Garland Science

# HUMAN EVOLUTIONARY GENETICS

second edition

Jobling · Hollox · Hurles · Kivisild · Tyler-Smith

# PROCESSES SHAPING DIVERSITY

CHAPTER FIVE

In *On the Origin of Species* Charles Darwin was primarily interested in the evolution of species over geological time. Others have been more concerned with processes operating on genetic diversity, within a single species, over a timescale of generations. These two scales of evolutionary change are often referred to as **macro**- and **microevolution**. While it is often assumed that species-level evolution is just an extrapolation of population-level evolution, the reconciliation of these two fundamental evolutionary levels is by no means complete. In this chapter, we will show that because microevolutionary processes shape genetic diversity, we can measure them by studying allele frequencies within populations. We will then have enough grounding in population genetic theory to understand its application to human evolutionary studies, particularly in Chapter 6, but also throughout the book. For further details on population genetic theory we recommend a specialist textbook, such as Hartl and Clark's *Principles of Population Genetics*.[16]



## 5.1 BASIC CONCEPTS IN POPULATION GENETICS

### Why do we need evolutionary models?

We study evolutionary processes by considering how allele frequencies within a population change in time and space. By understanding the mechanisms through which evolutionary processes act, we can produce mathematical models that approximate reality. Such models are necessary to understand the subtle interplay between the processes, and allow us to infer past processes from modern diversity. Using mathematical models that represent simplified versions of reality we can estimate parameters from the data, such as population growth rate, the age of an allele, or the migration rate between two populations. Models also allow us to test different hypotheses about the past. Put simply, if the model does not fit the observed data well, at least one of the assumptions underlying the model must be wrong. Alternatively, we can make several models and test which one best fits the observed data: for example, does a prehistoric migration between two ancestral populations, or divergence during a period of isolation, better explain the current patterns of genetic diversity? There are a variety of methods for testing **goodness-of-fit**, some of which are explored in the next chapter (see Box 6.4 for more about likelihood-based methods), where we also give several examples of how real inferences about human evolution can be derived from analysis of data using mathematical models.

One of the strengths of many population genetic models is their generality: they can be applied to data from any species that share broad characteristics. For example, some models applied to humans might be equally applicable to all

other species that reproduce sexually and do not self-fertilize. However, models require us to make assumptions that may not be true of all species. This problem drives mathematical models of evolution to become ever more sophisticated, abandoning simplifying assumptions one by one, and introducing new parameters that provide a better fit to biological reality. Nonetheless, even in the data-rich field of modern human genetics, no amount of data can compensate for an inappropriate model.

The concept of a **population** is central. We must define a population before we can measure the frequency of an allele within it. In addition, we are often interested in reconstructing past demographic events, and demography is a property of populations, not of individuals. It is for these reasons that this discipline is known as **population genetics**. Furthermore, many studies of human genetic diversity group individuals from a number of closely situated but distinct locations into a single population, often defined by political boundaries that may be only a few human generations old. An ecological approach to sampling, such as using regular grid squares, is rarely, if ever, adopted for humans (**Section 10.2**). This sampling of groups, rather than of individuals, leads to their being considered as a natural unit of investigation.

One type of model we will encounter is a mathematical approximation of populations, their interactions, and mating structures. When the term "population" is being used it is important to be clear how it was defined and whether it refers to individuals grouped together for the sake of analysis, or an idealized group, assumed to be adhering to the assumptions of a mathematical model (for example, randomly mating). In other words, does the term refer to a practical or theoretical entity?

The other types of mathematical model are those describing the molecular processes of mutation and recombination, which, as we saw in Chapter 3, differ between DNA sequences and genomic regions. These enable us to go beyond allelic definitions and allow us to make the connection between molecular diversity and population processes.

### The Hardy–Weinberg equilibrium is a simple model in population genetics

The **Hardy–Weinberg equilibrium** (HWE) model describes the relationship between allele frequencies and genotype frequencies in a randomly mating population. In diploid organisms such as humans, two alleles, $A_1$ and $A_2$, at the same locus, with allele frequencies $p$ and $q$ respectively, can be sorted to make three possible genotypes: $A_1A_1$, $A_1A_2$, and $A_2A_2$. If we know the frequency of these two alleles in a population ($p$ and $q$) we can predict the proportions of the genotypes in the succeeding generation by combining gametes (which contain single alleles) at random—a postulate known as the Hardy–Weinberg principle.[15] Thus the proportion of each genotype in the next generation is:

$$A_1A_1 = p^2,\ A_1A_2 = 2pq, \text{ and } A_2A_2 = q^2$$

If the genotype proportions in the next generation are calculated in this manner, and are found to be indistinguishable from those in the parental generation, then no evolution (defined as a change in allele frequencies) is occurring, and the population is at HWE. At the time of its discovery, the existence of this equilibrium was important as it showed that mating alone need not alter allele frequency.

For us to be able to estimate genotype proportions from one generation to the next in this way, the population must be made up of an infinite number of randomly mating, sexually reproducing diploid organisms. However, for HWE to be observed the idealized population must have certain additional properties, including:

- No selection
- No mutation

- No overlap between generations
- No migration
- No substructure

If the genotype proportions are not in HWE we might reasonably conclude that at least one of these assumptions has been broken.

How do we use the HWE model to test a hypothesis about human evolution? We can test the observed genotype frequency for an allele at a single nucleotide polymorphism (SNP, **Section 3.2**) against that expected from HWE given the allele frequencies deduced from the data. If the observed data do not fit the model well, one of the assumptions (for example, no selection) of the model does not apply to the SNP. Ability to digest **lactose** in milk as an adult is determined by a single SNP in Europeans (**Section 15.6**). Our hypothesis is that the trait, and therefore alleles at the responsible SNP, will be subject to natural selection.

Using genotype frequency data from over 3000 British people,[9] we can calculate allele frequencies for $p$ and $q$ as 0.747 and 0.253 respectively. **Table 5.1** shows that we can calculate the expected genotype frequencies using Hardy–Weinberg proportions, and compare them with the observed genotype frequencies using a goodness-of-fit test (in this case the $\chi$-squared test). Following the calculation in Table 5.1, there is no significant difference between observed and expected genotype frequencies given the HWE model, showing that the assumptions of the HWE model have not been broken. So we would infer that this SNP is not subject to natural selection.

However, we would be wrong; indeed, as shown by other tests (**Section 15.6**), this SNP displays some of the strongest evidence of positive selection for any variant in the genome. So why doesn't testing for departure from HWE detect selection at this SNP? The answer is that this test is very weak, and is poor at rejecting the null hypothesis (no selection). There are two reasons for this:

- Very strong natural selection is required to distort genotype frequencies sufficiently to be detected by goodness-of-fit tests.
- One round of random mating in the absence of natural selection restores genotype frequencies to HWE. Therefore the selective events of the past are very effectively erased and the HWE is capable of detecting selection only in the current generation.

Departures from HWE are generally rare in humans, and would only be observed as a result of selection if extreme differential mortality occurred within a single generation, as in survivors of kuru (**Box 5.1**). More often, they can result from population structure (and hence departure from random mating) generated, for example, by regarding samples from different continents as a single population.

**TABLE 5.1:**
**TESTING OBSERVED GENOTYPE COUNTS AGAINST HARDY–WEINBERG EQUILIBRIUM EXPECTATION**

| Genotypes | Observed genotype counts | Expected genotype frequencies | Expected genotype counts | | $(O–E)^2/E$ |
|---|---|---|---|---|---|
| TT | 1881 | 0.567 | 1897.6 | | 0.145 |
| CT | 1236 | 0.378 | 1264.4 | | 0.639 |
| CC | 227 | 0.064 | 214.1 | | 0.777 |
| | | | | Sum | 1.561 |
| | | | | $p$ (1 df) | 0.21 |

### Box 5.1: Kuru disease in the Fore of Papua New Guinea

Hardy–Weinberg equilibrium (HWE) of genotypes is expected in an outbreeding species such as humans, so well-established instances of deviations from HWE, Hardy–Weinberg disequilibrium, are very unusual and particularly interesting. A nonsynonymous SNP (rs1799990) at codon 129 of the human prion protein gene (*PRNP*) encodes either methionine or valine, and heterozygosity confers resistance to the acquired neurodegenerative disease **kuru**[29, 32] (OMIM 245300).

Kuru is caused and transmitted by a **prion** encoded by *PRNP*, and first came to the attention of Western medicine in the 1950s, when the Eastern Highlands of Papua New Guinea came under external administrative control. Inhabitants of this region included the Fore (Figure 1), who had a high incidence of kuru, with a peak mortality per year of around 2% in some villages. It was found that kuru is transmitted by consuming the brains of kuru-infected individuals, and that the Fore routinely ate deceased relatives at mortuary feasts. The men had the first choice of tissues, and left the less attractive brain, enriched for prions, to the women. Kuru was therefore more common among women than men. The practice subsequently stopped, so that young Fore do not engage in it.

Measuring the genotype frequencies of Fore women born before 1950, who had therefore been exposed to kuru-infected brains on multiple occasions yet were still surviving, showed a dramatic increase in frequency of heterozygotes. This increase is not seen in young modern Fore, nor in men born before 1950 who would have been less involved in brain consumption (Table 1). The departure from HWE is due to the selective mortality of *PRNP* homozygotes from kuru.

**Figure 1: A group of Fore men.**
In the 1950s and 1960s, the kuru epidemic killed a quarter of the female population in the South Fore, with few female survivors of marriageable age in some villages. [From Mathews JD (2008) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3679.]

**TABLE 1:**
***PRNP* CODON 129 GENOTYPES IN SUSCEPTIBILITY-STRATIFIED GROUPS FROM THE EASTERN HIGHLANDS OF PAPUA NEW GUINEA.**

| Fore group | Methionine homozygotes | Heterozygotes | Valine homozygotes | Departure from HWE, *p* value |
|---|---|---|---|---|
| Women born before 1950 | 16 | 86 | 23 | $2.1 \times 10^{-5}$ |
| Men born before 1960 | 34 | 111 | 60 | 0.15 |
| Young modern Fore individuals | 52 | 136 | 94 | 0.80 |

Indeed, departures from HWE as a result of biological effects are so rare and subtle that an apparent gross departure from HWE within a single population is routinely used to detect technical errors in genotyping and improve data quality prior to further analyses.[28]

The weakness of HWE as a test for events that alter allele frequency emphasizes the importance of more sophisticated population genetic models. These incorporate information about mutation rate, recombination rate, and population size: processes that are discussed in the rest of this chapter. How we use these improved models to test hypotheses in human evolutionary genetics is the subject of Chapter 6.

## 5.2 GENERATING DIVERSITY BY MUTATION AND RECOMBINATION

Mutation is the only process generating new alleles: indeed, by definition any change producing a new allele is called a mutation. It provides the raw material on which evolution can act. There are a broad variety of mutational changes, and these occur at widely varying rates (see Chapter 3). Each mutation is a

single change occurring in a single cell. Evolutionary consequences follow only from those changes that occur in the germ line, and not those in somatic tissues, because somatic mutations are not heritable. The dynamics of many types of mutations vary between the soma and the germ line. Because of the high fidelity of DNA polymerases and the operation of DNA repair mechanisms, germ-line mutations occur at low rates for individual nucleotides, although (given the size of the human genome) they are inevitable in every generation. Estimates of the human nucleotide mutation rate from different studies are given in Table 6.4, and estimates of the mutation rate of different classes of substitution are given in Table 3.1.

### Mutation changes allele frequencies

In the absence of other processes, a particular allele will decrease in frequency, because it will accumulate mutations changing it into different alleles. This phenomenon is known as mutation pressure. By knowing the mutation rate for the whole gene ($\mu$) and the initial allele frequency ($p_0$), assuming no back mutation, and ignoring stochastic processes, we can calculate this allele's frequency ($p_t$) $t$ generations later, by:

$$p_t = p_0 e^{-\mu t}$$

At low mutation rates, mutation pressure is a weak force that can only have appreciable impact over long time-scales. After 1000 generations, the wild-type sequence of a gene 1000 bp in size with a per-generation nucleotide mutation rate of $2 \times 10^{-9}$ will only decrease in frequency from 1.0 to 0.998.

### Mutation can be modeled in different ways

The example above introduced the model of a gene in which each new mutation creates a new allele; in other words we discounted the possibility of **back mutations** and **recurrent mutations**. This is known as the **infinite alleles model**. If we consider a gene 1000 bp in length then the number of possible SNP alleles is enormous: $4^{1000}$. If the 1000-bp sequence has $n$ mutational changes in $n$ different nucleotides then the probability of a back mutation is small: $n/3000$.

However, if we consider the evolution of a polymorphic microsatellite, oscillating in size by whole numbers of repeats, we can see that the opportunity for back mutation and recurrent mutation is much greater than for SNPs. Thus the infinite alleles model does not always appear to be a close approximation of biological reality. We need different models for different types of mutation. The stepwise mutation model (SMM) provides a better fit to microsatellite evolution. According to this model, mutations increase and decrease allele length by one unit with equal probability (**Figure 5.1**).

Initially, the SMM considered single-step changes only, but there is good empirical evidence for a lower, but nevertheless appreciable, rate for multiple-step mutations and the model can be adapted to account for these.[10] There are, however, other known aspects of microsatellite evolution not incorporated within the SMM model (see also **Section 3.4**):

- A positive correlation between allele length and mutability
- A lower length threshold under which mutation rate becomes undetectable
- A possible small bias toward expansions of short alleles, resulting in an increase in size of the microsatellite

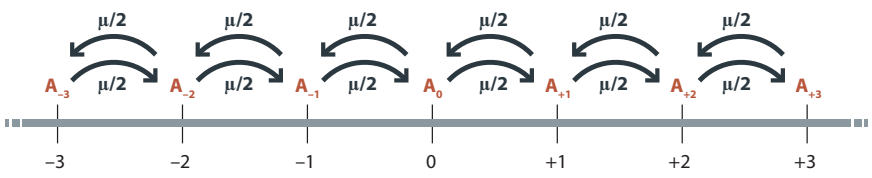



**Figure 5.1: The stepwise mutation model.**
The model considers only single-step mutations, and regards an increase or decrease as equally probable and independent of allele length. The average mutation rate is $\mu$, and any allele mutates to a smaller or larger allele with rate $\mu/2$.

- A possible preference for deletions rather than expansions in longer alleles; together with the previous point, this produces an equilibrium allele length distribution
- Very large expansions in triplet-repeat diseases, and consequent negative selection in these and other examples

Other types of mutations, such as genomic structural variation and GC-rich minisatellite mutations, fit neither of the above models.

If we are interested in aspects of sequence evolution involving the possibility of several changes occurring at the same site, then we need more complex models of mutation—for example, we may need to consider the probability that an A will mutate to a C and then subsequently back to an A again. These models come into play when considering sequence evolution over long time-scales, where back mutations result in the observed **sequence divergence** being an underestimate of the true number of mutational changes. We will come to applications of these models in Chapter 6.

In the simplest model all nucleotide substitutions occur at the same rate, while the most complex model allows a different rate for each nucleotide change. These models can be represented as a substitution scheme, and as a probability matrix, shown in Figure 5.2. The simplest example is known as the **Jukes–Cantor model** (JC), and one of the more complex models is the **general reversible model** (**REV**). There are a number of intermediate models that contain some, but not all, of the complexity of the REV model.

The frequency of each nucleotide clearly influences the probability of nucleotide changes averaged over an entire sequence. For example, an A to G transition may have the same rate as a C to T transition, but if there are twice as many As as Cs in a sequence then the probability of an A to G occurring within the sequence as a whole is not the same as that of a C to T. The JC model does not take potential bias in **base composition** into account, but the REV model does.

There are further aspects of sequence evolution known from empirical studies that are not accounted for in these models (Section 3.2). First, small (1–20 bp) insertion or deletion alleles (indels) occur on average once every 7.2 kb in the human genome.[34] Ignoring this kind of mutational change can have a large impact; for example, whether or not indels are removed prior

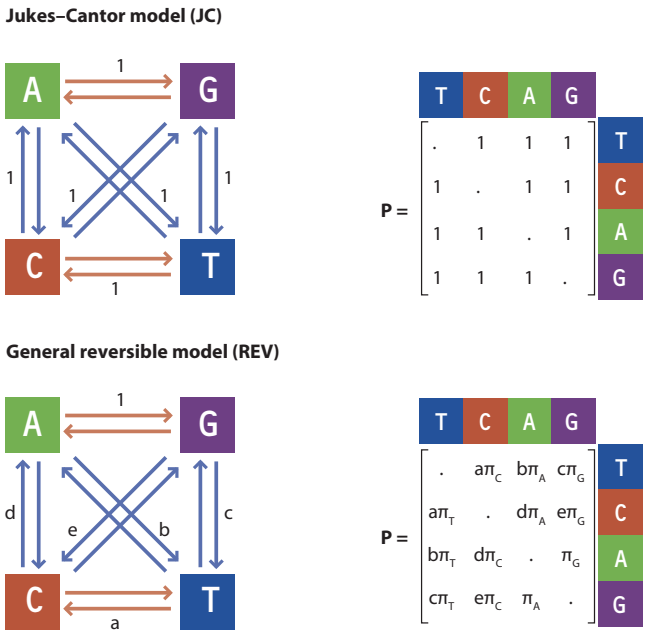



**Figure 5.2: Models of sequence evolution.**
The probability matrices of two different models of sequence evolution are shown. This matrix contains the relative rates of the different possible base substitutions, which are also shown on a substitution scheme that shows transitions in *red* and transversions in *blue*. The REV model includes the $\pi_l$ parameter, which is the frequency of that base in the sequence.

to sequence analysis makes a fourfold difference to the apparent sequence divergence between humans and chimpanzees, according to one way of measuring it (**Section 7.3**). The probability of a small indel event occurring is largely determined by the repetitive nature of the surrounding sequence, in a manner that is poorly understood and therefore difficult to model. Such changes are rarely found as polymorphisms in coding regions because they often disrupt the **reading frame**. Second, the phenomenon of the increased mutability of CpG dinucleotides departs significantly from the REV model (**Section 3.2**). The mutability of a nucleotide depends on its neighbor, so that not all Cs and Gs have the same probability of mutating. Both transitions and transversions have increased probability at CpGs.

Models have been developed that can accommodate rate variation among sites within a sequence. These fit such variations in rate to a statistical distribution. Some, like the **gamma distribution**, have a single modal value, whereas other models allow multimodal rate distributions that may provide a better fit to the rate variation among sites, as suggested by the increased mutability of CpGs described above.

### Meiotic recombination generates new combinations of alleles

Meiotic recombination occurs as a part of sexual reproduction, and enhances the ability of populations to adapt to their environments by combining advantageous alleles at different loci (**Figure 5.3**). By contrast, asexually reproducing species and nonrecombining portions of the human genome are prone to the operation of **Muller's ratchet**, the slow but inexorable accumulation of deleterious mutations. This process of degeneration may explain the low density of functional genes on the nonrecombining portion of the Y chromosome (**Appendix**).

Recombination generates new combinations of alleles on the same DNA molecule, known as haplotypes (**Section 3.8**), and in this way increases haplotype diversity. Consequently, recombination is capable of breaking up advantageous allelic combinations. This results in the theoretical possibility that outbreeding can result in a drop in fitness known as **outbreeding depression**.

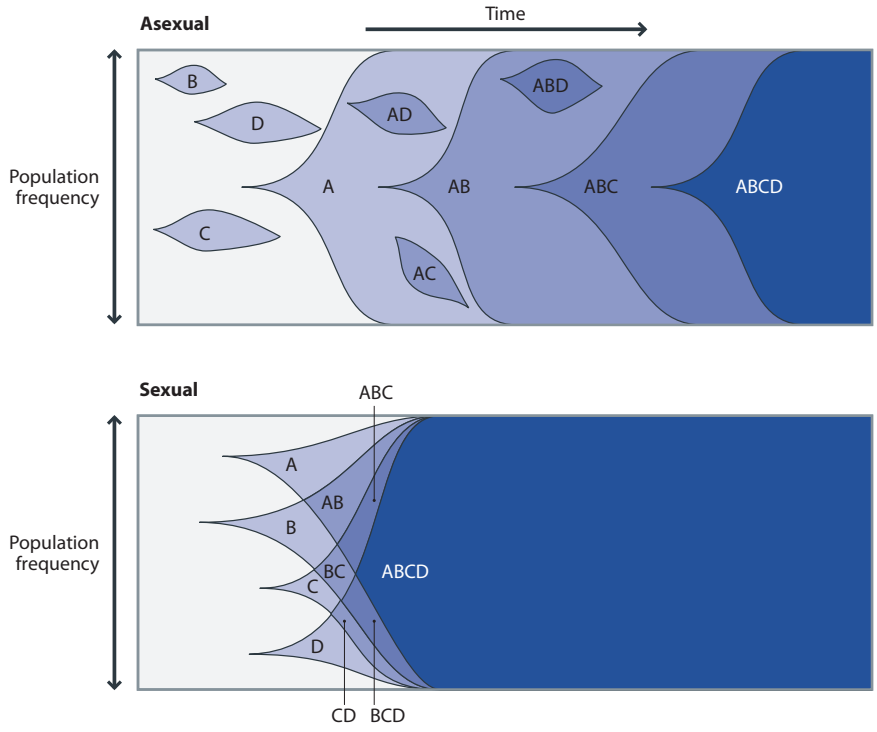



**Figure 5.3: The advantage of sexual reproduction.**
Four alleles (A–D) all increase the fitness of the organism, with the fittest having all four alleles. Only one allele at a time can prevail in an asexual organism, so they must be combined serially. By comparison, in a sexually reproducing organism these beneficial alleles can be combined in parallel. Thus it takes much less time to assemble the fittest genotype.

While alleles at loci on different chromosomes are randomly segregated during meiosis, alleles at loci closely linked on the same chromosome are not, as recombination between them occurs infrequently. Linked loci share a common evolutionary heritage: selection operating on one locus will affect diversity at the other. For example, an allele that rises to high frequency because of positive selection on a linked locus is said to be hitchhiking (**Section 6.7**). Conversely, negative selection at a locus also reduces diversity at linked loci, albeit at a slow rate, by a process known as **background selection**.[8]

### Linkage disequilibrium is a measure of recombination at the population level

Recombination can be studied at the population level by investigating whether specific alleles at different loci are correlated with one another more or less often than would be expected by chance. This nonrandom correlation is known as linkage disequilibrium (LD; **Section 3.8**, Box 3.5).

In an analogous fashion to the reduction in frequency of an allele by mutation pressure, recombination can reduce the frequency of a haplotype. Rather than monitor this process through the decline in frequency of the haplotype itself, we can follow the decay of LD using the statistic $D$ as follows. When a new mutation arises on a chromosome, it is linked to all other variant sites on the same chromosome forming a single haplotype. In other words, it will only be found associated with one allele at each of those other loci, and so is in complete LD with them ($D$ is at its maximal possible value). However, over several generations the frequency of the new mutant allele may grow; if so, recombination events will introduce the new allele onto copies of the chromosome with different alleles at the other variant sites (see the figure in Box 3.5). As a consequence, LD starts to decay. If we know the recombination rate per generation ($r$) between the newly mutated locus and a given locus, after a certain number of generations ($t$) we can track the decay of LD over time, by relating the present value of $D$ ($D_t$) to the initial value of $D$ ($D_0$) using the equation:

$$D_t = (1 - r)^t \times D_0$$

From this equation we can see that as time increases, $D_t$ tends to zero (linkage equilibrium). In addition, as we move along the chromosome away from the newly mutated locus, the interlocus recombination rate increases, meaning that $D_t$ will tend to zero even sooner. In an infinitely large population, LD would continue to decay over time as a result of an ever-increasing frequency of recombination between the newly mutated locus and any other locus. However, real populations are not infinitely large, and in **Section 5.6** we will explore why an inexorable decay of LD is an unrealistic expectation.

### Recombination results in either crossing over or gene conversion, and is not uniform across the genome

In comparison with models of mutation, models of recombination have traditionally been fairly simple. The simplest model is that the rate of recombination is uniform. In other words, the probability of a **crossover** occurring between a pair of sequence variants is determined only by the physical distance that separates them. The products of this type of recombination event are two new haplotypes containing contiguous stretches of alleles from each **ancestral haplotype** (**Figure 5.4a**).

Studies of recombination in humans and model organisms have revealed two biological properties of recombination that conflict with this simple model of recombination. First, not every recombination event results in a crossover (**Section 3.8**). A recombination intermediate can be resolved in one of two ways: a crossover, or a gene conversion event that converts a small segment of DNA (typically less than a kilobase) in one haplotype in a nonreciprocal way so that it is identical to that same segment in the other haplotype. Many

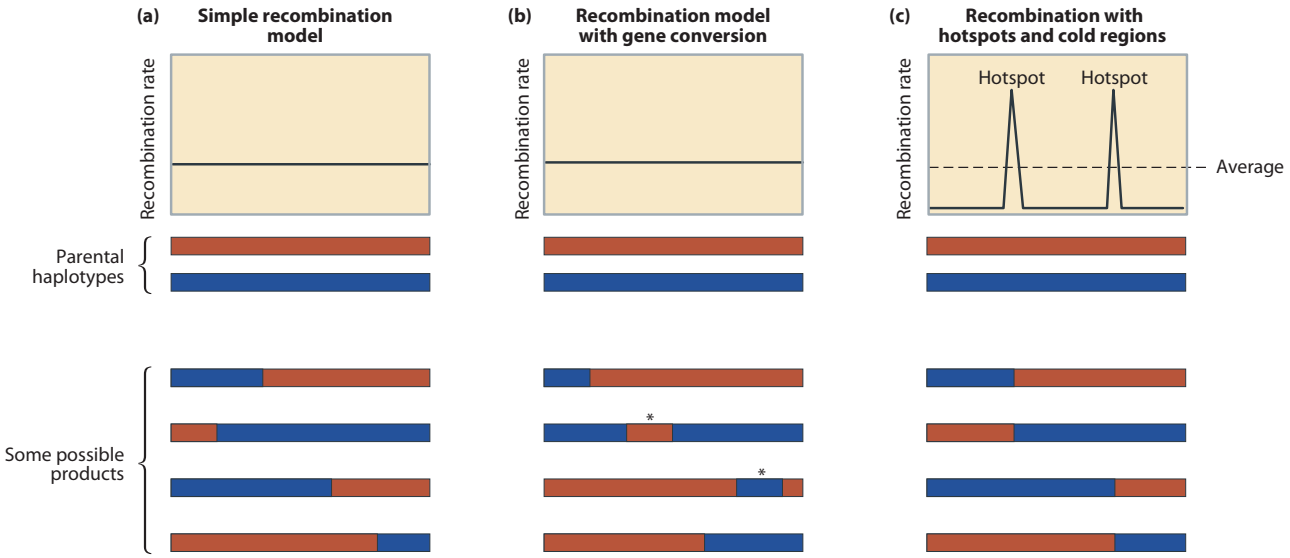



**Figure 5.4: Three models of recombination.** The recombination rate along the haplotype is shown above the parental haplotype and four typical recombined haplotypes for three models of recombination: (a) simple recombination, (b) recombination with gene conversion, and (c) recombination with hotspots and cold regions. Sections resulting from gene conversion are highlighted by asterisks.

recombination models used on large datasets, for example in the initial HapMap study, use methods that do not distinguish the effects of crossovers and gene conversions. Second, recombination rates are not uniform along a segment of DNA (**Section 3.8**). Crossovers appear to be concentrated in hotspots between which lie recombinationally inert, "cold" regions, and, at larger scales, recombination rates vary along the chromosome, often being relatively low near centromeres and high near telomeres. Hotspot position is different between individuals, and between populations, because of genetic variation in the *PRDM9* gene (**Section 3.8**).

Some models of recombination have been proposed that incorporate either one of these two additional complexities, but few, if any, models have combined the two. Incorporating gene conversion into recombination models requires knowledge of the ratio of gene conversions to crossover events, and the length of the gene-converted segment[46] (Figure 5.4b). Recombination rate heterogeneity can be modeled by considering the size and spacing of recombination hotspots, and the ratio of the recombination rates in hotspots and in cold regions (Figure 5.4c).

## 5.3 ELIMINATING DIVERSITY BY GENETIC DRIFT

No population is infinitely large, as is assumed by the Hardy–Weinberg theorem. Each generation represents a finite sample from the previous one, and variation in allele frequency between generations occurs through the stochastic process of sampling. This source of variation is known as **random genetic drift**.[44]

Intuitively, we might expect that the magnitude of genetic drift relates to the size of the population being sampled, and this can be shown to be true. **Figure 5.5** illustrates the change in allele frequency over 100 generations in simulated populations, starting with an initial allele frequency of 0.5. The allele rapidly becomes either fixed (100% frequency) or lost from the populations of constant size 20, whereas both alleles persist in the populations of constant size 1000, with more subtle variations in frequency. As genetic drift is a random process, it is impossible to predict which allele survives. A model that describes genetic drift in a finite population in combination with the other assumptions of the HWE (**Section 5.1**) is known as the **Wright–Fisher model**. This model is fundamental to many aspects of population genetics.

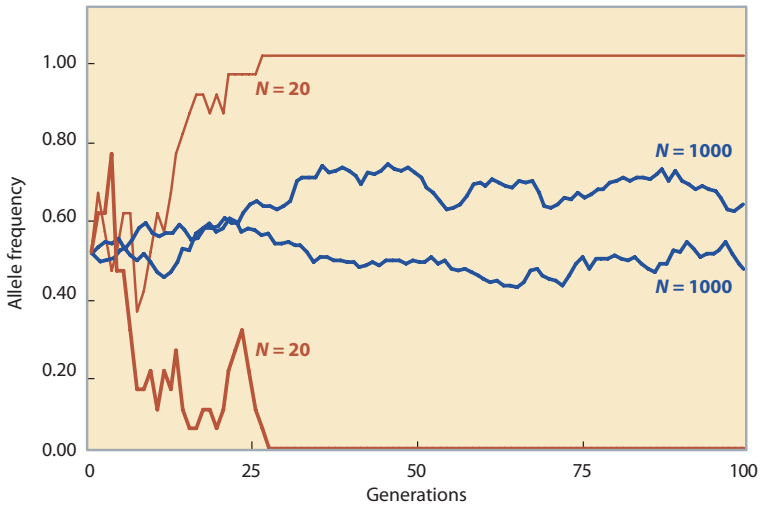



**Figure 5.5: Genetic drift in populations of different sizes.**
The results of simulations (over 100 generations) of allele frequencies of a binary polymorphism in diploid organism populations of size 20 or 1000 (each starting from a frequency of 0.5) are shown. The populations are of constant size and have non-overlapping generations; each generation is sampled randomly from the previous one (so individuals have an equal probability of contributing to the next generation). The allele rapidly becomes either fixed or lost from the populations of constant size 20, whereas more subtle variations are seen in the populations of constant size 1000.

## The effective population size is a key concept in population genetics

The Wright–Fisher model, like the HWE model, contains many unrealistic assumptions when compared with real populations. First, generations overlap in real human populations; second, populations are rarely constant in size; and third, large populations do not exhibit random mating. These three factors differ in importance for any given population. Wright's concept of **effective population size** ($N_e$) allows us to compare the amount of genetic drift experienced by different populations.[7] $N_e$ for any population represents the size of an idealized Wright–Fisher population that experiences the same amount of genetic drift as the one under study. It measures the magnitude of genetic drift: the smaller $N_e$, the greater the drift. We can understand the impact of different properties of real populations on genetic drift through the changes they cause in this value.

There are, in fact, two genetic ways of defining effective population sizes: one is based on the sampling variance of allele frequencies (that is, how an allele's frequency might vary from one generation to the next), and the other utilizes the concept of inbreeding (that is, the probability that the two alleles within an individual are identical by descent from a common ancestor). Both of these properties of a finite population depend on the size of that population. There also can be nongenetic definitions, such as the number of breeding individuals inferred from demographic studies. For the sake of simplicity in this chapter we treat these definitions interchangeably, but the reader should be aware that while under most simple population scenarios these definitions of effective population size give identical values for $N_e$, in more complex situations this is not always the case.

It is not easy to relate the effective population size ($N_e$) to the **census size** of a population ($N$), as there are many parameters that can affect this relationship, only some of which are relevant to humans. These are discussed later in this section. $N_e$ is almost always substantially less than the actual population size. For example, the introduction of overlapping generations alone into the population model[11] reduces $N_e$ to 25–75% of $N$. It is also important to remember that there is a distinction between **long-term effective population size** and recent effective population size. In descriptions of genetic diversity, in most genetic literature, and in this book, the value normally refers to long-term $N_e$. Recent effective population size, in humans, can be quite distinct from the pattern predicted from genetic diversity data, and reflects the very recent exponential expansion in human census size. We will discuss how human $N_e$ can be estimated in **Section 6.6**, and some estimates of human $N_e$ are given in Table 6.4.

**TABLE 5.2:**
**RELATIVE EFFECTIVE POPULATION SIZES FOR DIFFERENT CHROMOSOMES**

| | Y chromosome | X chromosome | Autosome |
|---|---|---|---|
| Wright–Fisher population | 1/4 | 3/4 | 1 |
| Extreme male reproductive variance | 1/8 | 9/8 | 1 |

### Different parts of the genome have different effective population sizes

Up to this point, effective population size has been considered at the level of individuals; however, not all genomic loci are equally represented in all individuals. If we consider a single mating couple as a microcosm of a species with equal sex ratios, they have between them four copies of each autosome, three copies of the X chromosome, two copies of mitochondrial DNA (mtDNA), only one of which will be inherited by the succeeding generation, and a single Y chromosome. Thus, given a 1:1 sex ratio, the effective population size of the Y chromosome and mtDNA will be only a quarter that of the autosomes, and a third that of the X chromosome. This assumes that the reproductive variances of males and females are equal, as in the Wright–Fisher model.

If, however, we also take into account the differences in effective population sizes between the sexes the relationships for the different loci become more complex. As we saw in Chapter 2, the Y chromosome is inherited paternally and mtDNA is inherited maternally, while the X chromosome is inherited twice as often from females as it is from males. The higher reproductive variance (that is, variation in number of offspring) of males than females reduces the $N_e$ of the Y chromosome relative to that of mtDNA, the X chromosome, and the autosomes, and increases the $N_e$ of mtDNA and the X chromosome relative to that of the autosomes. In cases of extreme male reproductive variance it is possible that the $N_e$ of the X chromosome may exceed that of the autosomes, up to a limit of 9/8 of autosomal $N_e$ (**Table 5.2**). In such extreme cases the $N_e$ of the Y chromosome approaches its lower limit of 1/8 that of the autosomes.[6] Such considerations may partially explain why the Y chromosome exhibits such a high degree of population differentiation (**Appendix**). However, discrepancies in generation times between the sexes also cause their effective population sizes to differ.[31] The sex with the shorter generation time will experience more genetic drift (all other factors being equal) as a result of more frequent episodes of sampling a new generation from the previous one. In humans, females appear to have the shorter generation time (**Section 6.6**), which should lower the $N_e$ of mtDNA relative to biparentally and paternally inherited loci. The relative importance of these opposing factors may differ from population to population. For example, analysis of detailed Icelandic genealogies indicates that in the last few centuries generation-time discrepancies between the sexes have outweighed any differences in reproductive variance, with the consequence that the effective population size of Icelandic mtDNA is less than that of the Y chromosome.[18]

### Genetic drift causes the fixation and elimination of new alleles

The concept of effective population size allows us to calculate the probability and rate of **fixation** (rise to 100% frequency) for a new allele in the absence of selection and mutation. Fixation itself is a rare event—a far more likely outcome for a new allele is that it will be lost. As intuition might suggest, with no favoring of either outcome, the fixation probability of an allele in the absence of selection is equal to its frequency in the population; a new allele would have a frequency of $1/2N$. Thus the smaller the population, the greater chance a new allele has

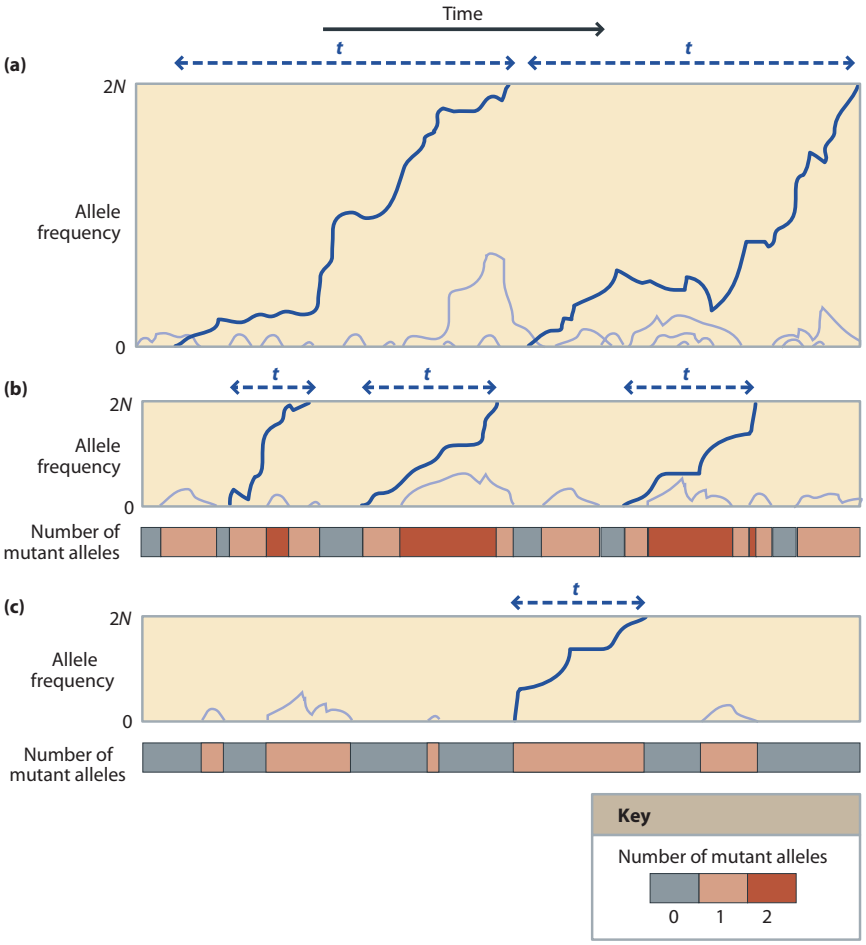



**Figure 5.6: Schematic view of the fixation of new alleles in three different populations.**
The change in allele frequency over time of new mutations in three different populations is shown. New alleles that arise and are then fixed are shown in *blue*, new alleles that are eliminated are shown in *gray*. The time taken for new alleles to fix ($t$) is longer in the larger population (a) than the smaller population (b). More new alleles are fixed in the smaller population than in the larger population. (c) A population of the same population size as (b), but with a lower mutation rate ($\mu$). The time to fixation in (c) is no different from that in (b), but the time between fixation of new alleles is greater, as is the proportion of time spent with no polymorphism.

of becoming fixed (see **Figure 5.6**). The average time to fixation ($t$) in generations has been shown to be:

$$t = 4N_e$$

Therefore a new allele in a smaller population will not only have a higher probability of becoming fixed, but it will also be fixed more rapidly than it would in a larger population. Nonetheless, fixation under the influence of drift alone is substantially slower than if selection were acting (**Section 5.6**).

### Variation in census population size and reproductive success influence effective population size

Few populations are constant in size for many generations, so what happens to the effective population size during these fluctuations? The long-term $N_e$ is approximately equal to the harmonic mean rather than the arithmetic mean of the population sizes over time (**Figure 5.7**). The harmonic mean is the reciprocal of the mean of the reciprocals:

$$1/N_e = (1/t)\Sigma_{i=1}^{t}(1/N_i) \text{ for } t \text{ generations}$$

In practice, this means that $N_e$ is disproportionately affected by the smaller population sizes. So in the recently expanded human population, the long-term effective population size (and hence the amount of neutral variation) is still largely determined by the smaller ancestral population sizes in our past. Table 6.4 gives estimates of $N_e$ in humans, and Figure 12.6a shows estimates of the population growth over the past 100 KY.

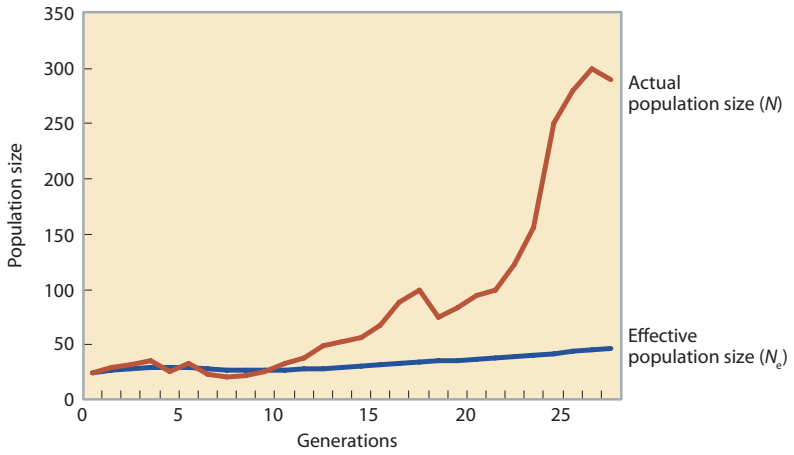



**Figure 5.7: $N_e$ in a population of variable size.**
The harmonic mean of the census size barely changes despite a recent population expansion.

This dependence of present-day variation on past small population sizes brings us to two important population processes that shape the genetic diversity apparent in many human populations: size-reduction **bottlenecks** and **founder effects**. In many respects the two processes are similar because both involve reduced population size, but the difference between them can be seen in **Figure 5.8**. Founder effects relate to the process of colonization and the genetic separation of a subset of the diversity present within the source population. In contrast, bottlenecks refer to the reduction in size of a single, previously larger, population and a loss of prior diversity.

The Wright–Fisher model assumes that all parents have an equal chance of contributing to the next generation. This results in a **Poisson distribution** of numbers of offspring. However, in real human populations there is often substantial variation in the contribution of individuals to the succeeding generation. To put it another way, there is a higher variance in the number of offspring than that expected under a Poisson distribution (where the variance equals the mean). This can be due to social causes, and need not be attributed solely to

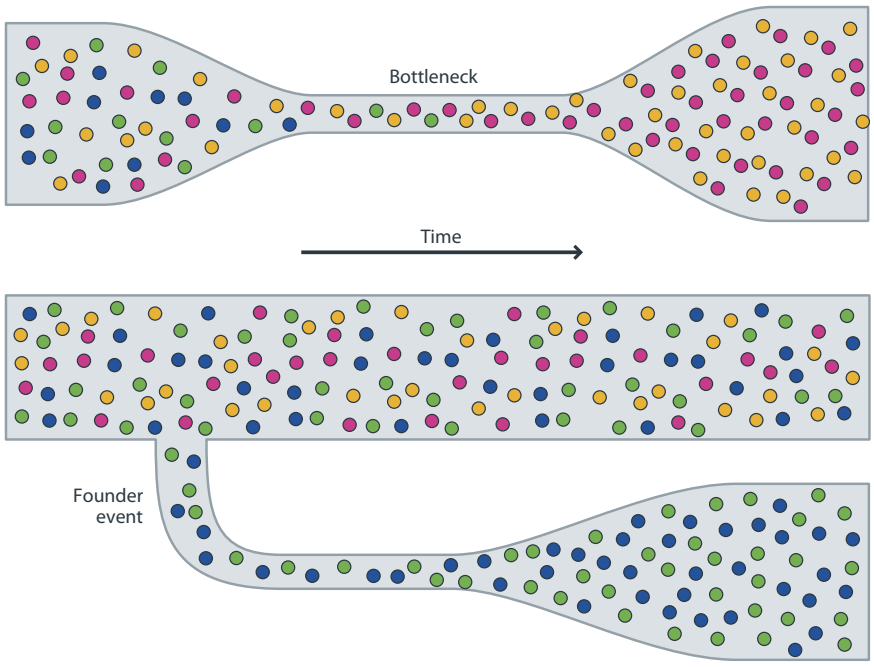



**Figure 5.8: Bottlenecks and founder events.**
Circles of different colors represent different alleles. Both bottlenecks and founder events result in a loss of allelic diversity.

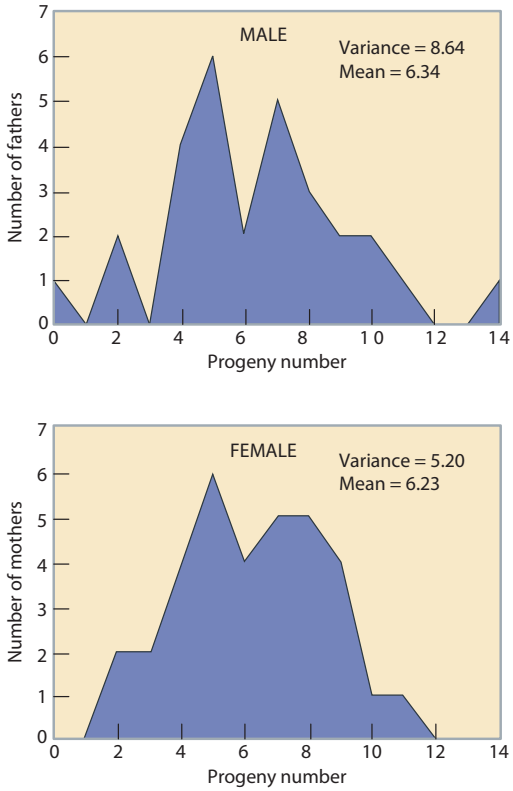



**Figure 5.9: Difference in reproductive variance for male and female Aka pygmies.**
The two sexes have almost the same mean value, but different variances. [Data from Hewlett BS (1988) in Human Reproductive Behaviour: A Darwinian Perspective. Betzig L, Borgerhoff Mulder M, Turke P (eds). Cambridge University Press.]

differences in **fertility**. The higher the **reproductive variance**, the lower the effective population size, because parental contributions become more and more unequal.

Because reproductive variance often differs between the sexes (see **Figure 5.9**), males and females may have different effective population sizes. Most anthropological studies show males to have higher reproductive variance than females, which is expected to result in a lower male effective population size. For example, using demographic data from Inuit hunter-gatherers of Greenland, the effective population size of males is estimated to be half the male **census population size**, whilst for females it is 70–90% of the female census population size, and this is due to a higher male reproductive variance.[31] This has implications for the effective population sizes of portions of the genome with different inheritance patterns.

There is a further reduction in effective population size when reproductive variance is correlated between generations, for example, when children of large families tend to have large families of their own. Biologically, this inheritance of **fecundity** could happen when a gene conferring greater fertility is polymorphic within a population. Alternatively, social mechanisms of inherited fertility may operate in structured societies where access to resources is both unequal and inherited. Whatever the cause, inheritance of family size has been noted in many human populations, from different types of demographic data (**Box 5.2**). In 1932, Huestis and Maxwell used completed questionnaires from University of Oregon students to demonstrate a significant correlation between the number of siblings of their parents and the number of children of their parents.[22] Alternatively, genealogical records can detail past inheritance of family size, as has been demonstrated with the Saguenay-Lac Saint Jean population in Quebec[2] and the British nobility.[39]

### Population subdivision can influence effective population size

Previously we have considered only randomly mating populations; however, most human populations are not so homogeneous. In one respect, all human mating is nonrandom because it usually involves a conscious choice, but in our present context "random" means only that mating is random with respect to the genetic make-up of each individual. A population may be nonrandomly mating because it consists of smaller, partially isolated **subpopulations**, also known as **demes**. Alternatively, nonrandom mating may also occur because mate choice is not blind to genetic relatedness.

**Population subdivision** is often modeled in terms of a **metapopulation** that comprises partially isolated subpopulations. This isolation eventually leads to partial **genetic differentiation** as genetic drift operates independently within each subpopulation. Members of the same subpopulation are therefore more closely related, on average, than are members of different subpopulations. Depending on the nature of the **population structure**, the effective population size of the metapopulation can be increased or decreased relative to a randomly mating population of the same size. If there are substantial levels of extinction and recolonization of subpopulations then the effective population size of the metapopulation can be dramatically reduced relative to the census size.

If subpopulations are not completely isolated, then the migration of individuals between them results in gene flow, reducing differentiation. So to understand the impact of population subdivision on genetic drift we must model: (1) the number, size, and spatial arrangement of the subpopulations; and (2) gene flow by migration. These models are considered in greater depth in **Section 5.5**. One aspect shared by all these models is the specification of a measure of population structure that is used to estimate parameters such as the rate of gene flow,

#### Box 5.2: The $F_{ST}$ statistic

$F_{ST}$ is a statistic that was developed independently by Sewall Wright and Gustave Malécot in the 1940s and 1950s, and is possibly the most widely used statistic in population genetics. It is in fact one of a family of statistics called the **fixation indices** that measure the deviation of observed heterozygote frequencies from those expected under Hardy–Weinberg theorem.[20, 45]

$F_{ST}$ measures the apportionment of genetic variation between subpopulations; in other words, it compares the genetic diversity found within subpopulations (the "S" of the subscript) to the genetic diversity of the total population (the "T" of the subscript). It can also be regarded as measuring the proportion of genetic diversity due to allele frequency differences among subpopulations.

$F_{ST}$ varies between 0 and 1, can be defined in a number of different ways, and can be estimated from genetic diversity data by a variety of methods, most commonly:

$$F_{ST} = \frac{(H_T - H_S)}{H_T}$$

where $H_T$ is the expected heterozygosity of the entire population and $H_S$ is the mean expected heterozygosity across subpopulations.

For use as a genetic distance, $F_{ST}$ can be formulated to compare two populations (known as pairwise $F_{ST}$) and can be defined as:

$$F_{ST} = \frac{V_p}{p(1-p)}$$

where $p$ and $V_p$ are the mean and variance of gene frequencies between the two populations respectively.

Wright suggested that qualitative guidelines shown in **Table 1** could be used to interpret $F_{ST}$ values. Using these guidelines, humans, with a genomewide average $F_{ST}$ value of around 0.05, show little to moderate genetic variation.

**TABLE 1: SEWALL WRIGHT'S QUALITATIVE GUIDELINES FOR INTERPRETING $F_{ST}$**

| $F_{ST}$ values | Level of genetic differentiation |
|---|---|
| Less than 0.05 | little |
| Between 0.05 and 0.15 | moderate |
| Between 0.15 and 0.25 | great |
| Greater than 0.25 | very great |

or the effective population size of the metapopulation. Some have argued that such estimates have little, if any, relevance to reality, because all current models are oversimplistic, and contain important assumptions that are violated by all human populations.

Perhaps the best-known measure of population structure is $F_{ST}$ (Box 5.2). When gene flow is high and there is little differentiation between subpopulations $F_{ST}$ is close to zero. When subpopulations are highly differentiated, then genetic diversity of the metapopulation is much greater than in any subpopulation and $F_{ST}$ is close to 1. $F_{ST}$ values between pairs of populations can also be considered to be a measure of genetic distance between them.

Subpopulation divergence results in an excess of **homozygotes** in the metapopulation and a corresponding deficiency of **heterozygotes**, a phenomenon known as the **Wahlund effect**.

### Mate choice can influence effective population size

Nonrandom mating also results from individuals choosing their mates via some assessment of their mutual similarity. If individuals choose partners on the basis of shared phenotypic characteristics such as socioeconomic status, IQ, or skin color, this is known as **assortative mating**.

Assortative mating can be based on physical, **psychometric**, or cultural traits. Physical traits that are selected include similar attractiveness, age, and ethnicity. The last can be demonstrated by the statistical analysis of census data, whereas the first is trickier as it relies upon a subjective notion of beauty. Nevertheless, it has been argued that individuals do choose to mate with those of a similar level of attractiveness to themselves. Psychometric traits thought to have been selected during assortative mating include IQ, and the presence of a mental disorder. Other relevant traits include religion, deafness, and educational qualifications.

**Disassortative mating** (or negative assortative mating) results when partners are chosen on the basis of their phenotypic differences rather than similarities. Disassortative mating at a locus generates a greater heterozygote frequency, and assortative mating a lower heterozygote frequency, than that expected under random mating. Assortative mating augments genetic drift by decreasing the effective population size, whereas the opposite is true for disassortative mating. One of the best-known traits proposed as a candidate for disassortative mating is resistance to infectious diseases. Much of an individual's resistance is encoded in the MHC region of the genome (**Box 5.3**). This region contains several closely linked and highly polymorphic genes that are involved in immunological recognition and response. Disassortative mating is one of a number of plausible explanations for the surprisingly high degree of polymorphism in the MHC region.

Inbreeding and **outbreeding** occur when mating happens between individuals who are respectively more, or less, related than would be expected by random mating. The more closely related the two partners are, the higher the chance that they will pass on the same deleterious recessive allele to their offspring. Thus there is a fitness cost to inbreeding known as **inbreeding depression**. The degree of inbreeding, or **consanguinity** (from the Latin "of the same blood"), is measured by the **coefficient of kinship** ($f$), which is the probability that two alleles from two different individuals are identical by descent. An alternative measure is the **coefficient of relatedness**, which is simply equal to $2f$. Incest represents the extreme of inbreeding, and usually refers to sexual intercourse between close relatives. The definition of close relatives is usually regulated by religion or the state, with first-cousin marriage ($f = 0.0625$) allowed by many religions but marriage between closer relations generally proscribed. Incest taboos are nearly universal and may represent an adaptive behavioral strategy to minimize inbreeding depression. Nevertheless, there is evidence

of institutional incest in certain dynasties: for example, sibling marriage was expected of the ruling Egyptian Pharaohs during the eighteenth and nineteenth dynasties (~1400–1700 years ago).[3]

Current surveys record a significant rate of consanguineous marriages in certain countries: it has been estimated that 10% of marriages in the global population are between partners related as second cousins or closer ($f \geq 0.0156$). Modern studies suggest that consanguineous marriages result in an increase in the female reproductive life span and the consequent higher average number of children, at least for first-cousin marriages, may outweigh the negative effects of inbreeding depression (estimated at about 4% more pre-reproductive deaths in offspring of first-cousin marriages).[4] Given the small size and extensive dispersal of prehistoric hunter-gatherer groups, it may be reasonable to suppose that there were similar levels of inbreeding throughout human evolution, although studies on present-day hunter-gatherer societies suggest social factors may promote breeding between, rather than within, tribes.[19]

### Genetic drift influences the disease heritages of isolated populations

Some human populations exhibit high incidences of multiple genetic diseases that are rare in surrounding populations. They appear to have a distinct heritage of genetic disease. These populations also show high frequencies of usually rare, but neutral, alleles. Often these groups are known to have undergone demographic processes that have resulted in small effective population sizes: for example, founder effects [for example, Finns (**Section 16.2**), Afrikaners in South Africa] and **endogamy** (within-group marriage, for example, Roma).

However, in some cases, where there is good evidence that a disease allele has been imported into a population by a single founder, it appears that insufficient time has elapsed for genetic drift alone to account for the high frequency. An example is the increase in carrier frequency for the disorder of amino acid metabolism, tyrosinemia I (OMIM 276700), from 1/5000 to 1/22 within 12 generations in the Saguenay-Lac Saint Jean population of Quebec. In such cases it is tempting to invoke some form of selective process. However, a more sophisticated appreciation of the demographic factors underpinning genetic drift often provides an adequate explanation. In the Quebec case, inheritance of family size, which was well documented in the genealogical records for this population, increases genetic drift sufficiently to account for the observed carrier frequencies.

## 5.4 THE EFFECT OF SELECTION ON DIVERSITY

Natural selection, as defined by Darwin and elaborated by Fisher, is the differential reproduction of individuals of different genotypes in sequential generations. Genotypic variation produces individuals with varying capacities to survive and reproduce in different environments. Selection can occur at any stage on the long journey from the formation of a genotype at fertilization to the bearer of that genotype generating their own viable progeny, including:

- Survival into reproductive age—viability and mortality
- Success in attracting a mate—sexual selection
- Ability to fertilize—fertility and gamete selection (**meiotic drive**)
- Number of progeny—fecundity

The sum of these is the ability of an individual genotype to survive and reproduce, its **fitness**, which is partly dependent on the environment. The important factor is the relative fitness of a genotype compared with other genotypes competing for the same resources. Relative fitness is measured by a **selection coefficient** ($s$), which compares a genotype with the fittest genotype in the population. A selection coefficient of 0.1 represents a 10% decrease in fitness compared with the fittest genotype.

**Box 5.3: The major histocompatibility complex**

### What is it?

When a tissue is transferred from one individual to another, it may be rejected or accepted by the host immune system; this is known as histocompatibility. Although a number of loci throughout the genome are involved in histocompatibility, in humans the major determinants are found in a large gene cluster on chromosome 6 known as the **major histocompatibility complex** (**MHC**). The different MHC-encoded proteins that can be recognized by the immune system are cell-surface proteins each known as a human leukocyte antigen (HLA). HLAs include proteins that are expressed on all nucleated cells.

### How is the locus arranged?

Due to its medical importance, the gene-dense MHC locus was one of the first large regions to be sequenced during the Human Genome Project. The 3.6-Mb locus is divided into three regions, called classes (see **Figure 1**). Ancient gene duplication events have generated several expressed HLA genes and many pseudogenes within the class I and II regions.[21] These HLA genes are involved in the development of **adaptive immunity** through the presentation of bacterial and viral **antigens** to T lymphocytes. Different alleles at each individual gene vary in their ability to present antigens from different pathogens.

### How diverse is it?

The most unusual feature of the MHC is the huge amount of variation contained within it. The HLA Sequence Database (http://www.ebi.ac.uk/imgt/hla/) currently contains over 8000 allele sequences from 35 different loci within the MHC (see Figure 1; and Figure 3.14).

As well as the sheer number of alleles, the differences between them are often many times greater than at other loci (many alleles at the same HLA locus differ by 5–17% of all nucleotides, whereas most alleles at other loci differ by less than 1%), indicating that their common ancestry is ancient. Many of these alleles are so old that they pre-date the human–chimpanzee split: that is, a human allele may be more closely related to a chimpanzee allele than to an alternative human allele, a characteristic known as **trans-species polymorphism**. In addition, there is also variation in the **copy number** of HLA-DRB genes.

### Why is it so diverse?

High MHC diversity within modern humans could be explained by selection for diversity or by an elevated mutation rate. However, trans-species polymorphism can only be explained by the operation of selection in preventing the fixation of alleles over time. Three alternative selection pressures have been proposed, although they are not mutually exclusive:

- Heterozygote advantage—individuals with heterozygous MHC haplotypes are better able to resist infectious disease as a result of having a broader spectrum of antigen-binding specificities.
- Frequency-dependent selection—low-frequency alleles are favored if pathogens have evolved to evade immune detection in individuals carrying the higher-frequency alleles.
- Disassortative mating—mate preference for dissimilar MHC haplotypes would maintain a highly diverse MHC in a population.

A role for selection is supported by the concentration of variation in the exons coding for the antigen-binding groove of the protein,[23] presence of very old haplotypes,[41] and peaks of noncoding variation around variable genes, suggesting hitchhiking along with the balancing selection operating at these loci. Evidence for disassortative mating is contradictory; in support of this idea women show some evidence of preferring body odor of MHC-dissimilar men, and genetic evidence from a moderately inbred Anabaptist group, the Hutterites, suggests that married couples are more likely than by chance to be MHC-dissimilar. However, some studies have failed to support the evidence for a female body-odor preference, and studies of other, more outbred, populations find no evidence of disassortative mating at the MHC.[17]

### What role does recombination play in generating diversity?

Recombination within heterozygous HLA genes creates new alleles, and interallelic and intergenic gene conversion generates additional variation. The MHC exhibits a high degree of linkage disequilibrium (LD) that most likely results from the localization of recombination events to certain hotspots within the locus, between which lie long regions of low recombination. As a consequence, linked MHC genes are frequently co-inherited in haplotype blocks (Section 3.8), making it easy to identify disease-related haplotypes, but difficult to locate disease-related alleles to a single gene. For example, the tightly linked class II loci, DRB1, DQA1, and DQB1, are often found to be in complete LD.

### But what about selection?

Through linkage analysis and **association studies**, numerous MHC haplotypes have been associated with susceptibility to, and protection against, different diseases.

These include infectious diseases (for example, the protective effect of HLA-B*53:01 against malaria), autoimmune disorders [for example, susceptibility to multiple sclerosis (OMIM 126200) conferred by HLA-DRB1*15:01], and other diseases [for example, HLA-DQB1*06:02 predisposes to narcolepsy (OMIM 605841)]. Particularly interesting is the HLA-B*57 allele, which is significantly protective against HIV but is very strongly associated with the inflammatory disease ankylosing spondylitis (OMIM 142830). Because of the medical importance of these associations, the MHC haplotype project has sequenced eight full MHC haplotypes that are commonly associated with type 1 diabetes (OMIM 222100) and multiple sclerosis, to help disentangle these associations and determine their functional basis.

## Can HLA variation be used to explore the human past?

The current geographical distribution of HLA alleles is shaped to some degree by events in the human past. High diversity at the protein level facilitated extensive study of these loci prior to the advent of DNA-based methods. However, the association of different MHC haplotypes with many different diseases raises the possibility that the spatial distribution of HLA alleles may be shaped not by population history, but by different selective environments. Selection can be expected to skew the frequencies not only of disease-related alleles, but also of alleles at any linked loci.

## How are different alleles named?

The **nomenclature** for the different alleles has been complicated by the use of two different methods to define alleles. Initially, **serological** methods that detect some but not all variation at the protein level were used to identify alleles. More recently, direct analysis of DNA sequences at this locus has revealed that multiple alternative DNA sequences can encode the same serologically defined allele. Thus the nomenclature has evolved to include information on the gene at which the allele is found, the serological allele, and the underlying DNA sequence. For example, the serological allele HLA-A1 (the first allele at the A locus within the HLA) can be encoded by the DNA allele HLA-A***01**:01 or HLA-A***01**:02. The first two numbers (shown here in bold) define the serological allele, and the second two numbers define different nonsynonymous changes that yield different proteins with the same **immunoreactivity**. The colon between these numbers is a convention introduced in 2010 and many reports show allele names without it. A third level of numbers can be added to indicate any synonymous changes that might be present. So, for example, the two alleles that give the same HLA-A*01:01 protein sequence but differ by a mutation that does not cause an amino acid change, are defined as HLA-A*01:01:01 and HLA-A*01:01:02. A fourth level of numbers can be added to indicate any nucleotide differences in the noncoding regions. As an added complication, some of the serological alleles have been given new names for the purposes of the DNA naming system; so HLA-DR17 has become HLA-DRB1*03, of which there are 84 nonsynonymous alleles (HLA-DRB1*03:01 to HLA-DRB1*03:84), some of which have synonymous variants (for example, HLA-DRB1*03:05:01 and HLA-DRB1*03:05:02).

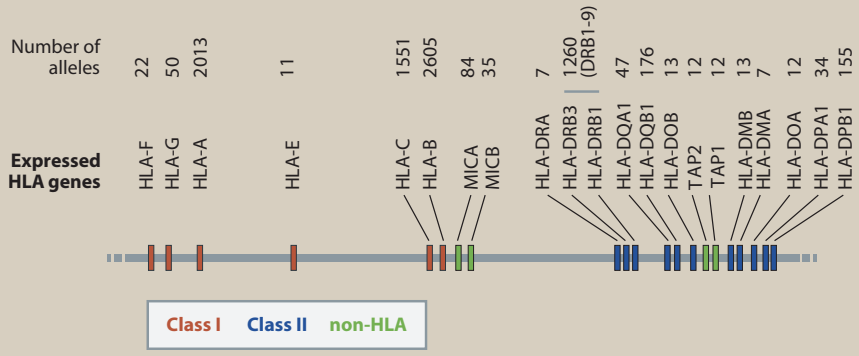



**Figure 1: Structure of the MHC region on chromosome 6, showing the location of the genes and the number of alleles reported in the HLA Sequence Database.**

Mutations that reduce the fitness of the carrier are subject to **negative selection**, also known as **purifying selection**, whereas mutations that increase fitness undergo positive, or **diversifying selection**. However, to understand the dynamics of selection at diploid loci we must consider the impact of mutants on the fitness of the genotypes, and not on the individual alleles. The two alleles within a diploid genotype can interact to determine the phenotypic fitness of an organism in different ways. This in turn affects the efficiency of natural selection in fixing or eliminating novel alleles. For example, a novel deleterious allele will be eliminated more rapidly from the population if it reduces the fitness of a heterozygote. Alternatively, a new allele may increase the fitness of a heterozygote relative to that of both homozygotes. The two homozygous genotypes may exhibit different reductions in fitness ($s_1$ and $s_2$). Such selection is known as **overdominant selection** (also known as **heterozygote advantage**) and creates a **balanced polymorphism**. By contrast, **underdominant selection** operates where new alleles reduce the fitness of the heterozygote alone. Several different selective regimes are summarized in **Table 5.3**.

Overdominant selection is not the only mechanism by which balanced polymorphisms can be generated, but is one of a number of processes described collectively as balancing selection (Box 6.6). An alternative mechanism is **frequency-dependent selection**, whereby the frequency of a genotype determines its fitness. If a genotype has higher fitness, relative to other genotypes, at low frequencies, but lower fitness at higher frequencies, an intermediate equilibrium value will be reached over time. Box 5.3 describes the MHC region, where genes have been suggested to be under both frequency-dependent and overdominant selection.

Other classic examples of balanced polymorphisms in humans are those that protect against **malaria** when heterozygous but have a reduced fitness compared with wild-type when homozygous, as a result of red blood cell disorders. A number of these types of balanced polymorphisms have arisen in different areas of malarial **endemicity**. The best known is the sickle-cell anemia allele of the β-globin gene, $Hb^S$ (OMIM 603903), which dramatically reduces fitness when homozygous. Malarial endemicity is not spread equally across the world, and as a consequence these balanced polymorphisms exhibit a limited geographical range that closely parallels that of malaria (**Figure 5.10**, **Section 16.4**).

Even small selective forces are capable of causing appreciable changes in allele frequencies over many generations. The inter-generational change in allele frequencies can be calculated by incorporating the selection coefficients described in Table 5.3 into the Hardy–Weinberg theorem. **Figure 5.11** compares the selection dynamics of a low-frequency advantageous allele under positive and **co-dominant** selection. It can be seen that selection achieves the most

**TABLE 5.3:**
**DIFFERENT TYPES OF SELECTION, AND THEIR EFFECTS ON GENOTYPE FITNESS**

| Type of selection | Genotype fitness | | |
|---|---|---|---|
| | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
| Simple negative/positive selection ($A_2$ is recessive) | 1 | 1 | $1 - s$ |
| Simple negative/positive selection ($A_2$ is dominant) | 1 | $1 - s$ | $1 - s$ |
| Co-dominant selection | 1 | $1 - s$ | $1 - 2s$ |
| Overdominant selection | $1 - s_2$ | 1 | $1 - s_1$ |
| Underdominant selection | 1 | $1 - s$ | 1 |

Note: $s$ = selection coefficient.

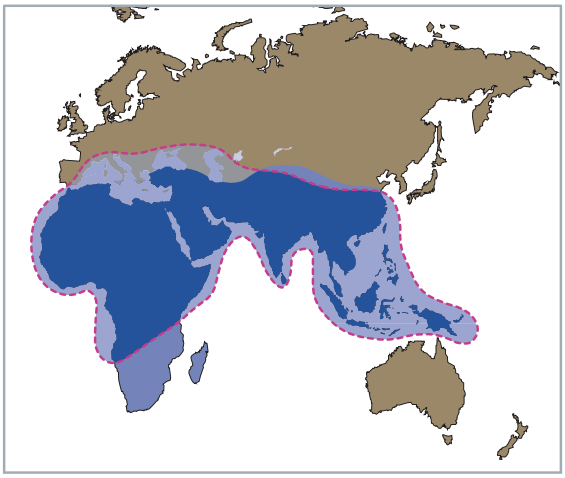


**Figure 5.10: The overlapping geographical distributions of red blood cell disorders and malaria.**
*Blue* indicates regions of current malarial incidence and the red dashed line shows the distribution of red blood cell disorders.

rapid changes in allele frequencies when alleles are at intermediate frequencies. However, in **Section 5.6** we shall see that other processes acting on allele frequencies may outweigh small selective forces.

### Mate choice can affect allele frequencies by sexual selection

In cases of limiting numbers of available partners, selection can operate at the level of **mate choice**. For humans, where the levels of investment of males and females in their offspring are unbalanced, the availability of females is the factor limiting male reproduction and therefore females can exercise mate choice. Desirable traits may be those that indicate health, access to resources, and ability or willingness to invest in offspring. An alternative mechanism of sexual selection in response to limited female mating resources is that of competition between males:

> We may conclude that the greater size, strength, courage, pugnacity, and energy in man, in comparison with woman, were acquired during primeval times, and have subsequently been augmented, chiefly through the contests of rival males for the possession of the females. (Charles Darwin, *Descent of Man*, 1871)

If these attractive or competitive traits are to some degree genetically determined, then the loci responsible are said to be under sexual selection. Mate choice can be differentiated from assortative mating on the basis that specific preferences are shared among all members of the same sex.

Darwin invoked sexual selection (Box 15.4) to explain the presence of secondary sexual characteristics among humans. Others have proposed that the human mind itself is largely a result of this selective process.[33] So far, there

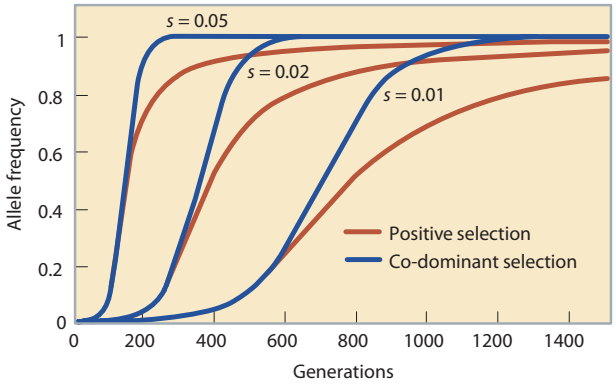



**Figure 5.11: Positive and co-dominant selection for an advantageous allele.**
The selection dynamics of a low-frequency, advantageous allele are compared under positive and co-dominant selection. Selection achieves the most rapid changes in allele frequencies when alleles are at intermediate frequencies.

have been very few studies in humans, mainly because the traits possibly under sexual selection in humans show complex multi-genic inheritance that has yet to be elucidated. Studies of sexual selection in humans have been limited to observing indirect correlations between phenotypes and reproductive success, an example being the higher number of offspring born to taller men.[38] Sexual selection, whether genetically or culturally determined, could be expected to lower the effective population size of the selected sex by increasing the reproductive variance.

## 5.5 MIGRATION

Unlike genetic drift, mutation, and selection, migration cannot change specieswide allele frequencies, but it is capable of changing allele frequencies in populations. It thus belongs to a second tier of population processes that shape human genetic diversity. As noted previously, gene flow counteracts genetic differentiation and is modeled within the framework of a larger, subdivided, metapopulation.

First, we must be clear on some definitions, because they are often used interchangeably in the literature. Colonization is the process of movement into previously unoccupied land, thus entailing a founder effect. By contrast, **migration** is the movement from one occupied area to another. Gene flow is the *outcome* of a migrant contributing to the next generation in their new location. Thus, to observe gene flow directly we not only need to monitor the movement of migrants but also their reproductive success. Estimates of gene flow have, therefore, relied upon indirect methods that assess allele frequency differences among populations using simplified models.

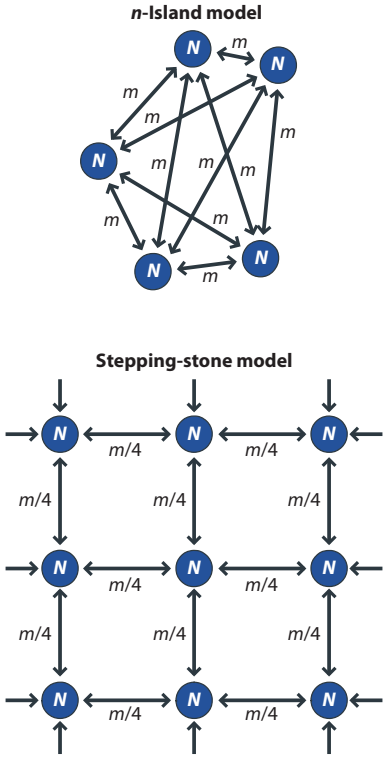



**Figure 5.12: The *n*-island and stepping-stone models of gene flow.** Each diagram represents one of a family of models: the N-island model, and the two-dimensional stepping-stone model, also known for obvious reasons as the lattice model. *N*, population size; *m*, rate of exchange of genes per generation.

### There are several models of migration

Perhaps the simplest model of gene flow is the ***n*-island model** devised by Sewall Wright. A metapopulation is split into islands of equal size $N$, which exchange genes at the same rate per generation, $m$ (where $m$ represents a proportion of the population migrating rather than an actual number). Under the assumptions of this model the rate of migrant exchange can be related directly to $F_{ST}$ (Box 5.2), by the equation:

$$F_{ST} = 1/(1 + 4Nm)$$

The assumptions of the $n$-island model include:

- No geographical substructure apart from the division into islands: all islands are equivalent
- Each population persists indefinitely
- No mutation
- No selection
- Each population has reached equilibrium between mutation and drift
- The migrants are a random sample from the source island

The **stepping-stone model** seeks to remove one obvious flaw of the $n$-island model—the lack of geographical substructure. The stepping-stone model introduces the idea of geographical distance by only allowing the exchange of genes between adjacent discrete subpopulations. Figure 5.12 shows a comparison of the $n$-island and stepping-stone models. The stepping-stone model also assumes equal rates of migration between subpopulations. Both kinds of model have been used to show that even very low rates of migration between subpopulations are capable of retarding their genetic differentiation.

Migration can be modeled as occurring within a continuous population, rather than discrete subpopulations, by considering that mating choices are limited by distance, and that these distances are typically less than the overall range

of the population. This is the basis for **isolation by distance** (IBD) models. Within such models, genetic similarity develops in neighborhoods as a function of dispersal distances. These can be thought of either as the difference between birthplaces of parent and offspring, or marital distances. Different mathematical functions have been used to relate the decline in frequency of these dispersals to geographical distance. Once the system has reached equilibrium between gene flow and the differentiation caused by genetic drift, genetic similarity declines over distance in a predictable fashion. The stepping-stone model described above is a discontinuous example of IBD.

These migration models are mathematically tractable and can be generalized to many species. However, for many human populations (unlike those of other species) we often have detailed data on parameters such as migration rates, migration distances, and marital distances. The migration matrix model uses this detailed information and thus can incorporate different migration rates and asymmetric migration between subpopulations. In this way, a more complex and realistic relationship between distance and migration is obtained. Nevertheless, it seems unlikely that present-day migration rates have been constant for long enough to allow the system to reach a state of equilibrium, which is required by calculations using such models. The uneven pattern of most human habitation falls between the models of discrete subpopulations and uniform continuity assumed by the stepping-stone and isolation by distance models respectively. Furthermore, migration processes are far more complex than the current models allow. Migration processes often include long-distance movements as well as smaller-scale mating choices. The choice to migrate is taken by individuals on the basis of multiple "push" and "pull" factors, so that migration rates are rarely, if ever, symmetric between two populations. Migrants are seldom a random sample of their source population; they are often age-structured, sex-biased, and related to one another. The latter property of migrants is known as **kin-structured migration** and is well documented both ethnographically[12] and archaeologically.[1] In light of these complications, we should be cautious in attempting to estimate parameters of population structure and be skeptical of their relationship to reality.

### There can be sex-specific differences in migration

If we consider possible differences between the sexes in their migration behavior, an intuitive hypothesis on observing the modern world might be that men tend to migrate over longer distances than women: intercontinental migrants tend to be male-biased and recent history documents explorers, traders, and soldiers as being almost exclusively male. Involuntary migration, particularly slavery, is often sex-biased: for example the Atlantic slave trade involved mostly males, while the Indian Ocean/Red Sea slave trade involved mostly females. However, when considering the impact of migration on genetic diversity, we must not only examine long-distance migration patterns but also small-scale local migrations.

Marital residence patterns are critical to investigating local migration patterns. **Patrilocality** describes the phenomenon by which a female from one village, when marrying a man from a different village, takes up residence in the man's village. In contrast, **matrilocality** describes the situation when the husband moves to the wife's village. It has been estimated that roughly 70% of modern societies are patrilocal.[5, 35] In other words, in the majority of societies, mtDNAs are moving between villages each generation, whereas Y chromosomes are staying put. Similarly, the X chromosome is more mobile than the autosomes, as it passes down the female line twice as often as it does down the male line.

How might we determine which of the above current phenomena—the apparent male bias of long-distance migration or the female bias of intergeneration marital movement—has had a greater role in shaping modern genetic diversity?

To resolve this issue, we can compare the geographical patterning of genetic diversity of chromosomes that have different inheritance patterns. Migration

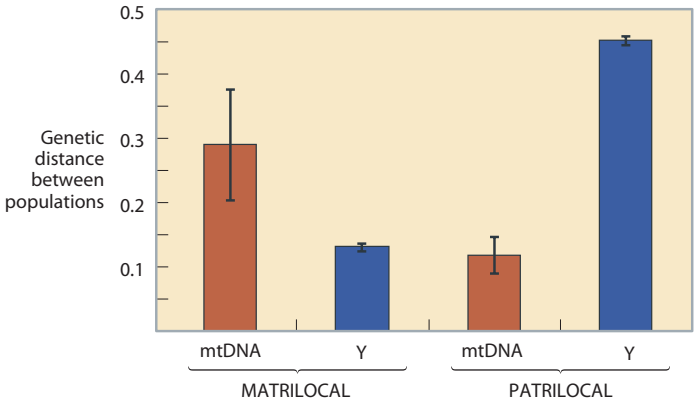



**Figure 5.13: Influence of sex-specific migration on genetic diversity.** [Data on three matrilocal and three patrilocal groups among the hill tribes of northern Thailand, from Oota H et al. (2001) *Nat. Genet.* 29, 20.]

reduces population differentiation, so by studying the relationship between geographical distances and genetic distances among the same set of populations, we can identify which loci appear to have been experiencing greater gene flow. These comparisons assume that migration rate is the only factor determining population differentiation; however, genetic drift also influences levels of population subdivision, and the effective population sizes of the Y chromosome and mtDNA need not be equal (Table 5.2). Consequently it has been argued that these differences between loci reflect differences in genetic drift as a result of sex differences in reproductive variance, and not migration rates. However, populations with a matrilocal pattern of marital residence show greater mtDNA than Y-chromosomal genetic differentiation (**Figure 5.13**). This suggests that a sex-specific difference in migration rate, and not drift, is responsible for the different patterns of genetic differentiation of mtDNA and the Y chromosome.

The sex-specific migration rate differs dramatically between populations and cultures, with a slight bias toward male migration overall (**Figure 5.14**). This suggests that long-range migration patterns (male-dominated) have contributed most to different patterns of diversity between mtDNA and Y chromosomes, but local effects can outweigh this overall pattern. Indeed, culture has an important effect: in sub-Saharan Africa, hunter-gatherer populations show matrilocality while pastoralist and agricultural populations show patrilocality.[43]

## 5.6 INTERPLAY AMONG THE DIFFERENT FORCES OF EVOLUTION

Thus far, we have examined individually some of the important factors influencing the level of variation in a population. Mutation, recombination, and

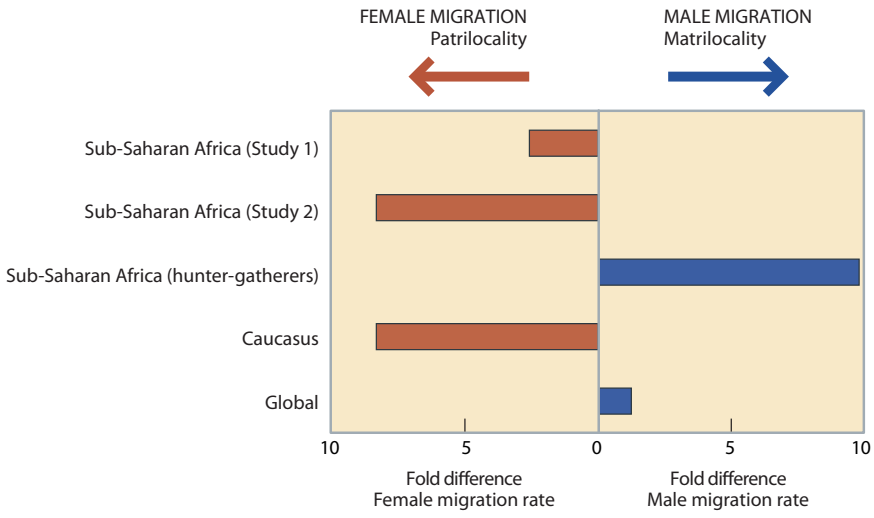



**Figure 5.14: Sex-specific differences in migration in humans.** Bars labeled Studies 1 and 2 reflect pattern in migration of sub-Saharan populations irrespective of lifestyle. [Study 1 is Destro-Bisol G et al. (2004) *Mol. Biol. Evol.* 21, 1673. Study 2 is Wilder JA et al. (2004) *Nat. Genet.* 36, 1122.]

migration increase diversity, random genetic drift decreases it, and selection can do either. In this section, we investigate how these opposing forces interact with one another.

In the previous section we saw that in a subdivided population the opposing forces of migration and drift can reach an equilibrium state whereby differentiation among subpopulations, as measured by $F_{ST}$, remains constant over time. It is only by assuming that this equilibrium has been attained that we can estimate migration rates from $F_{ST}$ values in real populations.

## There are important equilibria in population genetics

### *Mutation–drift balance*

In the simplest model of a population with no selection or migration and the usual assumptions of constant size and random mating, diversity will reach an equilibrium value where the number of novel variants (generated by mutation) entering the population is balanced by the number lost by drift. This is known as **mutation–drift balance** or **mutation–drift equilibrium**. There is a simple analogy to illustrate this point: imagine a water tank fed by a dripping tap at the top, with another tap at the bottom to let water flow out (**Figure 5.15**). Water will accumulate in the tank until the amount entering from the tap at the top (= mutations) is balanced by the amount lost through the tap at the bottom (= drift), leading to a stable water level (= diversity). If the mutation rate increases (more water in) or decreases (less water in), diversity at equilibrium will increase or decrease correspondingly. Similarly, if drift increases (opening the bottom tap) or decreases (closing the bottom tap), diversity will decrease or increase. This equilibrium value of diversity is known as the **population mutation parameter** (**θ** or **theta**), and is discussed in more detail in Chapter 6.

### *Recombination–drift balance*

Earlier in this chapter we saw that in an infinitely large population, linkage disequilibrium (LD, Chapter 3) decays over time as a result of recombination generating new haplotypes. However, random genetic drift is continually removing haplotypes from the population. As a consequence LD can reach an equilibrium value in finite populations. This equilibrium value of LD is determined by the **population recombination parameter** (**ρ** or **rho**). This parameter combines information on effective population size and recombination rate ($c$) using the equation:

$$\rho = 4N_e c$$

The precise relationship between different measures of LD (Box 3.5) and ρ is complex, but when ρ is large

$$r^2 \approx 1/\rho$$

Thus we can see that LD decreases as ρ increases, for example as a result of a larger effective population size (lower genetic drift) or higher recombination rate.

In real populations, it is apparent that LD is not simply an equilibrium between recombination and drift, but can be greatly affected by selection, mutation, gene conversion, and demography (Chapter 6). Because demography influences LD, analysis of LD within a population can allow inferences to be made on the prehistoric demography of that population (for an example of how patterns of LD can reveal ancient admixture events, see **Section 14.4**).

If we are examining LD over a large genomic region containing many polymorphisms, it is unclear how best to combine the information from measures of LD based on comparisons between individual pairs of variants (that is, $D$, $D'$ or $r^2$, Box 3.5). Therefore attention has focused on estimating ρ itself for these kinds of data, as this gives a single measure of LD for the entire region. Estimating ρ requires the use of population models and is computationally intensive, but it allows the other forces that shape LD (for example, demography and mutation)

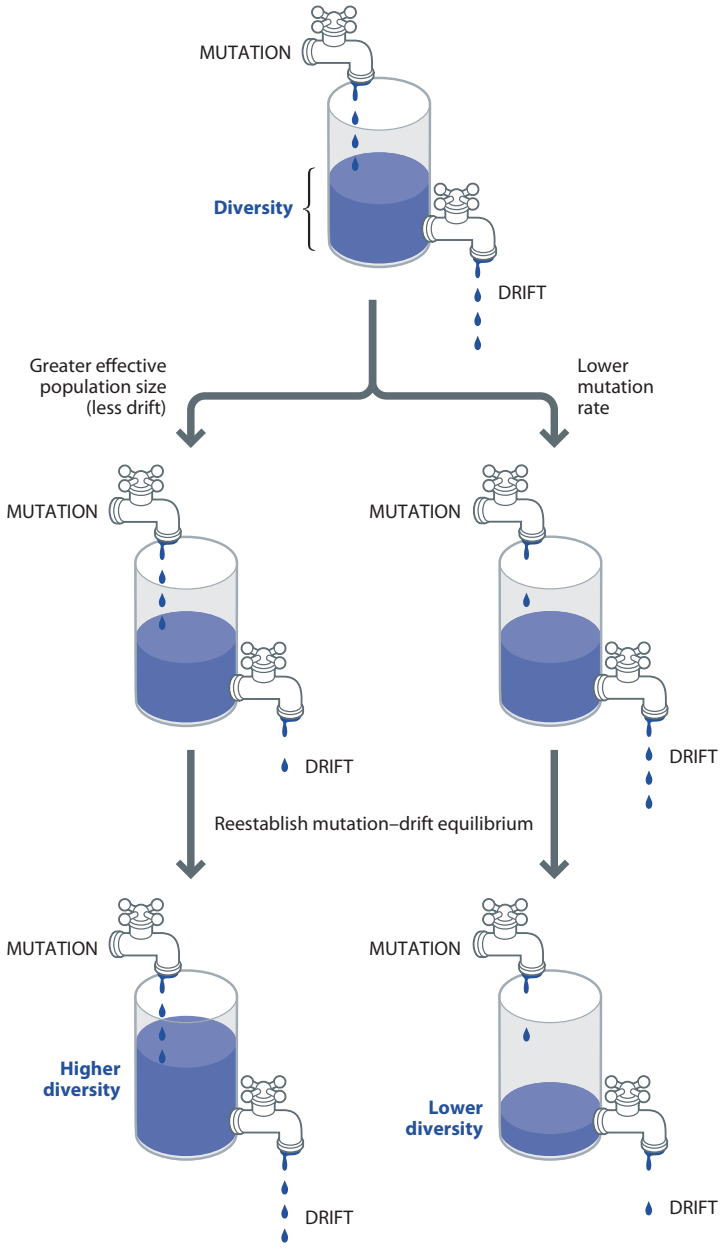



**Figure 5.15: A metaphorical depiction of the relationship between mutation rate, drift, and diversity.**
A change in either the mutation rate or effective population size changes the diversity at mutation–drift equilibrium—see Section 5.6. (Original metaphor by John Relethford.)

to be taken into account.[40] An additional advantage of studying ρ is that it allows *c*—the recombination rate across the region—to be estimated, and compared between different regions of the genome.

### *Mutation–selection balance*

Some deleterious mutational events have sufficiently high mutation rates that within a large population they occur several times within a single generation, and can be considered recurrent mutations. Mutation and selection are opposing forces determining the frequency of such mutant alleles in the population. The rate at which new alleles are generated by mutation can be balanced by the eventual elimination of each mutant allele by negative selection so that the average number of examples of a given mutant allele reaches an equilibrium value within the population. Lowering the selective cost of a mutation, or increasing its mutation rate, will increase this equilibrium value. Mutant alleles

that cause monogenic diseases are often likely to be in mutation–selection balance and this is discussed in Chapter 16.

Rather than considering a single recurrent mutation in isolation, if we consider all deleterious alleles together, a balance between mutation and selection may operate over the genome as a whole, such that at equilibrium each genome contains a certain number of deleterious alleles. This has implications for the study of complex diseases where both genetic variation and environmental variation play a role (Chapter 17).

### Does selection or drift determine the future of an allele?

So far in this section we have not considered the interplay between selection and drift, and their relative weight in influencing allele frequencies. For example, the selection dynamics described in **Section 5.4** assume an infinitely large population. What happens in finite populations where random genetic drift is also operating?

Because drift operates more effectively in smaller populations, stronger selection is required to influence fixation or elimination of alleles. Whether drift or selection predominates depends on a number of factors, which include:

- The effective population size
- The selection coefficient
- The type of selection
- The frequency of the allele under selection

Equations relating these parameters exist for different types of selection. They can be used to determine whether an allele is likely to be under the influence of selection. Relating these parameters together allows us to draw four important conclusions:

- Selection often substantially increases the probability that an **advantageous allele** becomes fixed compared to a **neutral allele**; in humans, most new advantageous alleles are still far more likely to be eliminated than fixed.
- If new alleles are almost exclusively deleterious, the optimal allele can persist unchanged over very long time-scales. This conforms to the hypothesis that functional constraint on important proteins such as histones underlies their extreme lack of variability among diverse species.
- The time taken to fix an advantageous allele is much shorter than that to fix a neutral allele.
- A general rule for diploid loci is that for selection to be effective then the following relationship should hold:

  $$s > 1/2N_e$$

  where $s$ is the selection coefficient.

  For haploid loci that are transmitted by only one sex, with one-quarter the effective population size of diploid loci, the relevant rule is:

  $$s > 2/N_e$$

The use of this last rule can be seen in the following example. A polymorphic inversion on the human Y chromosome, present in roughly 70% of British males, protects against XY translocations during meiosis. The offspring resulting from these rearrangements (XX males and XY females) are infertile; infertility is evolutionary death for the individual. However, these translocations occur only at low rates, such that the selection coefficient ($s$) of this inversion has been calculated as 1/90,000.[24] Given that the Y chromosome is a haploid locus and the effective population size ($N_e$) of humans has been estimated at around 10,000 (Table 6.4), the selective advantage of this inversion is not sufficiently large to overcome the effects of drift: $1/90{,}000 < 2/10{,}000$).

## 5.7 THE NEUTRAL THEORY OF MOLECULAR EVOLUTION

Before information on molecular diversity became available it was hypothesized that **genetic load** (that is, the accumulation of deleterious alleles) would mean that only a limited amount of polymorphism would be compatible with a sustainable population, because most new mutations would be selected against. Muller famously predicted that only one in a thousand genes would be heterozygous. According to this view, polymorphisms were stable entities, maintained by balancing selection. In contrast, the substitution of one nucleotide for another in a DNA sequence of a species was the result of positive selection for new mutations spreading through the population to **fixation**. The process of DNA substitution during evolution and the processes affecting the frequency of polymorphisms within a species were therefore thought to be independent.

The first studies, in the 1960s, to measure genetic diversity directly used protein electrophoresis to measure the frequency of **allozymes** in populations. Their finding was startling at the time: the amount of polymorphism uncovered within human genomes, and those of other species, was many times greater than was expected. To explain this, Motoo Kimura developed the **neutral theory of molecular evolution**, often referred to as simply the **neutral theory**.[25, 36] Neutral theory states that the fate of mutations is largely determined by random genetic drift rather than selection. The theory holds that negative selection is the prevailing mode of selection that eliminates deleterious mutations whereas cases of positive and balancing selection are rare. The vast majority of polymorphisms observed in populations are transient, awaiting eventual fixation or elimination by genetic drift. The theory therefore predicts that most polymorphisms have little or no effect on fitness. Kimura showed that, for a polymorphism where $2 N_e s \leq 1$ (a selective advantage of <0.00005 in humans) then genetic drift would determine the fate of that polymorphism.

The publication of the theory caused a polarized debate in population genetics. A consensus may now have been reached, suggesting that the neutral theory does not entirely explain the observed genetic variation and adaptation in humans and other species, but is nevertheless a useful conceptual framework for thinking about genetic variation and evolution. In particular, it provides a powerful null model against which empirical data can be tested for evidence of selection, and it is in that context that we show it being applied to human evolutionary genetics in Chapter 6, and throughout this book.

### The molecular clock assumes a constant rate of mutation and can allow dating of speciation

Perhaps the most important result of the neutral theory is that it presented a unified model of allele frequencies in a population and molecular substitution rates in an evolutionary lineage. The neutral theory shows that the rate of sequence evolution is driven only by the rate of mutation. Assuming the rate of mutation is constant, the rate of evolution is approximately constant over all evolutionary lineages. Therefore measuring the number of differences between the DNA sequences of two species can, in principle, be used to date the divergence of those two species,[27, 47] if calibration points of dated lineage divergences are known. This requires accurate, independent dating of the lineages by other disciplines, most notably **paleontology**. We will address applications of the molecular clock hypothesis and **genetic dating** in the next chapter; however, here we need to consider whether this is a reasonable approximation of the evolutionary process.

The mutation rate and the substitution rate can be shown to be equal, independent of population size, by a simple mathematical proof. The rate of nucleotide substitution ($k$) is equal to the rate at which new mutations are generated ($2N\mu$), multiplied by their probability of fixation ($u$). In a population of size $N$, diploid

loci have a population size of $2N$, and the rate of nucleotide substitution ($k$) is therefore:

$$k = 2N\mu u$$

Remember that for a neutral mutation the probability of fixation is its frequency, which for a new mutation is the reciprocal of the population size ($1/2N$).

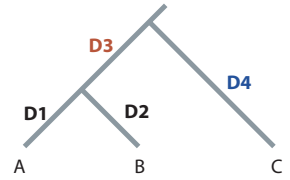
$$k = 2N(1/2N)\ \mu$$

$$\text{therefore—}k = \mu$$

The regularity of mutation might not translate into a regularity of evolution if selection plays a dominant role in determining the survival of new mutations in some lineages. Indeed, the regularity of molecular evolution contrasts with the non-uniform change of morphological evolution.

Fossil dates used to calibrate estimates of species divergence dates may be unreliable, or have unacceptably broad confidence limits. The **relative rates test** does not require absolute divergence times, but simply the knowledge of the order in which a number of lineages diverged from one another. This test (**Figure 5.16**) compares the rate of evolution in two lineages by relating them to a third, which is known to be an **outgroup** (a lineage more distantly related to the other two lineages than they are to one another).

Although the lineages used in the relative rates test are often individual species, the test can also be used on nonrecombining regions of the genome within a species (where the phylogeny is known). The number of mutational events in each branch of the tree relating the three lineages is calculated. The significance of any differences between the mutational distances shown in Figure 5.16 can be assessed by a number of different methods, such as the **likelihood ratio test** (Box 6.3).




If molecular clock true (**D1** = **D2**) then:
$\mathbf{D_{AC}} = \mathbf{D_{BC}}$ (**D1**+**D3**+**D4** = **D2**+**D3**+**D4**)
Thus, test to see if $\mathbf{D_{AC}} - \mathbf{D_{BC}} = \mathbf{0}$

**Figure 5.16: Testing the molecular clock with the relative rates test.**
The rate of evolution in two lineages, leading to A and B, is compared with that in an outgroup lineage, C. D1 to D4 are mutational distances on different branches of the phylogeny relating the three species.

### There are problems with the assumptions of the molecular clock

Comparisons of rates of evolution between species have often shown significantly varying rates of nucleotide substitution on different evolutionary lineages. Where the same rates are found across multiple loci it suggests that different mutation rates rather than selection are the cause of rate inequalities. There are several possible explanations, which relate the mutation rate to biochemical processes within the cell (**lineage effects**). The different processes that have been proposed to cause lineage effects need not be mutually exclusive, but could operate in concert (**Figure 5.17**).

The lineage effect that has received most attention is the **generation time hypothesis**. This assumes that most mutations occur during DNA replication in the germ line. As a consequence, the mutation rate is determined by the number of replications during a certain period of time. If species have the same number

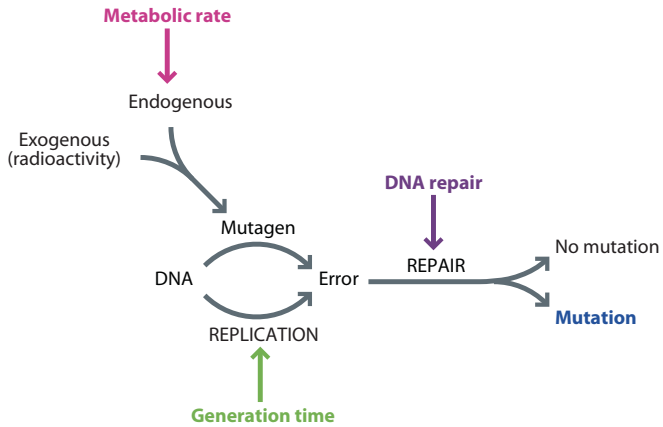



**Figure 5.17: Different sources of lineage effects.**
A number of processes are involved in generating a new mutation. Potential sources of lineage effects are shown in different colors.

of cell divisions per generation, then the mutation rate becomes dependent on the generation time. This hypothesis of replication errors causing mutations also underpins the hypothesis of **male-driven evolution** (see **Box 5.4**).

The generation time hypothesis has gained support from studies of mammalian species that show a better correlation of mutational distance with generation time than with calendar time. Many studies have demonstrated that while generation times within mammalian groups have the order rodents < monkeys < humans, evolutionary rates have the reverse order, both for indels and base substitutions. Similarly the mutation rate among higher primates seems to be negatively correlated with generation time. The mutation rate increase in rodents does not appear to be linear with respect to generation time as there is only a two- to fourfold increase in annual mutation rates relative to humans despite a fortyfold difference in generation times; however, the difference in replications per generation can account for some of this discrepancy. Human males have four times more replications per generation than do male mice, and this would be expected to lessen rate differences. Further problems for the generation time hypothesis appear in the difference in the substitution rate ratio of synonymous (silent) and nonsynonymous (amino-acid-changing) mutations (**Section 3.2**) between primates and rodents, suggesting a potential role for selection. In addition, there remain a number of cases where calendar time, rather than generation time, is better correlated with rates of evolution, even after correcting for the differing number of replications per generation among the different species.

### Box 5.4: Male-driven evolution

J. B. S. Haldane first proposed that the greater number of genome replications in the male germ line should result in the male mutation rate being higher than that of the female.[14] His hypothesis is based on the same assumption—that mutation is caused by errors in replication—as the generation time hypothesis discussed in Section 5.7.

In humans (Figure 3.27), the number of replications required during oogenesis is constant, at about 22, whereas the number of replications in the male germ line increases with age. Thirty replication events are required to generate spermatogonial stem cells at puberty (~13 years old), which then go through ~23 replications per year, before the final five replications required to make mature spermatozoa. Thus a male reproducing at 25 will be using DNA that has gone through $30 + 23 \times (25 - 13) + 5 = 311$ replication cycles. This is about 14 times as many as for the oocyte DNA.

The ratio of male to female mutation rates ($\alpha$—also known as the alpha factor) can be calculated by comparing the number of mutations that have accrued in autosomal, Y-chromosomal, and X-chromosomal sequences over the same time period. The ratios of these numbers can be related to $\alpha$ by the following equations:

$$\text{X/autosome} = 2/3(2 + \alpha)/(1 + \alpha)$$
$$\text{Y/autosome} = 2\alpha/(1 + \alpha)$$
$$\text{Y/X} = 3\alpha/(2 + \alpha)$$

Accurate estimates of $\alpha$ have been derived from analysis of large regions of genomic sequence. The estimates shown in **Table 1** incorporate a consideration of the diversity apparent in the human–chimpanzee common ancestor, which allows more accurate estimation of $\alpha$ than the equations given above.[42] Recent direct measurement of mutations from genome sequences of families[26] suggest a value of $\alpha$ of about 3.9.

**TABLE 1:**
**ESTIMATES OF $\alpha$ (95% CONFIDENCE INTERVALS) FROM HUMAN–CHIMPANZEE SEQUENCE COMPARISONS**

| Correction for ancient diversity | X/autosome | Y/autosome | Y/X |
|---|---|---|---|
| None | 7.58 (7.04–8.20) | 1.77 (1.64–1.94) | 2.68 (2.50–2.89) |
| 2× modern human diversity | 6.92 (6.38–7.58) | 3.04 (2.70–3.46) | 4.03 (3.67–4.43) |
| 4× modern human diversity | 6.11 (5.58–6.78) | 11.2 (8.04–17.6) | 8.24 (7.01–9.84) |

The **metabolic rate hypothesis** is based on the idea that most mutations result from the presence of **endogenous mutagens**.[30] Free radical by-products of aerobic respiration are the prime suspects, and, therefore, organisms with higher metabolic rates should produce more mutagenic free radicals. Differences in metabolic rates have been used to explain rate differences that were previously difficult to reconcile with generation time differences. In addition, considerations of metabolic mutagens may explain why rate differences are generally more pronounced among mitochondrial sequences than nuclear ones, as most oxidative free radicals are produced within mitochondria themselves.

Alternative sources of lineage effects could lie in the enzymatic mechanisms that act to repair the effects of mutagenic processes, rather than the processes themselves. While most attention has focused on varying efficiencies of **DNA repair**, another source could be differences in the many pathways that mop up mutagens of different kinds before they are able to damage DNA. At present, the relative efficiencies of these pathways in different lineages are too poorly characterized to allow these hypotheses to be tested.

In addition to rate variation amongst lineages there are often also differences between the rates of nonsynonymous and synonymous substitution. To explain this weakness of the molecular clock hypothesis, two alternatives have been proposed. The **episodic selection** model suggests that episodic selective pressures at nonsynonymous sites, created by environmental changes, distort the molecular clock.[13] The **nearly neutral** model suggests that nonsynonymous changes are slightly deleterious rather than neutral, so the interplay of drift and selection is sensitive to fluctuations in population sizes.[37] The idea is that negative correlation between population size and generation time allows the rate of nonsynonymous changes to operate largely independently of any generation-time effect. Small populations tend to have longer generation times and drift predominates, frequently fixing nonsynonymous mutations that occur at low rates. By contrast, negative selection predominates at nonsynonymous sites in large populations, so fixation occurs infrequently although mutations are generated more rapidly as a result of shorter generation times. These factors cancel out such that large and small populations have similar rates of evolution with respect to calendar time. This debate has not been resolved.

## SUMMARY

- Mutation and recombination increase human diversity by generating new alleles and new haplotypes respectively. Genetic drift reduces diversity, and results from the random sampling of one generation from the preceding one, causing random change in allele frequencies. Selection can operate in a number of different ways to increase, decrease, or maintain diversity.
- Migration increases population-specific diversity by introducing new alleles.
- The different forces acting on allele frequencies could in principle balance one another out so that, with sufficient time, diversity within a population reaches an equilibrium value. Human populations, however, are not at equilibrium.
- The effective population size ($N_e$) can be very different from the census population size and is affected by past population size fluctuations, variance in reproductive success, and population structure. The effective population size varies between the Y chromosome, X chromosome, mtDNA, and autosomes, because of their different patterns of inheritance.
- The discovery of high levels of natural polymorphism led to the development of the neutral theory, which states that the fate of most mutations in the human genome is determined by genetic drift.
- The concept of a molecular clock is useful for dating splits in gene and species trees but its constant rate assumption is often violated by the finding of different mutation rates in different phylogenetic lineages, which can be explained by a number of factors, known as lineage effects.

## QUESTIONS

**Question 5–1:** A SNP in the tensin gene (*TNS1*) was genotyped in 184 Maasai from Kinyawa in Kenya. The genotype counts were 4 AA, 35 AG, and 145 GG. Test whether or not this SNP is in Hardy–Weinberg equilibrium in this population.

**Question 5–2:** Using the genotype frequency data from the 1000 Genomes browser (browser.1000genomes.org), calculate whether the rs1799990 SNP in the *PRNP* gene is in HWE in the GBR, LWK, and CLM populations.

**Question 5–3:** Using 11,600 as an estimate of human long-term $N_e$ (from Table 6.4), how strong does selection need to be to overcome the effect of genetic drift in changing the frequency of a particular allele on

(a) An autosome

(b) The Y chromosome

Comment on the value of $s$ in both cases.

**Question 5–4:** Briefly compare and contrast these forms of selection: negative, balancing, purifying, and positive.

**Question 5–5:** What are the consequences of polygyny for the relative diversity of the sex chromosomes and mitochondrial DNA?

**Question 5–6:** In humans, why is long-term effective population size so different from real census population size?

**Question 5–7:** Giving examples, explain the consequences of a major reduction in population size followed by a rapid expansion for

(a) Neutral variation

(b) The ability of selection to drive an allele to fixation

Reading Chapter 13 will help with your answer.

**Question 5–8:** Discuss the consequences of kin-structured migration on population genetic parameters, giving examples in humans.

**Question 5–9:** Discuss the evidence for selection at HLA genes.

## REFERENCES

The references highlighted in purple are considered to be important (for this chapter) by the authors.


1. **Anthony DW** (1990) Migration in archeology: the baby and the bathwater. *Am. Anthropol.* **92**, 895–914.
2. **Austerlitz F & Heyer E** (1998) Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proc. Natl Acad. Sci. USA* **95**, 15140–15144.
3. **Bittles A & Black M** (2010) Consanguineous marriage and human evolution. *Annu. Rev. Anthropol.* **39**, 193–207.
4. **Bittles AH & Neel JV** (1994) The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.* **8**, 117–121.
5. **Burton ML, Moore CC, Whiting JWM et al.** (1996) Regions based on social structure. *Curr. Anthropol.* **137**, 87–123.
6. **Caballero A** (1995) On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics* **139**, 1007–1011.
7. **Charlesworth B** (2009) Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205.
8. **Charlesworth B, Morgan M & Charlesworth D** (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.
9. **Davey Smith G, Lawlor DA, Timpson NJ et al.** (2008) Lactase persistence-related genetic variant: population substructure and health outcomes. *Eur. J. Hum. Genet.* **17**, 357–367.
10. **Di Rienzo A, Peterson A, Garza J et al.** (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl Acad. Sci. USA* **91**, 3166–3170.
11. **Felsenstein J** (1971) Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68**, 581–597.
12. **Fix AG** (2004) Kin structured migration: causes and consequences. *Am. J. Hum. Biol.* **16**, 387–394.
13. **Gillespie JH** (1984) The molecular clock may be an episodic clock. *Proc. Natl Acad. Sci. USA* **81**, 8009–8013.
14. **Haldane J** (1947) The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann. Eugen.* **13**, 262–271.
15. **Hardy GH** (1908) Mendelian proportions in a mixed population. *Science* **28**, 49–50.
16. **Hartl DL & Clark AG** (2007) Principles of Population Genetics, 4th ed. Sinauer Associates.
17. **Havlicek J & Roberts SC** (2009) MHC-correlated mate choice in humans: a review. *Psychoneuroendocrinology* **34**, 497–512.
18. **Helgason A, Hrafnkelsson B, Gulcher JR et al.** (2003) A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am. J. Hum. Genet.* **72**, 1370–1388.
19. **Hill KR, Walker RS, Božičević M et al.** (2011) Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science* **331**, 1286–1289.
20. **Holsinger KE & Weir BS** (2009) Genetics in geographically structured populations: defining, estimating and interpreting $F_{ST}$. *Nat. Rev. Genet.* **10**, 639–650.
21. **Horton R, Wilming L, Rand V et al.** (2004) Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899.
22. **Huestis R & Maxwell A** (1932) Does family size run in families? *J. Hered.* **23**, 77–79.
23. **Hughes AL & Nei M** (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170.

24. **Jobling MA, Williams G, Scheibel at al.** (1998) A selective difference between human Y-chromosomal DNA haplotypes. *Curr. Biol.* **8**, 1391–1394.

25. **Kimura M** (1968) Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

26. **Kong A, Frigge ML, Masson G et al.** (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475.

27. **Kumar S** (2005) Molecular clocks: four decades of evolution. *Nat. Rev. Genet.* **6**, 654–662.

28. **Leal SM** (2005) Detection of genotyping errors and pseudo SNPs via deviations from Hardy–Weinberg equilibrium. *Genet. Epidemiol.* **29**, 204–214.

29. **Lindenbaum S** (2008) Understanding kuru: the contribution of anthropology and medicine. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3715–3720.

30. **Martin AP & Palumbi SR** (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90**, 4087–4091.

31. **Matsumura S & Forster P** (2008) Generation time and effective population size in polar Eskimos. *Proc. Biol. Sci.* **275**, 1501–1508.

32. **Mead S, Whitfield J, Poulter M, et al.** (2008) Genetic susceptibility, evolution and the kuru epidemic. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3741–3746.

33. **Miller G** (2000) The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature. Heinemann.

34. **Mills RE, Luttig CT, Larkins CE et al.** (2006) An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res.* **16**, 1182–1190.

35. **Murdock GP** (1967) Ethnographic Atlas. University of Pittsburgh Press.

36. **Nei M, Suzuki Y & Nozawa M** (2010) The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289.

37. **Ohta T** (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286.

38. **Pawlowski B, Dunbar R & Lipowicz A** (2000) Evolutionary fitness. Tall men have more reproductive success. *Nature* **403**, 156.

39. **Pearson K, Lee A & Bramley-Moore L** (1899) Mathematical contributions to the theory of evolution. VI. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philos. Trans. R. Soc. Lond. A* **192**, 257–330.

40. **Pritchard JK & Przeworski M** (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14.

41. **Raymond CK, Kas A, Paddock M et al.** (2005) Ancient haplotypes of the HLA class II region. *Genome Res.* **15**, 1250–1257.

42. **Taylor J, Tyekucheva S, Zody M et al.** (2006) Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol. Biol. Evol.* **23**, 565–573.

43. **Wilkins JF** (2006) Unraveling male and female histories from human genetic data. *Curr. Opin. Genet. Dev.* **16**, 611–617.

44. **Wright S** (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.

45. **Wright S** (1951) The genetical structure of populations. *Ann. Eugen.* **15**, 323–354.

46. **Yin J, Jordan MI & Song YS** (2009) Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics* **25**, i231–i239.

47. **Zuckerkandl E & Pauling L** (1965) Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins (V Bryson & HJ Vogel eds), pp 97–166. Academic Press.