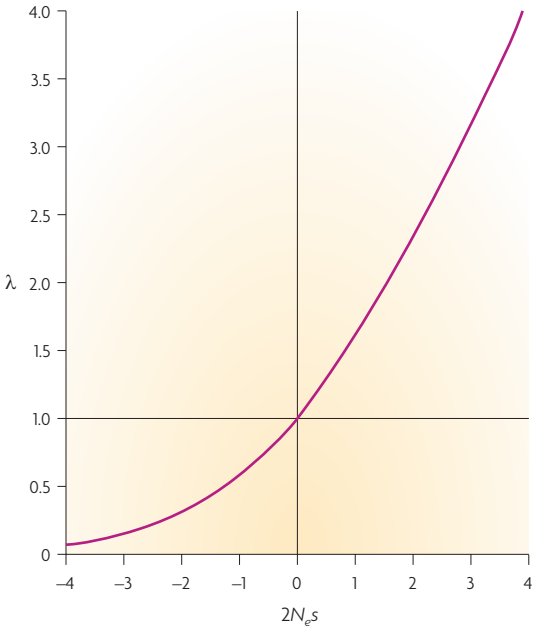
Box 5 | **Fixation probabilities**

The probability of fixation of a mutation is the chance that it will spread through the population and become fixed. In a finite population, even deleterious mutations can become fixed by drift, and favourable ones can be lost. The results of some fairly complex calculations[17,19,64] can be illustrated with the simple case of selection at a biallelic autosomal locus with semi-dominance, such that the relative fitnesses of A_1A_1, A_1A_2 and A_2A_2 are 1, 1 + 0.5s and 1 + s, respectively. s is the selection coefficient, and is negative if A_2 is deleterious and positive if it is advantageous.

If the population size is *N*, and the effective population size is N_e, the probability that a newly arisen mutation to A_2 from A_1 survives in the population and eventually replaces A_1 is given by:

$$Q \approx \frac{N_e s}{N} \frac{1}{\{1 - \exp(-2N_e s)\}} \quad (9)$$

The dependence of *Q* on $N_e s$ is illustrated in the figure. λ is the fixation probability of a semi-dominant mutation, expressed relative to the neutral value (1/2*N*). This is given by *Q* (from the equation above) divided by 1/(2*N*). This also represents the evolutionary rate of substitution of mutations with selection coefficient s, relative to the rate for neutral mutations[8].



The effects of spatial structure on neutral variation. Spatial structure was first studied by classical population genetic methods, extending the methods of BOX 2 to include the effects of geographic subdivision of a metapopulation into partially isolated, local populations[5,83–85]. More recently, the study of neutral variability in a spatially structured population has been simplified by extending the structured coalescent approach described above to a metapopulation consisting of a set (*d*) of discrete local populations (demes) that are interconnected by migration[7] or that are affected by local extinctions of demes and recolonization[7,86].

A useful result applies to the case of 'conservative' migration, that is, when migration among demes leaves their relative sizes unchanged; the mean allele frequency across demes is also unchanged[27,87,88] (the classical island and stepping stone models[83,89,90] are examples of this). Provided that all demes experience some migration events, the mean coalescence time for a pair of alleles sampled from the same deme (T_S) is given by the sum of the effective population sizes over all demes (N_{eT}), so that the mean within-deme nucleotide site diversity is the same as for a panmictic population with this effective population size. This suggests that the mean within-deme nucleotide site diversity for a species is the most appropriate measure to compare the properties of different species.

We might also be interested in describing aspects of variability such as the total amount of variability in a metapopulation, as measured by the mean pairwise nucleotide site diversity among a pair of alleles sampled at random from the metapopulation (π_T) and the corresponding mean coalescence time (T_M) corresponding to what we can call the total effective size of the metapopulation: $N_{eM} = T_M/2$. In contrast to T_S, the value of T_M is highly dependent on the details of the migration process, and can be greatly increased when migration is restricted.

For more general migration models, it is hard to derive an expression for T_M. However, when the number of demes is very large, it is approximately the same as the mean coalescence time for a pair of alleles sampled from two distinct, randomly chosen populations. Wakeley and his collaborators have shown that this large deme number approximation often yields a simple approximate general formula for T_M[7,86,91–93].

Standard tests for departures from neutral equilibrium utilize patterns of variability to detect departures from those predicted by the standard coalescent model; tests of this kind are widely used in studies of DNA sequence variation[7,18]. If such departures are detected, the occurrence of selection or of demographic events, such as changes in population size, is implied. In the case of a metapopulation with a large number of demes, if a sample of *k* alleles is taken by sampling each allele from a separate population, these obey the same coalescent process as alleles sampled from a panmictic population, described in BOX 3. Tests of this kind for a metapopulation are thus best carried out by sampling only one allele from a given population. Similar results also apply to measures of linkage disequilibrium in spatially structured populations. If a single haplotype is sampled from each local population studied, under conservative migration the expected level of linkage disequilibrium between a pair of sites with recombination frequency *r* is controlled by $4N_{eT}r$ in the same way as by $4N_e r$ in the case of a panmictic population[7,94].

The effects of spatial structure on variants under selection. We can also ask how to determine the fixation probability of a mutation under selection in a

Metapopulation
A population consisting of a set of spatially separate local populations.

metapopulation. With semi-dominant or haploid selection (BOX 5), the fixation probability of a new mutation in a structured population consisting of a set of Wright–Fisher populations connected by conservative migration is determined by the product of the selection coefficient and N_{eT} in the same way as by $N_e s$ in a single, panmictic population[87,95,96]. Recent work suggests that an approximate diffusion equation can be derived for more general selection and migration models, using the large deme number approximation just discussed[97–100]. This is useful, as it implies that spatial structure might not have much effect on the fixation probabilities of weakly selected mutations, which are likely to have intermediate dominance coefficients[101], so that the standard models of molecular evolution apply even to highly subdivided populations. Predictions of the effects of differences in effective population sizes on rates of sequence evolution for species[79,81] should therefore use estimates of N_e based on mean within-population diversities.

With dominance, however, population structure can cause important departures from the panmictic results[98,102,103]. Fixation probabilities are reduced for recessive or partly recessive deleterious mutations, and increased for recessive or partly recessive advantageous mutations, relative to the value for a panmictic population with an effective size of N_{eS}. The reverse is true for dominant or partially dominant mutations. The overall effect of population subdivision on the rate of evolution thus depends on both the level of dominance of new mutations, and on the extent to which advantageous or deleterious mutations contribute.

The effects of genetic structure. Investigations of DNA sequence variability have shown that presumptively neutral diversity is not constant across the genome. For example, silent site DNA sequence variability is elevated in the neighbourhood of the highly polymorphic major histocompatibility (MHC) loci of mammals[104], and of the self-incompatibility (SI) loci of plants[105,106]. Conversely, in *D. melanogaster*[14,107,108], humans[109] and some plant species[110], silent site variability correlates positively with the local rate of genetic recombination, and is extremely low in regions where there is little or no recombination. In addition, as already noted, species or populations with high levels of inbreeding often exhibit reduced levels of variation compared with outcrossing relatives[40,41,110], to a much greater extent than the two-fold reduction predicted on a purely neutral model (see above).

The most likely explanation for these patterns, with the possible exception of human populations[109,111], is that N_e is affected by selection occurring at closely linked sites or, in inbreeding populations, sites that rarely recombine with physically distant targets of selection because of the reduced evolutionary effectiveness of recombination in a highly homozygous genome[28]. The concepts and methods used to study the effects of spatial structuring of populations can be used to understand stable genetic structure, whereby different genotypes are maintained in the population, either by long-term balancing selection, or by recurrent mutation to deleterious alleles.

The effects of balancing selection. Long-term balancing selection refers to the situation in which two or more variants at a locus are maintained in the population by forms of selection such as heterozygote advantage or frequency-dependent selection, for much longer than would be expected under neutrality. There is clear evidence for such selection in the cases of the MHC and SI loci mentioned above. What is the effect of balancing selection on neutral variability at linked sites? Consider an autosomal site with two variants, A_1 and A_2, maintained by balancing selection in a randomly mating population with effective population size N_e. A neutral site recombines with the A site at rate r. The flow of neutral variants by recombination between the haplotypes carrying A_1 and A_2 is similar to conservative migration between demes[25,28,112]. High equilibrium levels of differentiation between A_1 and A_2 haplotypes are expected at closely linked neutral sites, for which $N_e r$ is much greater than 1, that is, in the situation equivalent to low migration. This is reflected in a local elevation in the effective population size, equivalent to the elevation of N_{eM} over N_{eT}, producing a local peak of diversity close to the target of balancing selection, as is observed in the cases mentioned above. Coalescence times in this case can be much greater than the time during which the species has existed[113]. Neutral variants that distinguish the selected alleles might then persist across the species boundaries. This is called *trans*-specific polymorphism, and is seen, for example, in the SI polymorphisms of plants[114].

This suggests that polymorphisms maintained by long-term balancing selection could be discovered by scanning the genome for local peaks of silent site diversity and/or polymorphisms that are shared between species. Such scans using the human and chimpanzee genomes have so far been largely negative, suggesting that there are rather few cases of long-term balancing selection[115,116], although some convincing examples have been discovered[117].

Background selection and other Hill–Robertson effects. Another important type of genetic structuring in populations is caused by deleterious alleles maintained by recurrent mutation[118]. These reduce neutral diversity at linked sites because the elimination of a deleterious mutation carried on a particular chromosome also lowers the frequencies of any associated neutral or nearly neutral variants. This process of background selection is one example of the general process known as the Hill–Robertson effect; see REF. 13 for a recent review. This can be understood in terms of N_e as follows. Selection creates heritable variance in fitness among individuals, which reduces N_e (REF. 119). A site that is linked to a selected variant experiences an especially marked reduction in its N_e, because close linkage maintains the effects for many generations[120,121]. In addition to reducing levels of variability, this reduction in N_e impairs the efficacy of selection (see the discussion of fixation probability above). This probably accounts for the observation that the level of adaptation at the sequence level, as well as sequence diversity, often seems to be reduced in low recombination regions of the *Drosophila* genome[14,15,82,122–124].

Semi-dominant or haploid selection
With a diploid species, semi-dominant selection occurs when the fitness of the heterozygote for a pair of alleles is intermediate between that of the two homozygotes; haploid selection applies to haploid species, and is twice as effective as semi-dominant selection with the same selection coefficient.

Dominance coefficient
(h). Measures the extent to which the fitness of a heterozygote carrier of a mutation is affected, relative to the effect of the mutation on homozygous carriers.

Heterozygote advantage
The situation in which the fitness of a heterozygote for a pair of alleles is greater than that of either homozygote. This maintains polymorphism.

Frequency-dependent selection
Situations in which the fitnesses of genotypes are affected by their frequencies in the population. Polymorphism is promoted when fitness declines with frequency.

Background selection
The process by which selection against deleterious mutations also eliminates neutral or weakly selected variants at closely linked sites in the genome.

Hill–Robertson effect
The effect of selection on variation at one location in the genome and on evolution at other, genetically linked sites.

Selective sweep
The process by which a new favourable mutation becomes fixed so quickly that variants that are closely linked to it, and that are present in the chromosome on which the mutation arose, also become fixed.

Another important example of a Hill–Robertson effect is the effect on linked sites of the spread of a selectively favourable mutation. This was called a hitchhiking event by Maynard Smith and Haigh[125], and is now often referred to as a selective sweep[126]. The expected reduction in N_e caused by a single selective sweep is very sensitive to the ratio r/s, where s is the selective advantage to the favourable mutation and r is the frequency of recombination between this mutation and the site whose N_e is being considered, and the reduction in N_e is small unless r/s is much lower than 1 (REFS 125,127). This effect is transient, in the absence of further sweeps in the same region, and resembles the effect of a population bottleneck, as variability will start to recover once the favourable mutation has become fixed[128,129].

The selective sweep model can be extended to allow a steady rate of substitution of favourable variants, at sites scattered randomly over the genome[130–133]. Using empirical estimates of the proportion of amino-acid divergence between species that is due to positive selection, this model provides a good fit to data on sequence variability in *D. melanogaster*[134]. N_e for a typical locus seems to be reduced by a few per cent as a result of ongoing adaptive substitutions of amino-acid mutations. The abundance of weakly selected deleterious amino-acid variants in *Drosophila* populations seems to be sufficiently high for background selection to further reduce N_e for genes with normal levels of recombination by a few per cent[135].

Hill–Robertson effects mean that N_e for a particular location in the genome is highly dependent on its recombinational environment, and that no region is entirely free of the effects of selection at nearby sites, even in genomic regions with normal levels of recombination. Large genomic regions that lack recombination, such as the Y chromosome and asexual or highly self-fertilizing species, are expected to experience the most extreme reductions in N_e (REFS 118,136). This probably accounts for the evolutionary degeneration of Y chromosomes[123,124,137], and the lack of evolutionary success of most asexual and highly inbreeding species[138,139].

Conclusions

From a modest beginning, when Sewall Wright dealt with the process of genetic drift in a population with two sexes, the concept of effective population size has been extended to the status of a unifying principle that encompasses the action of drift in almost any imaginable evolutionary scenario. Over that time, there has been a considerable shift in theoretical methodology, with current formulations using the powerful technology of coalescent theory, and approximations based on separating drift into processes acting on different timescales.

One important advance is that we now have a much clearer appreciation of the role of selection in shaping the effective population size at genetically linked sites than we did 10 years ago. Already, we can be fairly sure that no nucleotide in the compact genome of an organism such as *D. melanogaster* is evolving entirely free of the effects of selection on its effective population size; it will be of great interest to see whether this applies to species with much larger genomes, such as humans, when we make use of the avalanche of data on DNA sequence variation and evolution that will be produced by new sequencing technologies.

However, it is important to note that N_e has some limitations as a tool for understanding patterns of evolution and variation. It is extremely useful for describing expected levels of genetic diversity, and for evaluating the effects of different factors on the efficiency of selection. But certain aspects of genetic variability, such as the distribution of frequencies of individual nucleotide variants across different sites, cannot simply be described in terms of N_e. A given reduction in variability caused by a population bottleneck, a selective sweep or background selection might well be associated with different variant frequency distributions, and so cannot be described by a simple reduction in N_e (REFS 128,129,136,140). Models that describe all aspects of the data are needed in these cases; the challenge is to extend existing models to include increasingly refined estimates of parameters, such as the incidence of selective sweeps and the distribution of selection coefficients against weakly deleterious mutations, into models that can be tested against the data.

1. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931). **A classic founding paper of theoretical population genetics, which introduces the concept of effective population size.**
2. Wright, S. Inbreeding and homozygosis. *Proc. Natl Acad. Sci. USA* **19**, 411–420 (1933).
3. Wright, S. Size of population and breeding structure in relation to evolution. *Science* **87**, 430–431 (1938).
4. Wright, S. *Statistical Genetics in Relation to Evolution (Actualites Scientifiques et Industrielles, 802: Exposés de Biométrie et de la Statistique Biologique. XIII)* 5–64 (Hermann et Cie, Paris, 1939).
5. Wright, S. *Evolution and the Genetics of Populations* Vol. 2 (Univ. Chicago Press, Chicago, Illinois, 1969).
6. Crow, J. F. in *Statistics and Mathematics in Biology* (eds Kempthorne, O., Bancroft, T. A., Gowen, J. W. & Lush, J. L.) 543–556 (Iowa State Univ. Press, Ames, Iowa, 1954).
7. Wakeley, J. *Coalescent Theory. An Introduction* (Ben Roberts, Greenwood Village, Colorado, 2008). **A broad-ranging treatment of coalescent theory and its use in the interpretation of data on DNA sequence variability.**
8. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983). **A somewhat partisan account of how population genetics theory can be used to interpret data on molecular evolution. It provides an excellent summary of the use of the concept of effective population size, and describes results from the use of diffusion equations.**
9. Crow, J. F. & Morton, N. E. Measurement of gene-frequency drift in small populations. *Evolution* **9**, 202–214 (1955).
10. Frankham, R. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.* **66**, 95–107 (1995). **This reviews evidence for much lower effective population sizes than census sizes in natural populations.**
11. Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**, 2142–2149 (2002).
12. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nature Rev. Genet.* **7**, 645–653 (2006).
13. Comeron, J. M., Williford, A. & Kliman, R. M. The Hill–Robertson effect: evolutionary consequences of weak selection in finite populations. *Heredity* **100**, 19–31 (2008). **A review of the theory and data on the effects of selection at one genomic site on variability and evolution at other sites in the genome.**
14. Presgraves, D. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15**, 1651–1656 (2005). **Reviews data supporting a correlation between recombination rate and neutral or nearly neutral variability in *D. melanogaster*, and presents evidence for reduced efficacy of selection when recombination rates are low.**
15. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* **24**, 114–123 (2008).
16. Fisher, R. A. On the dominance ratio. *Proc. Roy. Soc. Edinb.* **52**, 312–341 (1922).
17. Fisher, R. A. The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinb.* **50**, 205–220 (1930).
18. Hein, J., Schierup, M. H. & Wiuf, C. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, Oxford, 2005).

19. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
20. Caballero, A. Developments in the prediction of effective population size. *Heredity* **73**, 657–679 (1994).
21. Wang, J. L. & Caballero, A. Developments in predicting the effective size of subdivided populations. *Heredity* **82**, 212–226 (1999).
22. Nagylaki, T. *Introduction to Theoretical Population Genetics* (Springer, Berlin, 1992).
23. Ewens, W. J. *Mathematical Population Genetics. Theoretical Introduction* Vol. 1 (Springer, New York, 2004).
24. Vitalis, R. Sex-specific genetic differentiation and coalescence times: estimating sex-biased dispersal rates. *Mol. Ecol.* **11**, 125–138 (2002).
25. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
26. Hey, J. A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theor. Pop. Biol.* **39**, 30–48 (1991).
27. Nagylaki, T. The strong-migration limit in geographically structured populations. *J. Math. Biol.* **9**, 101–114 (1980).
28. Nordborg, M. Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514 (1997).
29. Rousset, F. Genetic differentiation in populations with different classes of individuals. *Theor. Pop. Biol.* **55**, 297–308 (1999).
30. Laporte, V. & Charlesworth, B. Effective population size and population subdivision in demographically structured populations. *Genetics* **162**, 501–519 (2002).
This uses the fast timescale approximation to provide a general framework for deriving formulae for effective population size.
31. Nordborg, M. & Krone, S. M. in *Modern Developments in Population Genetics. The Legacy of Gustave Malécot.* (eds Slatkin, M. & Veuille, M.) 194–232 (Oxford Univ. Press, Oxford, 2002).
32. Brotherstone, S. & Goddard. Artificial selection and maintenance of genetic variance in the global dairy cow population. *Phil. Trans. R. Soc. B* **360**, 1479–1148 (2005).
33. Frankham, R., Ballou, J. D. & Briscoe, D. A. *Introduction to Conservation Genetics* (Cambridge Univ. Press, Cambridge 2002).
34. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, New Jersey, 1994).
35. Nunney, L. The influence of age structure and fecundity on effective population size. *Proc. Roy. Soc. Lond. B* **246**, 71–76 (1991).
36. Nunney, L. The influence of mating system and overlapping generations on effective population size. *Evolution* **47**, 1329–2341 (1993).
37. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).
38. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185–1195 (1997).
39. Pollak, E. On the theory of partially inbreeding populations. I. Partial selfing. *Genetics* **117**, 353–360 (1987).
40. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, 1289–1299 (2005).
41. Cutter, A. D. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* **172**, 171–184 (2005).
42. Wright, S. J. *et al.* Testing for effects of recombination rate on nucleotide diversity in natural populations of *Arabidopsis lyrata*. *Genetics* **174**, 1421–1430 (2006).
43. Cutter, A., Baird, S. E. & Charlesworth, D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174**, 901–913 (2006).
44. Moriyama, E. N. & Powell, J. R. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**, 261–277 (1996).
45. Andolfatto, P. Contrasting patterns of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster* and *D. simulans*. *Mol. Biol. Evol.* **18**, 279–290 (2001).
46. Hutter, S., Li, H. P., Beisswanger, S., De Lorenzo, D. & Stephan, W. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide nucleotide polymorphism data. *Genetics* **177**, 469–480 (2007).
47. Singh, N. D., Macpherson, J. M., Jensen, J. D. & Petrov, D. A. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. *BMC Evol. Biol.* **7**, 202 (2007).
48. Pool, J. E. & Nielsen, R. The impact of founder events on chromosomal variability in multiply mating species. *Mol. Biol. Evol.* **25**, 1728–1736 (2008).
49. Sundström, H., Webster, M. T. & Ellegren, H. Reduced variation on the chicken Z chromosome. *Genetics* **167**, 377–385 (2004).
50. Felsenstein, J. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68**, 581–597 (1971).
51. Charlesworth, B. *Evolution in Age-structured Populations* 2nd edn (Cambridge Univ. Press, Cambridge 1994).
52. Charlesworth, B. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet. Res.* **77**, 153–166 (2001).
53. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **143**, 579–587 (1991).
54. Wright, S. Breeding structure of species in relation to speciation. *Am. Nat.* **74**, 232–248 (1940).
55. Voight, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA* **102**, 18508–18513 (2005).
56. Boyko, A. *et al.* Assessing the evolutionary impact of amino-acid mutations in the human genome. *PLoS Genet.* **5**, e1000083 (2008).
57. Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* **15**, 790–799 (2005).
58. Wang, J. Estimation of effective population sizes from data on genetic markers. *Phil. Trans. R. Soc. B* **360**, 1395–1409 (2005).
A review of methods for using information on genetic variants in populations to estimate effective population size.
59. Waples, R. S. & Yokota, M. Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**, 219–233 (2007).
60. Jorde, P. E. & Ryman, N. Unbiased estimator for genetic drift and effective population size. *Genetics* **177**, 927–935 (2007).
61. Coyer, J. A., Hoarau, G., Sjotun, K. & Olsen, J. L. Being abundant is not enough; a decrease in effective size over eight generations in a Norwegian population of the seaweed, *Fucus serratus*. *Biol. Lett.* **4**, 755–757 (2008).
62. Wall, J. D. & Przeworski, M. When did the human population size start increasing? *Genetics* **155**, 1865–1874 (2000).
63. Roman, J. & Palumbi, S. R. Whales before whaling in the North Atlantic. *Science* **301**, 508–510 (2003).
64. Kimura, M. Diffusion models in population genetics. *J. App. Prob.* **1**, 177–223 (1964).
65. Kimura, M. & Crow, J. F. The measurement of effective population size. *Evolution* **17**, 279–288 (1963).
66. Ethier, S. & Nagylaki, T. Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Adv. Appl. Prob.* **12**, 14–49 (1980).
67. Nagylaki, T. Models and approximations for random genetic drift. *Theor. Pop. Biol.* **37**, 192–212 (1990).
68. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930; Variorum Edn, Oxford Univ. Press, 1999).
Fisher's summing up of his fundamentally important contributions to population genetics theory.
69. Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino-acid mutations in humans. *Genetics* **173**, 891–900 (2006).
70. Keightley, P. D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261 (2007).
71. Loewe, L. & Charlesworth, B. Inferring the distribution of mutational effects on fitness in *Drosophila*. *Biol. Lett.* **2**, 426–430 (2006).
72. Maside, X., Weishan Lee, A. & Charlesworth, B. Selection on codon usage in *Drosophila americana*. *Curr. Biol.* **14**, 150–154 (2004).
73. Comeron, J. M. & Guthrie, T. B. Intragenic Hill–Robertson interference influences selection on synonymous mutations in *Drosophila*. *Mol. Biol. Evol.* **22**, 2519–2530 (2005).
74. Comeron, J. M. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proc. Natl Acad. Sci. USA* **103**, 6940–6945 (2006).
75. Cutter, A. D. & Charlesworth, B. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr. Biol.* **16**, 2053–2057 (2006).
76. Li, W.-H. Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345 (1987).
77. Bulmer, M. G. The selection–mutation–drift theory of synonomous codon usage. *Genetics* **129**, 897–907 (1991).
78. Paland, S. & Lynch, M. Transitions to asexuality result in excess amino-acid substitutions. *Science* **311**, 990–992 (2006).
79. Woolfit, M. & Bromham, L. Population size and molecular evolution on islands. *Proc. R. Soc. B* **272**, 2277–2282 (2005).
80. Fry, A. J. & Wernegreen, J. J. The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**, 1–10 (2005).
81. Charlesworth, J. & Eyre-Walker, A. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc. Natl Acad. Sci. USA* **104**, 16992–16997 (2007).
82. Haddrill, P. R., Halligan, D. L., Tomaras, D. & Charlesworth, B. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8**, R18 (2007).
83. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).
84. Malécot, G. *The Mathematics of Heredity* (W. H. Freeman, San Francisco, California, 1969).
85. Maruyama, T. *Stochastic Problems in Population Genetics. Lectures in Biomathematics 17* (Springer, Berlin, 1977).
86. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893–905 (2001).
87. Nagylaki, T. Geographical invariance in population genetics. *J. Theor. Biol.* **99**, 159–172 (1982).
88. Nagylaki, T. The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604 (1998).
89. Kimura, M. 'Stepping stone' model of a population. *Ann. Rep. Nat. Inst. Genet.* **3**, 63–65 (1953).
90. Wilkinson-Herbots, H. M. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37**, 535–585 (1998).
91. Wakeley, J. Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871 (1999).
This explains the large deme number approximation, and applies it to problems in human population genetics.
92. Wakeley, J. The coalescent in an island model of population subdivision with variation among demes. *Theor. Pop. Biol.* **59**, 133–144 (2001).
93. Matsen, F. A. & Wakeley, J. Convergence to the island-model coalescent process in populations with restricted migration. *Genetics* **172**, 701–708 (2006).
94. Wakeley, J. & Lessard, S. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* **164**, 1043–1053 (2003).
95. Maruyama, T. On the fixation probabilities of mutant genes in a subdivided population. *Genet. Res.* **15**, 221–226 (1970).
96. Maruyama, T. A simple proof that certain quantitities are independent of the geographic structure of population. *Theor. Pop. Biol.* **5**, 148–154 (1974).
97. Cherry, J. L. & Wakeley, J. A diffusion approximation for selection and drift in a subdivided population. *Genetics* **163**, 421–428 (2003).
98. Cherry, J. L. Selection in a subdivided population with dominance or local frequency dependence. *Genetics* **163**, 1511–1518 (2003).
99. Cherry, J. L. Selection in a subdivivided population with local extinction and recolonization. *Genetics* **164**, 789–779 (2003).
100. Whitlock, M. C. Fixation probability and time in subdivided populations. *Genetics* **164**, 767–779 (2003).
101. Garcia-Dorado, A. & Caballero, A. On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* **155**, 1991–2001 (2000).

102. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**, 1855–1861 (2000).
103. Roze, D. & Rousset, F. Selection and drift in subdivided populations: a straightforward method for deriving diffusion approximations and applications involving dominance, selfing and local extinctions. *Genetics* **165**, 2153–2166 (2003).
104. Shiina, T. *et al.* Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**, 1555–1570 (2006).
105. Richman, A. D., Uyenoyama, M. K. & Kohn, J. R. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science* **273**, 1212–1216 (1996).
106. Kamau, E., Charlesworth, B. & Charlesworth, D. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. *Genetics* **176**, 2357–2369 (2007).
107. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. *Nature* **356**, 519–520 (1992).
108. Shapiro, J. A. *et al.* Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl Acad. Sci. USA* **104**, 2271–2276 (2007).
109. Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
110. Roselius, K., Stephan, W. & Städler, T. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753–763 (2005).
111. Spencer, C. C. A. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, 1375–1385 (2006).
112. Hudson, R. R. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–45 (1990).
113. Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* **168**, 2363–2372 (2004).
114. Charlesworth, D., Kamau, E., Hagenblad, J. & Tang, C. *Trans*-specificity at loci near the self-incompatibility loci in Arabidopsis. *Genetics* **172**, 2699–2704 (2006).
115. Asthana, S., Schmidt, S. & Sunyaev, S. A limited role for balancing selection. *Trends Genet.* **21**, 30–32 (2005).
116. Bubb, K. L. *et al.* Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**, 2165–2177 (2006).
117. Baysal, B. E., Lawrence, E. C. & Ferrell, R. E. Sequence variation in human succinate dehydrogenase genes: evidence for long-term balancing selection on *SDHA*. *BMC Biol.* **5**, 12 (2007).
118. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
119. Robertson, A. Inbreeding in artificial selection programmes. *Genet. Res.* **2**, 189–194 (1961).
120. Santiago, E. & Caballero, A. Effective size of populations under selection. *Genetics* **139**, 1013–1030 (1995).
121. Santiago, E. & Caballero, A. Effective size and polymorphism of linked neutral loci in populations under selection. *Genetics* **149**, 2105–2117 (1998).
122. Marais, G. & Piganeau, G. Hill–Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. *Mol. Biol. Evol.* **19**, 1399–1406 (2002).
123. Bartolomé, C. & Charlesworth, B. Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. *Genetics* **174**, 2033–2044 (2006).
124. Bachtrog, D., Hom, E., Wong, K. M., Maside, X. & De Jong, P. Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biol.* **9**, R30 (2008).
125. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
126. Berry, A. J., Ajioka, J. W. & Kreitman, M. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics* **129**, 1111–1117 (1991).
127. Barton, N. H. Genetic hitchhiking. *Phil. Trans. R. Soc. B* **355**, 1553–1562 (2000).
128. Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. The hitchiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* **140**, 783–796 (1995).
129. Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429 (1995).
130. Stephan, W., Wiehe, T. H. E. & Lenz, M. W. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Pop. Biol.* **41**, 237–254 (1992).
131. Stephan, W. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* **12**, 959–962 (1995).
132. Kim, Y. Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. *Mol. Biol. Evol.* **21**, 286–294 (2004).
133. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchiking model. *Genetics* **155**, 909–919 (2000).
134. Andolfatto, P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* **17**, 1755–1762 (2007).
135. Loewe, L. & Charlesworth, B. Background selection in single genes may explain patterns of codon bias. *Genetics* **175**, 1381–1393 (2007).
136. Kaiser, V. B. & Charlesworth, B. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* (in the press).
137. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).
138. Normark, B. B., Judson, O. P. & Moran, N. A. Genomic signatures of ancient asexual lineages. *Biol. J. Linn. Soc.* **79**, 69–84 (2003).
139. Barrett, S. C. H. The evolution of plant sexual diversity. *Nature Rev. Genet.* **3**, 274–284 (2002).
140. Gordo, I., Navarro, A. & Charlesworth, B. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**, 835–848 (2002).
141. Cotterman, C. W. *A calculus for statistico-genetics.* Thesis, Ohio State Univ. (1940).
142. Malécot, G. *Les Mathématiques de l'Hérédité* (Masson, Paris, 1948).
143. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43 (1982).
144. Kimura, M. Theoretical foundations of population genetics at the molecular level. *Theor. Pop. Biol.* **2**, 174–208 (1971).
145. Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O. & Li, W. H. Nucleotide diversity in gorillas. *Genetics* **166**, 1375–1383 (2004).
146. Charlesworth, J. & Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23**, 1348–1356 (2006).
147. Alter, S. E., Rynes, E. & Palumbi, S. R. DNA evidence for historic population size and past ecosystem impacts of gray whales. *Proc. Natl Acad. Sci. USA* **104**, 15162–15167 (2007).
148. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418**, 323–326 (2002).

Acknowledgements
B.C. thanks the Royal Society for support from 1997–2007.

FURTHER INFORMATION
Brian Charlesworth's homepage:
http://www.biology.ed.ac.uk/research/institutes/evolution/homepage.php?id=bcharlesworth

ALL LINKS ARE ACTIVE IN THE ONLINE PDF