1 Genes, populations and evolution

In this chapter we examine how allele frequencies vary over time and between subpopulations, as a result of drift, mutation and selection. This is relevant to understanding how genetic differentiation between human subpopulations has arisen, and to understanding how allelic association between linked loci varies with demographic history.

1.1 Genetic drift

Although the human population is now very large, during most of human evolution the total human population has been small. Since migration out of Africa began about 100 000 years ago, the human population has been subdivided into endogamous subpopulations or demes. The modern expansion of population dates from the invention of agriculture about 10 000 years ago. Small isolated subpopulations still exist exist in remote regions of the world such as the Arctic, and on small islands. In present-day human populations, the levels of heterozygosity, and the distances over which allelic association can be detected are largely determined by past history when population size was much smaller.

We begin by introducing the concept of genetic drift. In a population of infinite size, allele frequencies will not change from one generation to the next unless there is selection. In a population of finite size, allele frequencies fluctuate at random from one generation to the next, even where there is no selection and each gene copy has an equal chance of being transmitted to the next generation. **Drift** is the change in allele frequencies that results from the accumulation of these random fluctuations over successive generations. The smaller the population, the more pronounced will be the effects of drift. In a finite population undergoing drift, all but one of the allelic variants at a locus will eventually be eliminated, if there is no mutation to generate new alleles and no selection pressure to maintain the polymorphism. This follows from the argument about coalescence that was obtained in the last chapter. Just as all the descent trees of all gene copies coalesce on a single lineage if followed back far enough in a finite population, so if a finite population of gene copies is followed forward in time for long enough, all lineages but one will eventually become extinct. Unless the population is small, however, the number of generations required for one of the allelic variants to be fixed and the others eliminated can be very large.

On average, drift reduces heterozygosity (gene diversity). In the long run this decrease in heterozygosity is balanced by the rate at which new alleles are generated by mutation. The balance between the effects of drift and mutation upon heterozygosity will be determined by the size of the population and the mutation rate of the loci under study (very low for single-nucleotide polymorphisms, higher for microsatellite loci). Another effect of drift, described in Chapter 3, is to generate allelic association between linked loci, as haplotypes drift away from their equilibrium frequencies and some haplotypes are eliminated altogether. In the long run, this increase of allelic association between linked loci is balanced

by the decay of allelic association through recombination. The balance between the effects of drift and recombination upon allelic association will be determined by the size of the population and by the recombination fraction between loci. At very short distances – less than 50 kilobases – the effects of drift dominate.

1.2 Mutations

1.2.1 Mutation rate

The definition of the rate at which spontaneous mutations occur depends upon whether we are examining mutations at the level of single nucleotides (producing single-nucleotide polymorphisms), single codons (some of which will produce variants in protein sequence) or entire genes (as in the case of a Mendelian disorder where mutations at many different possible sites can produce a loss-of-function mutation). Estimates of the mutation rate in humans have been based mainly on observations of Mendelian disorders or classical protein polymorphisms.

To estimate the mutation rate in humans, four main approaches have been used.

1.2.2 Estimation from the epidemiology of Mendelian disorders, assuming equilibrium between mutation and selection

The earliest attempts to estimate mutation rates in humans were made for Mendelian traits that have a dominant mode of inheritance. Even if no information is available on what proportion of cases have an affected parent, it is possible to estimate the mutation rate μ from the prevalence K of the trait at birth in the population and the selection coefficient s, which for a Mendelian dominant disorder is simply the proportion by which fitness is reduced in affected individuals compared with unaffected individuals. If all affected individuals die before reproductive age or are infertile, the selection coefficient is 1. In a population of N individuals, the rate at which new mutant alleles enter the population is $2N\mu$ per generation, and the rate at which they are lost is 2N.Ks (if the mutant alleles do not undergo mutations back to the wild-type allele).

At equilibrium, $K = \frac{2\mu}{s}$

On the same principle, the equilibrium frequency of alleles for a Mendelian recessive disorder, assuming random mating and no difference between the fitness of heterozygotes as

$$\sqrt{\mu/S}$$

where

On this basis Haldane (1932) calculated the mEstimates of mutation rates based on the prevalence of recessive traits at birth are unreliable: the frequency of recessive traits is increased by inbreeding, and the fitness of heterozygotes is not usually known.

1.2.3 Estimating the prevalence at birth of Mendelian dominant disorders in births to unaffected parents

For example, in a study in Denmark, 10 heterozygotes for achondroplastic dwarfism were found in a survey of 94000 births. Eight of these cases were offspring of unaffected parents, and were assumed to be new mutations. From this we can estimate the rate of mutations that disrupt this gene as about 4 x 10^{-5} per generation.

Unless mutations in the base sequence or protein structure are typed directly, errors in estimates of the mutation rate are likely to arise from phenocopies (individuals whose phenotype resembles that produced by the mutation but who do not have a genetic defect) and from genetic heterogeneity (mutations at several different loci). These errors tend to inflate estimates of the mutation rate.

1.2.4 $Mutationrateatthenucleotidelevel$

Recent estimates of mutation rate for single nucleotides are estimated to be about 2×10^{-8} per generation() (Neel, Satoh, Goriki, Fujita, Takahashi, Asakawa, & Hazama 1986).

Drake JW 1998 Nachman MW 1998

On this basis, each human zygote (with 3 x 10^9 base-pairs) will contain about 100 new mutations

1.2.5 Mutation rate at the level of single codons or entire genes

A nucleotide substitution within a codon that does not change the amino acid in the peptide is known as a **synonymous substitution** or silent mutation. The mutation rate at the codon level is the probability that a codon chosen at random undergoes a non-synonymous mutation: one that results in the replacement of the amino acid in the peptide. The relationship between the mutation rate μ_n at the nucleotide level and the mutation rate μ_c at the codon level can be calculated as follows. If we neglect the very small probability that more than one nucleotide substitution will occur, the probability of a nucleotide substitution in a codon is approximately $3\mu_n$. The probability that a nucleotide substitution in a codon will be nonsynonymous - that it will result in an amino acid replacement - is about $^3/_3$. The mutation rate at the codon level (μ_c) is therefore approximately $\frac{3}{4} \times 3\mu_n = 2.25\mu_n$

It is estimated that electrophoresis detects about one-third of all amino acid variants.

Estimating the mutation rate from the rate of substitution

1.2.6 $Mutationratesestimatedfromproteinelectrophoresis$

The development of electrophoretic methods to detect variant peptide sequences made it possible to estimate mutation rates at the level of gene, codons and nucleotide. An early estimate by this method rate was made by Kimura and

Ohta (1965). In a total of 320 000 individuals whose haemoglobin was examined by electrophoresis, 62 individuals were found to be heterozygous for variants of the haemoglobin α or β chain. Of 18 individuals heterozygous for variants of the α or β chains whose parents were studied, two were found to have new mutations. We can thus estimate that for the α and β chains combined, the detectable mutations occur at a rate of

$\frac{2}{18} \times \frac{62}{320\,000 \times 2} = 10^{-5}$ per generation

As only about one quarter to one third of variants in amino acid sequence are detectable by electrophoresis, we can estimate that the combined mutation rate for the α and β chains (considered as a single locus) is about 3 x 10^{-5} per generation. As these chains consist of 141 and 146 amino acids, the mutation rate at the codon level can be estimated as about 10^{-7} per generation. If $\mu_c$is about 10^{-7} per generation, the mutation rate at the nucleotide level is therefore about 4 x 10^{-8} per generation.

In a study in Hiroshima (excluding those heavily exposed to radiation from the atomic bomb), Neel et al examined parent-offspring trios and typed 36 polypeptides by electrophoresis(.) (Neel et al. 1986). Three new mutations were detected among 540 000 transmitted genes. This yielded an estimated detectable mutation rate at the gene level of 6 x 10^{-6} per generation. On the assumption that electrophoresis detects about one-third of variants in amino acid sequence, the rate for mutations causing amino acid substitutions in polypeptides was estimated to be about 1.2 x 10^{-5} per generation. Allowing for synonymous mutations, they estimated the mutation rate at the nucleotide level to be about 10^{-8} per generation.

1.2.7 *Relationofmutationratetoparentalageandsex*

In females, germ cell division is complete by the time of birth and meiosis occurs only when an oocyte matures. The number of divisions from zygote to egg is estimated to be about 24. In males, germ cells divide continuously and many cell divisions occur before a spermatocyte is produced. It is estimated that there about 200 cell divisions before producing a spermatocyte at age 20, and 770 cell divisions before producing a spermatocyte at age 45.

Watson (1965) estimated that the rate of mutation per cell division is constant. If the rate of mutation per cell division is constant, we can predict that the mutation rate *per generation* will be higher in men than in women, and will be dependent on paternal age because spermatocytes from older men will have accumulated more mutations. Crow() (1997) has reviewed the evidence for these two predictions. If two unaffected parents produce an offspring affected b determine the parent of origin of autosomal mutations. In new cases of a disorder such as achondroplasia that is inherited as a Mendelian dominant with complete penetrance, the mutation can be presumed to have arisen in the germ line of one of the parents. By typing markers adjacent to the locus of the mutation it is possible to determine which parent the mutation was inherited from. In studies of achondroplasia and Apert syndrome (acrocephalosyndactyly), the mutations were found to be predominantly of paternal origin. The risk of these mutations

increases with paternal age: in cases of Apert syndrome, the average paternal age was 6.1 years older than the average for the population, whereas there was no effect of maternal age or birth order when paternal age was included in the mode(.) (Erickson & Cohen, Jr. 1974).

Estimates of the mutation rate in males can also be derived from estimates of the rate of gene substitution on the Y chromosome. The mutation rate per generation in humans and higher primates is estimated to be 3 to 6 times higher in males than in females.{Li, Ellsworth, et al. 1996 4803 /id}. Further support for this "generation-time" hypothesis comes from comparisons with mice and rats, in which the male-to female ratio for mutation rate per generation is only about 2, similar to the corresponding sex ratio in numbers of germ cell divisions per generation.{Li, Ellsworth, et al. 1996 4803 /id}

In a systematic review of available data on the relationship of risk of mutations to parental age, about two-thirds of Mendelian dominant disorders, including achondroplasia, Apert syndrome and Marfan syndrome, were found to show a strong paternal age effect.{Risch, Reich, et al. 1987 4801 /id}. For the remaining one-third, including multiple exostoses and neurofibromatosis, the relationship to paternal age was weak or nonexistent. It has been suggested that the disorders which show a strong relationship to paternal age are those which result from base substitutions, which accumulate during germ cell proliferation. Disorders which are not related to paternal age may result from deletions or gene duplications, which may occur at meiosis rather than accumulating during germ cell proliferation. For instance most cases of achondroplasia or Apert syndrome result from single-nucleotide substitutions at specific sites. In contrast, most cases of neurofibromatosis result from deletions (mostly of maternal origin) rather than base substitutions (mostly of paternal origin).

Crow points out that one way to reduce the accumulation of deleterious mutations would be for all males to freeze sperm samples at puberty, and to use these samples for procreative purposes.

1.3 The Wright-Fisher model

The evolution of allele frequencies through drift and mutation can be predicted by a model developed by Fisher (1930) and Wright (1931). This model is based on an **idealized population** of size N, in which each new generation of N individuals is formed by drawing a random sample (with replacement) of two gene copies N times from the 2N gene copies in the previous generation. "Sampling with replacement" means that it is possible for an individual to inherit two copies of the same gene copy. This sampling process will generate random variation in allele frequencies between successive generations. The smaller the population, the larger will be this random variation of allele frequencies across generations. We can thus use the variation of allele frequencies to define a measurement of population size.

1.3.1 Effective population size

We define the **effective population size** N_e as the size of an idealized population that has the same variance of allele frequencies between successive generations as that in the population under study. This is distinguished from the **census population size,** which is simply the population size as determined by counting individuals.

As humans (unlike plants) are not self-fertilizing, it is impossible for an individual to inherit two copies of the same gene copy as in Wright's model of an idealized population. However if the population consists of equal numbers of males and females mating at random, it is easy to show that the effective population size is the same as if each new generation were formed by sampling two gene copies N times from the 2N gene copies in the previous generation (Wright 1931).

In an idealized population of size N, where the initial allele frequencies at a biallelic locus under study are p_0 and q_0, the variance of the allele frequencies p_1 and q_1 in the next generation is given by the standard formula for the variance of a binomial proportion based on 2N observations

$$V(p_1) = V(q_1) = \frac{p_0 q_0}{2N}$$

In a population of ½N men and ½N women mating at random, the 2N gene copies transmitted to the next generation of N individuals consist of two samples of size N, one from men and one from women. The variance of the allele frequencies p_1 and q_1 in the total sample of 2N gene copies transmitted to the next generation is

$\mathrm{Var}[\frac{1}{2}(p_m + p_f)] = \frac{1}{4}[\mathrm{Var}(p_m) + \mathrm{Var}(p_f)] = \frac{1}{4}\left(\frac{p_0 q_0}{N} + \frac{p_0 q_0}{N}\right) = \frac{p_0 q_0}{2N}$

which is the same as the variance of p_1 and q_1 in an idealized population of size N

Strictly this definition, the **variance effective size**, is only one of three possible ways of defining the effective population size. Two other properties of finite populations can be used to define an effective population size. One is that in a finite population mating at random, there is a non-zero probability that the two gene copies transmitted to an individual are identical by descent even when people avoid obviously consanguineous mating. This is because in a finite population, all individuals will share common ancestors if their lineages are followed back far enough. Thus we can define the **inbreeding effective size** as the size of an idealized population in which the probability that two gene copies sampled at random (with replacement) are identical by descent is the same as the probability in the population under study that the two gene copies transmitted to an individual are identical by descent. In an idealized population of 2N gene copies, the probability that two gene copies are identical by descent is 1/2N. Half the reciprocal of this probability is N.

Alternatively, we can define the **extinction effective size** as the size of an idealized population in which the rate at which alleles are lost from the population as a result of random drift is equal to that in the population under

study. Except in some extreme situations, where there is random mating within the subpopulation both these alternative definitions of effective population size are equivalent to the variance effective size, defined above.

The effective population size of human populations is generally only about one-quarter of the total population size that would be recorded in a complete census of all living individuals. This is because fertility varies between individuals, and because generations overlap: only those individuals who are of reproductive age produce offspring. Several simple formulae have been derived that allow the effective population size to be calculated from the census size - the total number of individuals in the population.

1.3.2 *Unequalnumbersofmalesandfemales*

If the population consists of N_m males and N_f females, the effective population size is

$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

If N_m and N_f are equal, the effective population size is simply the total number of males and females as shown above.

1.3.3 *Varyingnumbersofprogeny*

If G is a random variable representing the number of progeny from one of the N parents in the population, and the population size is stable (implying that E[G] = 2), the effective population size is approximately

$$N_e = \frac{4N}{V\,(G) + 2}$$

If the number of progeny follows a Poisson distribution, V(G) = E (G) =2 and the effective population size is approximately equal to N, the number of parents in the population. In most human populations, the variance of progeny number is larger than the mean, and the effective population size is proportionately reduced.

If variation in fertility is inherited, and the heritability (defined later) is h^2, the effective population size is reduced further, and the equation above becomes

$$N_e = \frac{4N}{(1 + 3h^2)\,V\,(G) + 2}$$

It is estimated that in humans the combination of heritable and nonheritable variation in fertility reduces population size to approximately half the value that it would have if the number of progeny was a random variable with a Poisson distribution (Nei and Murata).

1.3.4 *Fluctuatingpopulationsize*

Where population size is fluctuating, the effective population size N_e is given by the **harmonic mean** (the reciprocal of the mean of reciprocals) of the population sizes in each generation, If N_i is the population size in the i^{th} of n generations

$$\frac{1}{N_e} = \frac{1}{n}\sum_i \frac{1}{N_i}$$

The same principle applies to the effective population size where drift is occurring independently in two or more populations of unequal size. For calculating the variance of allele frequencies between populations, the effective population size is the harmonic mean of the population sizes in each independently drifting population. As small values of N_i contribute disproportionately to the harmonic mean, the effective population size of most modern human subpopulations that have expanded in the last 10 000 years is largely set by their demographic history during the earlier hunter-gatherer phase, when population sizes were smaller.

1.3.5 *Overlappinggenerations*

If N_a is the number of individuals born per year who survive to reproductive age and τ is the mean age at reproduction, then

$N_e = \tau N_a$

In a typical human population where the mean age at reproduction is about 25 years and the number of individuals born each year who survive to reproductive age is about one-sixtieth of the total population, the effect of overlapping generations is to reduce the value of N_e to about 40 percent of the total population size (Nei and Imaizumi 1966).

1.4 Equilibrium between mutation and drift in a finite population

In a finite population, new alleles at each locus are continuously generated by mutation and lost from the population by drift. The balance between these two effects will determine the frequency with which neutral polymorphisms occur. If the mutation rate and the effective population size are constant, the heterozygosity, averaged over a large number of loci, will reach an equilibrium. This applies only if we are choosing a sample of loci that have been identified by methods that do not depend on the heterozygosity of the locus; for instance if each locus is defined by the base sequence coding for a peptide. If we choose a sample of restriction site polymorphisms or single-nucleotide polymorphisms that have been selected for their heterozygosity, obviously these loci will be more heterozygous than their equilibrium state.

If the mutation rate per generation at the locus is μ, the expected number of new mutations at this locus in a population of effective size N is $2N\mu$ in each generation. If there is no selection, the probability that an allelic variant will be fixed (given that one of the allelic variants at a locus must eventually be fixed)

is simply the frequency of that allele. Thus if a ncew mutation is neutral to selection, the probability that this mutation will eventually become fixed in the population is 1/2N, the initial frequency of the mutant allele. At equilibrium, in a population of effective size N, the rate at which new alleles are created by mutation balances the rate at which they are lost by drift. The average rate at which alleles are completely replaced in the population will then be

$$2N\mu\frac{1}{2N} = \mu$$

This is the rate of allele substitution - the rate at which alleles at the locus are completely replaced as a result of mutation and drift. Thus the rate of allele substitution is equal to the mutation rate, and independent of the effective population size. This result is useful in estimating the time that has elapsed since two subpopulations became separated, as explained later.

1.4.1 *Expectedhomozygosityundertheinfiniteallelesmodelofmutation*

Kimura and Crow (1964) derived a simple relationship to calculate the expected homozygosity (or heterozygosity) at equilibrium between mutation and drift from the mutation rate at the locus and the effective population size. This relationship is based on the assumption that every mutation produces a new allele that does not already exist in the population. This **infinite alleles model** may not be too unrealistic for sites that consist of single nucleotides, where the mutation rate is very low and only a small proportion of sites show any variation: where there is no pre-existing variation, every mutation will produce a new allele. The infinite alleles model may also be realistic when the locus under study is an entire gene, and alleles are typed by electrophoresis or immunological methods, as the number of possible peptide sequences is large. The infinite alleles model is not a realistic model for microsatellite loci where the mutation rate is high, alleles are detected by the number of tandem repeats they contain, and mutations that increase or reduce the number of repeats are likely to produce alleles that already exist in the population.

As in other analyses of drift, the derivation relies on the concept of effective population size. If the effective population size is N, we can consider the two alleles at a locus in a randomly-chosen individual as obtained by sampling without replacement from a population of 2N alleles.

Let μ be the mutation rate per generation, and J_t the expected homozygosity in generation t - the probability that an individual chosen at random is homozygous at the locus under study. This is simply the probability that two alleles chosen at random from generation t are identical by state.

A homozygous individual can be produced in two different ways: either (i) if two copies of the same allele are transmitted, and neither allele has mutated; or (ii) if two different alleles are transmitted, these two alleles are identical by state, and neither allele has mutated.

The probability that two copies of the same allele are transmitted, and nei-

ther allele has mutated is

$$\frac{1}{2N}\left(1-\mu\right)^2$$

The probability that two different alleles are transmitted to an individual in generation t, these two alleles are identical by state, and neither allele has mutated is

$$\left(1-\frac{1}{2N}\right)J_{t-1}\left(1-\mu\right)^2$$

We therefore have $J_t = \frac{1}{2N}\left(1-\mu\right)^2 + \left(1-\frac{1}{2N}\right)J_{t-1}\left(1-\mu\right)^2$

At equilibrium, $J_t = J_{t-1} = J$. The homozygosity at equilibrium (J) is given by

$$J = \frac{\left(1-\mu\right)^2 \frac{1}{2N}}{1-\left(1-\mu\right)^2\left(1-\frac{1}{2N}\right)} = \frac{\left(1-\mu\right)^2}{2N-\left(1-\mu\right)^2\left(2N-1\right)} \approx \frac{1}{1+4N\mu}$$

To express this another way, at equilibrium the effective number of alleles at the locus (1/J) is equal to 1 + 4Nμ.

This can alternatively be derived by an argument based on coalescence. In the previous generation, a pair of gene copies may coalesce with probability 1/2N, fail to coalesce or mutate with probability 2μug(for the event that either one mutates). The probability of identity is therefore the probability that the gene copies ultimately coalesce (rather than mutate). This probability is

$$\frac{1/2N}{1/2N+2\mu} = \frac{1}{1+4N\mu}$$

1.4.2 *Expectedhomozygositywherethenumberofpossibleallelesisfinite*

Where the locus is a single nucleotide, there are only four possible alleles at the locus and the infinite alleles model does not strictly apply. There is still a simple relationship between the mutation rate, the effective population size and the expected homozygosity at equilibrium. If there are k possible alleles at a locus, and mutations to all allelic states are equally probable, so that with total mutation rate μ per generation the probability that an allele will mutate to one of the remaining (k - 1) states is μ/(k - 1).

The homozygosity at equilibrium (J) is then $\frac{1+\frac{4\mathrm{N}_m}{\mathrm{k}-1}}{1+4\mathrm{N}m+\frac{4\mathrm{N}_m}{\mathrm{k}-1}}$

Where 4Nμ is small and k=4, as for single nucleotides in human populations, the value of this expression is very close to the simpler expression obtained under the infinite alleles model.

These expressions for expected homozygosity apply only to loci that have been identified by methods that do not depend upon the level of heterozygosity. For instance, we could not use these expressions to predict heterozygosity at SNP loci that have been identified by mining sequence data from haploid genomes to identify sites at which a single nucleotide differs sites in sequence data on

overlapping clones from haploid genomes. With this procedure, loci that have high heterozygosity have a higher chance of being detected and catalogued as SNPs.

We can however use these expressions for expected heterozygosity to predict the heterozygosity at the nucleotide level: the proportion of single nucleotides that will differ in two haploid genomes randomly chosen from unrelated individuals from the population, assuming that most of these variant sites are neutral to selection. More usefully, we can estimate the effective population size from the observed heterozygosity at the nucleotide level

1.4.3 *Estimation of effective population size from observed heterozygosity in human populations*

[Li and Sadler 1991, Przeworski 2000]

For single nucleotides, the mutation rate μ is estimated to be about 2×10^{-8} per generation. From data obtained during efforts to sequence the human genome using overlapping clones, the heterozygosity at the nucleotide level is estimated to be 1 in 1300 nucleotides, or about 0.00077. The homozygosity is therefore 0.99923.

Substituting J = 0.99923 and $\mu = 2 \times 10^{-8}$ into the equation for expected homozygosity under the infinite alleles model, we can estimate the effective size of the human population as about 10 000 individuals. This means that the level of heterozygosity in the human population as a whole is about what would be expected in a population equivalent In other words

1.4.4 *Proportion of loci that have a given level of heterozygosity*

We can also predict the proportion of loci that are polymorphic. If we define a polymorphic locus as one where the frequency of the commonest allele is less than or equal to 1 - q, the expected proportion of polymorphic loci is simply

1 - $q^{4N\mu}$

and the expected proportion of loci that are polymorphic (frequency of commonest allele less than of equal to 1 - q) is given by

$$1 - k\frac{\Gamma(A+B)}{\Gamma(A)\,\Gamma(B)}\left(\frac{1-q^{A}}{A} - \frac{1-q^{A+B-1}}{A+B-1}\right)$$

where A = 4Nμ/(k - 1), B = 4Nμ. and Γ(x) is the gamma function

If the number of possible alleles k is large, these reduce to the simpler expressions derived above for the infinite alleles model. Strictly this model does not hold for single nucleotides, as not all nucleotide substitutions are equally probable.

The expected proportion of loci at which we would expect to find a single nucleotide polymorphism for which the frequency of the commonest allele is less than or equal to 0.80 is

1 - $0.20^{0.004}$ = 0.006

[Calculate using correct formula for four alleles]

For comparison, recent experimental studies indicate that SNPs that have this level of heterozygosity occur about once every 1000 base-pairs in the human genome.

1.5 Loss of heterozygosity under drift in a subpopulation compared with the ancestral total population: the fixation index

From the effective population size we can predict the effect of drift upon heterozygosity, and model the genetic differentiation between human subpopulations that has been produced by drift. For instance if we study stable polymorphisms such as single-nucleotide polymorphisms where the mutation rate is low (only about 10^{-8} per transmission) and examine the effects of demographic change over relatively short periods (less than 10^4 generations) we can ignore the effects of mutation and fit a model based entirely on drift.

To exaimine how drift leads to variation in allele frequencies between endogamous subpopulations that have split off from an ancestral total population, we consider a subpopulation of effective size N, formed by drawing a sample (with replacement) of 2N gene copies from a total population with allele frequencies p_0, q_0 at a biallelic locus. We write p_1, q_1 for the corresponding allele frequencies in the first generation of the subpopulation.

The variance of allele frequencies in the first generation of the subpopulation is given by

$$E\left(p_1^2\right) - \left[E\left(p_1\right)\right]^2 = \frac{p_0 q_0}{2N}$$

Thus $E\left(p_1^2\right) = p_0^2 + \frac{p_0 q_0}{2N}$ since $E(p_1) = p_0$

The expected frequency of heterozygotes in the next generation is

$E(2p_1q_1)$

$= 2E(p_1) - 2E(p_1^2) = 2p_0 - 2\left(p_0^2 - \frac{p_0 q_0}{2N}\right)$

$$= 2p_0 q_0 \left(1 - \frac{1}{2N}\right)$$

and the expected frequency of heterozygotes after t generations is therefore

$$2p_0 q_0 \left(1 - \frac{1}{2N}\right)^t$$

Thus in a subpopulation of effective size N, derived from a total population in which the allele frequencies at a diallelic locus are initially $p_0$and q_0, the expected heterozygosity after t generations of drift is
$\left(1 - \frac{1}{2N}\right)^t$ times the value of $2p_0q_0$ in the ancestral total population.

At $t = \infty$, one of the two alleles will have become fixed and the expected overall frequency of heterozygotes in the subpopulation is zero.

The same argument can easily be extended to loci with more than two alleles. The proportion by which heterozygosity is reduced by drift in an endogamous

subpopulation compared with the ancestral total population from which the subpopulation separated, is called the **fixation index** (Wright 1931) and denoted by F_{ST}, where the subscript ST stands for "subpopulation-total". Under this definition,

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

where H_T is the heterozygosity in the ancestral total population from which the subpopulation was derived, and H_S is the heterozygosity in the subpopulation. The F_{ST} value for a subpopulation relative to a total population lies between 0 and 1.

[Earlier chapter should define gene diversity and gene identity (equal to heterozygosity and homozygosity for diploid organisms with random mating)].

The expected value of the fixation index in a subpopulation relative to the ancestral total population depends only on the number of generations over which the subpopulation has been drifting and the effective size N of the subpopulation.

$$E\left(F_{ST}\right) = \left(1 - \frac{1}{2N}\right)^t$$

This relationship of F_{ST} to the effective subpopulation size N and the number t of generations since separation from the ancestral total population applies only if F_{ST} is calculated from stable polymorphisms such as single-nucleotide polymorphisms, where the mutation rate is so low that the effect of mutations is negligible in comparison with the effect of drift on allele frequencies.

1.6 Fixation index as a measure of genetic differentiation between two subpopulations

As defined above, the fixation index F_{ST} measures the proportion by which drift has reduced the heterozygosity of a subpopulation, compared with the heterozygosity in the ancestral total population from which that subpopulation was derived. This is not exactly the same as the definition originally given by Wright, but it is the one that most easily leads to a more general definition of F_{ST} as a measure of genetic differentiation between two or more subpopulations.

We can extend this definition to two or more subpopulations by defining F_{ST} as the average proportion by which drift has reduced the heterozygosity in these subpopulations, compared with the heterozygosity in the ancestral total population. We can then use F_{ST} to measure the genetic differentiation between these subpopulations that has resulted from drift. To calculate F_{ST} directly for two or more subpopulations, we require the allele frequencies in each subpopulation (to calculate H_S as the average heterozygosity in subpopulations) and in the ancestral total population (to calculate H_T as the heterozygosity in the ancestral total population). possible, because the subdivision of a total population into subpopulations occurred long ago, and the ancestral total population is no longer available for study. When studying genetic differentiation between human subpopulations, typically we are not comparing a subpopulation with a total population, but instead studying two or more subpopulations that have

been derived from an ancestral total population that no longer exists. To estimate F_{ST} as defined above, we have then to estimate the heterozygosity H_T in the ancestral total population from the allele frequencies in the subpopulations. This leads to an alternative definition of F_{ST}. For simplicity, this definition is set out below for a locus with two alleles and for two subpopulations labelled X and Y.

We define the **gene identity within subpopulations** ($\mathbf{J}_S$) as the probability that two gene copies, each chosen at random from the same subpopulation (also chosen at random) , are identical by state. This is the average homozygosity (gene identity) of the two subpopulations. J_S is equal to 1- H_S, where H_S is the average heterozygosity of the two subpopulations.

We define the **gene identity between subpopulations** ($\mathbf{J}_T$) as the probability that two gene copies, each chosen at random from one of the two subpopulations (denoted X and Y), are identical by state. This is the expected homozygosity in individuals who have one parent from each of the two subpopulations. We can easily show that J_T is an estimate of the homozygosity in the ancestral total population from which the subpopulations separated. Suppose that the allele frequencies in this ancestral total population were p_0, q_0 in that divides at random into two subpopulations. The homozygosity of this ancestral total population - the probability that two gene copies chosen at random from the total population are identical by state - is $(p_0^2 + q_0^2)$. As the two subpopulations drift at random, their average heterozygosity declines but the probability J_T that two gene copies, each chosen at random from one of the two subpopulations, are identical by state is still $(p_0^2 + q_0^2)$, because the probabilities of each allele when one gene copy is chosen at random from each subpopulation are still p_0, q_0.

In the definition of F_{ST} given above, we can then replace H_T by $(1 - J_T)$, and H_S by
$(1 - J_S)$ to obtain

$$F_{ST} = \frac{J_S - J_T}{1 - J_T}$$

This definition of F_{ST} in terms of gene identity probabilities depends only upon the allele frequencies in the subpopulations.

For two subpopulations X and Y that have allele frequencies p_X, q_X and p_Y,q_Y the gene identity between subpopulations is given by

$J_T = p_X p_Y + q_X q_Y$

and the gene identity within subpopulations is given by

$J_S = 1 - \frac{1}{2}(2p_X q_X + 2p_Y q_Y) = 1 - p_X q_X - p_Y q_Y$

Substituting these expressions into the definition of F_{ST} yields, after some algebra

$F_{ST} = \frac{(p_X - p_Y)^2}{2(\bar{p} - p_X p_Y)}$ where $\bar{p} = \frac{1}{2}(p_X + p_Y)$

This expression for F_{ST} in terms of allele frequencies in subpopulations can be generalized to loci with more than two alleles, and to more than two subpopulations [Reynolds, also Pons and Chaouche 1995]. In practice, the allele frequencies p_X and p_Y in the two subpopulations X and Y are not known ex-

actly, and have to be estimated by genotyping samples of individuals from each subpopulation. To obtain unbiased estimates of the gene identity probabilities, it is necessary to correct for sampling error.

Where n individuals have been sampled from a homogeneous subpopulation and a biallelic locus has been typed, an unbiased estimate of the gene identity within this population can be calculated as

$$1 - \frac{n}{n-1} 2\hat{p}\hat{q}$$

where $\hat{p}$and $\hat{q}$ are the estimated allele frequencies. [check this please].

[Pons and Chaouche 1995]

To obtain reliable estimates of F_{ST} for a pair of subpopulations, the proportionate reduction in heterozygosity is estimated from allele frequencies at many loci. To estimate F_{ST} distances, , the most appropriate loci are stable polymorphisms that are likely to be neutral to selection, such as single nucleotide polymorphisms in noncoding regions of DNA are most suitable. Estimates of F_{ST} from classical polymorphisms, where the locus is an entire gene and alleles are typed by protein electrophoresis, are similar to estimates based on restriction site polymorphisms or SNPs.

Empirical evidence suggests that the proportion of markers that have extreme frequency differentials between human ethnic groups is larger than would be expected under a purely neutral model (.) (Bowcock et al. 1991).

1.7 Fixation index as a measure of genetic distance between subpopulations

From the definition of F_{ST} in terms of gene identity probabilities, we can derive a useful measure of **genetic distance** – the extent of genetic differentiation - between human subpopulations. Confusingly, the term genetic distance is sometimes used as a synonym for the map distance between two linked loci. We shall use it only for measures of genetic differentiation between subpopulations.

The relationship derived above shows that in subpopulations of effective size N after t generations of drift, the expected fixation index is given by

$$F_{ST} = 1 - \left(1 - \frac{1}{2N}\right)^t$$

If the effective population size N is large, we can approximate this expression for F_{ST} by an exponential function, using the equation

$$e^x = \lim_{a \to 0} (1 + a)^{\frac{x}{a}}$$

Substituting a = -1/2N and x = -t/2N into this equation gives

$$F_{ST} = 1 - \left(1 - \frac{1}{2N}\right)^t \approx 1 - e^{-t/2N}$$

This is an example of a diffusion approximation, in which the discrete changes in allele frequencies are approximated by a continuous process.

To obtain a fmeasure of genetic distance in terms of F_{ST} it is convenient to rearrange the equation $F_{ST} \approx 1 - e^{-t/2N}$ to give

$$\frac{t}{2N} \approx -\log_e (1 - F_{ST})$$

The expression $\log_e (1 - F_{ST})$ is known as the $\mathbf{F}_{ST}$ **distance**. This can take values between 0 (when $F_{ST} = 0$) and infinity (when $F_{ST} = 1$ as with two inbred strains of mice). Where F_{ST} is small, as it is between all human subpopulations, the F_{ST} distance $\log_e (1 - F_{ST})$ is close to the fixation index F_{ST}. If there has been no mutation and no selection pressure at the loci under study, the F_{ST} distance between two subpopulations should depend only on N, the effective size of the subpopulations (as the harmonic mean over both subpopulations and all generations) and t, the number of generations since separation.

If the "out-of-Africa" hypothesis is correct, we can model the history of human subpopulations as having been formed by separation from an ancestral total population in Africa. For instance, the first migration of anatomically modern humans out of Africa some 100 000 years ago must have subdivided the total human population into at least two endogamous subpopulations. Settlement of other regions within and outside Africa has led to further subdivision. F_{ST} distances between human subpopulations have been estimated by Cavalli-Sforza, using phenotypic markers such as serum proteins. The estimates are summarized in the table below:-

[Add table of F_{ST} distances]

The F_{ST} distances between non-African and sub-Saharan African populations are generally in the range 0.15 to 0.2. It is estimated that Africans and non-Africans separated about 100 000 years ago [Cavalli-Sforza], equivalent to about 5000 generations. From the expression for the F_{ST} distance above, we have $0.2 = 5000/2N$, where N is the population effective size (harmonic mean over both subpopulations and all generations). From this we can estimate N as about 12000. As we shall see later, the effective size of African populations appears to be much larger than the effective size of European and other non-African populations.

For a given value of the F_{ST} distance between two subpopulations, the distribution of F_{ST} values over a large number of loci can be tested for departure from the distribution expected under the neutral theory. This has been used to test for evidence of disruptive selection pressure on polymorphisms since subpopulations became separated.

1.8 Modelling drift using coalescent theory

An alternative way to model the evolution of genes is to use coalescent theory. Under neutrality, it is assumed that each lineage chooses one ancestor at random from the population in the previous generation.

For this example N is the size of the haploid population

T(j) is the time for the size of the genealogy to be reduced from j to (j - 1), measured in units of N generations.

In the diffusion limit, T(j) is exponentially distributed with parameter $\binom{j}{2}$

and $E(T_{MRCA}) = 2$ if the sample size is large

Thus if there are N copies of an allele in the population, the expected time back to the most recent common ancestor of these copies, assuming neutrality and no new mutations, is 2N generations.

Implications: where the allele is rare and the effective size of the total population is small (as for rare disease-causing mutations in Finland), the time back to the most recent common ancestor will generally be short . For a common allele, unless there has been extreme constraint of population size, the most recent common ancestor will be a long way back in the past. For example, in a population whose effective size (harmonic mean over the time back to coalescence is 10 000 individuals, the effective number of copies of an allele that has frequency 20% is 4000. The expected time back to the most recent common ancestor, assuming neutrality and no new mutations since the allele first arose, is therefore 2 × 4000 generations (about 160 000 years). This suggests that most single-nucleotide polymorphisms are very old, antedating the split between African and non-African subpopulations some 100 000 years ago. This is consistent with the empirical observation that most common single-nucleotide polymorphisms present in non-African populations are present in African populations also.

1.8.1 *Measures of genetic distance that depend upon the mutation rate: Nei's standard genetic distance*

A widely-used measure of genetic distance was defined by Nei (1975) as the standard genetic distance D.

This is defined by $D = -[\log_e J_{XY} - \frac{1}{2}(\log_e J_X + J_Y)]$

where J_{XY} is the gene identity between subpopulations X and Y, J_X is the homozygosity in subpopulations X, and J_Y is the homozygosity in subpopulation Y.

If p_{iX} and p_{iY} are the frequencies of alleles of type i in subpopulations X and Y,

$J_{XY} = \sum p_{iX} p_{iY}$, $\quad J_X = \sum p_{iX}^2$ and $J_Y = \sum p_{iY}^2$

D lies between 0 and infinity.

Nei's D was developed for use with classical protein polymorphisms, where each locus under study is an entire gene, and the alleles are variants in peptide sequence, detected by electrophoresis or immunological reactions. In this situation the terms $\log_e J_{XY}$, $\log_e J_X$ and $\log_e J_Y$ are related to the probability that at a single codon within the locus, the amino acid specified by the codon differs in two alleles (peptide sequences) chosen at random..

If c_X is the probability (assumed same for all codons) that the codon differs in two alleles chosen at random from population X, and the allele is made up of n codons, the homozygosity J_X at the locus is given by

$J_X = (1 - c_X)^n \quad \approx \exp(-nc_x)$ since c_x is small

It follows that $\log_e J_X \approx -nc_X$.

The expectation of this expression is the mean number of codons that differ in two alleles chosen at random. We can define c_Y and c_{XY} similarly, so that

$D \approx n [c_{XY} - \frac{1}{2}(c_X + c_Y)]$

D is thus an estimate of what Nei calls the **net codon differences per locus** between two subpopulations: the expected number of codons that differ in a pair of alleles chosen at random from subpopulation X and subpopulation Y, minus the average of the expected number of codons that differ in a pair of alleles chosen at random within each subpopulation. Note that the net codon differences per locus is proportional to n, the number of codons that make up the allele. The more amino acids in the peptide, the higher will be the value of D.

Under certain assumptions – that an infinite alleles model holds, that all alleles neutral to selection, and that both subpopulations X and Y are in equilibrium between neutral mutation and drift, D is proportional to the number of generations since the two subpopulations became separated.

Suppose that two subpopulations X and Y are formed from a total population. We write $J_{XY}(t)$ for the gene identity between subpopulations X and Y at t generations after separation, and μ for the mutation rate at the codon level per generation.

The probability that two alleles made up of n codons are identical by state in generation t+1, given that they are identical by state in generation t, is $(1 - \mu)^{2n}$ if the infinite alleles model applies

Thus $J_{XY}(t + 1) = J_{XY}(t).(1 - \mu)^{2n}$

If subpopulations X and Y are in equilibrium between mutation and drift, and all alleles are neutral to selection, $J_{XY}(0) = J_X = J_Y$

It follows that $J_{XY}(t) = (1 - \mu)^{2nt}$

$$D = -\log_e \frac{J_{XY}}{\sqrt{J_X J_Y}} = -\log_e \left[(1-\mu)^{2nt}\right] \approx 2n\mu t$$

If an estimate of the mutation rate μ at the codon level is available, and D has been estimated by typing a large number of classical protein loci, we can use this relationship to estimate t, the number of generations since two human subpopulations became separated.

For instance Nei and Roychoudhury (1974) estimated from typing 62 protein loci that the standard genetic distance D between Europeans and west Africans was 0.023. The rate at which electrophoretically detectable mutations occur per protein locus is estimated to be about 10^{-7} per year, or 2.5 x 10^{-6} per generation if the average generation time is 25 years. As this is the mutation rate per protein locus, rather than per codon, we can substitute it for $n\mu$.

This gives $t = \frac{D}{2n\mu} = \frac{0.023}{2 \times 2.5 \times 10^{-6}} = 4600$

If average generation time is 25 years, 4600 generations is equivalent to 115 000 years. This is consistent with other estimates of the time since the split between African and non-African subpopulations.

Although Nei's D statistic was developed for use with classical protein polymorphisms, it can be calculated for any type of locus where the infinite alleles

model applies, the alleles are neutral and there is an equilibrium between mutation and drift. Where the locus consists of an entire gene, these conditions are likely to apply because the gene consists of a large number of independent subunits (codons or nucleotides) in which mutations occur independently and a mutation in any subunit is enough to generate a new allele. Values of D calculated from stable polymorphisms such as SNPs cannot be interpreted in terms of the mutation rate or the number of generations since separation.

The value of D depends upon the rate at which detectable mutations occur. Thus much higher values of D would be obtained if allelic variants were distinguished by sequencing entire genes to detect all variants in base sequence, rather than by electrophoresis of proteins, which will detect only non-synonymous mutations in codons (and only about one-third of these).

1.9 Other measures of genetic distance between subpopulations that depend on drift

1.9.1 $Edwards's d coefficient$

Another measure of genetic distance that is closely related to F_{ST} is the d coefficient defined by Edwards (1971). This is based on a geometric model. Suppose that, for a locus with two alleles, we plot the allele frequencies p_X, q_X in subpopulation X and p_Y, q_Y in subpopulation Y on a graph with axes scaled as $\sqrt{p}$ and $\sqrt{q}$.

As p + q = 1, the two points that define allele frequencies in each subpopulation lie on the circumference of a quarter circle with radius 1. Edwards' d is defined as

$$d = \sqrt{1 - \cos\theta}$$

where θ is the angle between the radii connecting the two points to the origin. If the differences in allele frequencies between the two subpopulations X and Y are not large, then

$2d^2 = F_{ST}$

1.9.2 $Wahlund variance$

The fixation index F_{ST} measures the average proportion by which heterozygosity in subpopulations has been reduced since they separated from the total population. An alternative measure of genetic differentiation between subpopulations, which does not depend upon any assumptions about an ancestral total population, is simply to estimate the average proportion by which heterozygosity in these subpopulations is lower than in a total population formed by pooling these subpopulations into a single population. Where the total population is made up of endogamous subpopulations that have different allele frequencies, heterozygosity is lower than if there is random mating between all members of the total population. This proportionate reduction in heterozygosity that results from partition of a total population into two subpopulations is called the **standardized variance of allele frequencies** or the **Wahlund variance**,

after the geneticist who defined it in 1928. We shall use the symbol f for the Wahlund variance, to distinguish from the fixation index F_{ST}.

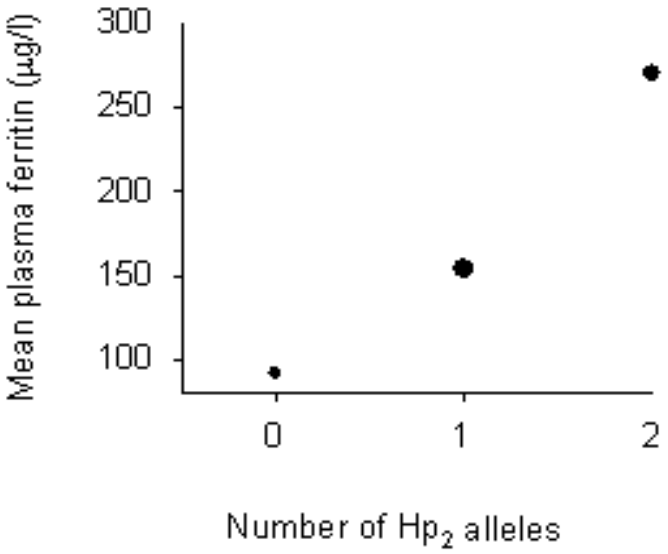
Suppose that the allele frequencies at a biallelic locus are p_X, q_X in subpopulation X and p_Y, q_Y in subpopulation Y. If equal numbers of individuals from the two subpopulations were combined to form a total population, the heterozygosity H_t of this total population would be given by

$H_t = 2\bar{p}\,\bar{q}$ where $\bar{p} = \frac{1}{2}(p_1 + p_2)$ *and* $\bar{q} = 1 - \bar{p}$.

Although each subpopulation is in Hardy-Weinberg equilibrium, the total population will not be in Hardy-Weinberg equilibrium. Instead, because there are two endogamous subpopulations with different allele frequencies, the mean heterozygosity (H_S) of the two subpopulations is given by

$H_S = {}^1\!/_2(2p_X q_X + 2p_Y q_Y)$

This expression is always less than or equal to $2\bar{p}\,\bar{q}$. This can be seen by plotting heterozygosity H - defined as the function 2p(1 - p) - against allele frequency p. This gives a curve that is concave downwards. Because the curve is concave downwards the y-coordinate of the mid-point of a straight line drawn between two points on the curve (the average heterozygosity $H_S = {}^1\!/_2(p_X q_X + p_Y q_Y)$ of two subpopulations with allele frequencies p_X and p_Y) must be less than the y-coordinate of the point on the curve that has the same x-coordinate as this mid-point (which is the heterozygosity of a pooled population with allele frequency ${}^1\!/_2(p_X + p_Y)$. This is an example of a mathematical principle known as the Jensen inequality, which states that for any function that is concave downwards, the mean of the function is less than or equal to the function of the mean; this general result is known in mathematics as the Jensen inequality



The reduction in heterozygosity that results from partition of the total population into two equally-sized subpopulations that have different allele frequencies is then

H_S – H_T

$= 2\bar{p}\,\bar{q} - (p_1 q_1 + p_2 q_2) = 2 \times \frac{1}{2}(p_1 + p_2) . \frac{1}{2}(q_1 + q_2) - (p_1 q_1 + p_2 q_2) = \frac{1}{2}(-p_1 q_1 + p_1 q_2 + p_2 q_1 - p_2 q_2)$

$= \frac{1}{2}[p_1 (p_1 - p_2) - p_2 (p_1 - p_2)] = \frac{1}{2}(p_1 - p_2)^2$

The reduction in heterozygosity H_S - H_t can then be expressed as a proportion of the heterozygosity H_t of the total population.

$$f = \frac{H_t - H_s}{H_t} = \frac{\frac{1}{2}(p_1 - p_2)^2}{2\bar{p}\,\bar{q}} = \frac{\frac{1}{4}(p_1 - p_2)^2}{\bar{p}\,\bar{q}}$$

Because the average heterozygosity of subpopulations is always lower than the heterozygosity of a total population formed by pooling these subpopulations, f always lies between 0 and 1. This definition of f as the proportionate reduction in heterozygosity of subpopulations compared with the total population formed by pooling these subpopulations (H_t - H_S)/H_t can be generalized to loci with more than two alleles, and to more than two subpopulations. Nei refers to this proportionate reduction in heterozygosity as the **coefficient of gene differentiation**, and uses the symbol G_{ST}.

The terms "standardized variance" and Wahlund variance refer to an alternative definition of f in terms of the variance of allele frequencies. If p_X and p_Y are the values of two observations sampled from an underlying distribution,

$\frac{1}{4}(p_1 - p_2)^2$is the sample variance $\frac{1}{2}\sum_{i=1}^{2}(p_i - \bar{p})^2$

Note that this is the sample variance - the sum of squared deviations from the sample mean divided by the number of observations- rather than an estimate of the variance of an underlying probability distribution. To calculate an unbiased estimate of the variance of an underlying probability distribution from a sample of n observations, we would divide the sum of squared deviations from the sample mean by (n - 1).

If we assign numeric values of 0 and 1 to the two alleles, the variance of the allelic value in the pooled population where the allele frequencies are $\bar{p}$ and $\bar{q}$ is simply the binomial variance $\bar{p}\,\bar{q}$. The Wahlund variance f is therefore the ratio of the variance of allele frequencies in the sample of two subpopulations to the variance $\bar{p}\,\bar{q}$of the mean allelic value $\bar{p}$ in the pooled population.

1.9.3 *RelationofWahlundvariancetofixationindex*

The distinction between Wahlund variance and fixation index is not always clearly made; for instance Cavalli-Sforza defines the two terms to be synonymous. This is set out here to minimize confusion.

For two or more subpopulations that have mean heterozygosity H_S :-

the Wahlund variance (f) is $\frac{H_t - H_S}{H_t}$

the fixation index (F_{ST}) is $\frac{H_T - H_S}{H_T}$

where H_t is the heterozygosity that would exist in a total population formed by pooling equal numbers of individuals from each subpopulation, and H_T is the heterozygosity in the ancestral total population from which the subpopulations are derived. H_t can be calculated directly from the allele frequencies in the subpopulations under study, whereas H_T (usually) has to be estimated from the gene identity J_{XY} between subpopulations, also a function of the allele frequencies in the subpopulations under study.

For s subpopulations f and F_{ST} are related by the equation $f = \frac{\left(1-\frac{1}{s}\right)F_{ST}}{1-\frac{F_{ST}}{s}}$

If the number s of subpopulations is infinite, this equation reduces to $f = F_{ST}$. In other words, when the number of subpopulations is infinite the heterozygosity H_t of a total population formed by pooling these subpopulations is the same as the heterozygosity H_T of the ancestral total population.

If there are only two subpopulations, this equation reduces to $f = F_{ST}/(2 - F_{ST})$. If F_{ST} is small (<0.25) as it is between all human subpopulations, this expression reduces to $f \approx {}^1\!/_2 F_{ST}$.

As noted earlier, more complicated formulae are required to estimate both f and F_{ST} to allow for sampling error in measuring the allele frequencies in subpopulations.

Either f or F_{ST} can be used to summarize variation in allele frequencies at a single locus, or the average variation over a large number of loci. One use of f values, as outlined later, is as a measure of the information about ancestry (which subpopulation) that is conveyed by typing an allele at a marker locus.

f is less useful as a measure of the average genetic distance between two or more subpopulations because it depends upon the number s of subpopulations studied, whereas F_{ST} depends only on the effective population size N and the number t of generations since separation.

1.9.4 *Approximating drift by a diffusion process*

For an initial allele frequency of p_0 in a population of size N, the probability distribution of the allele frequency after t generations of random drift can be derived by approximating the fluctuations in allele frequencies by a diffusion process (Kimura).

Equation for probability that an allele with initial frequency p_0 will become fixed.

1.10 Allelic association and drift

Under a neutral model, approximate expressions can be obtained for the expected value of the squared allelic correlation coefficient r.

Where the mutation rate is small compared with the recombination fraction θ, the expectation of r^2 (estimated from a large sample) is approximately

$$E\left(r^2\right) = \frac{1}{1 + 4N_e\theta}$$

We showed earlier that the effective population size N_e can be estimated from observed heterozygosity at the nucleotide level as about 10^4. If the ratio of map distance to physical distance is approximately 1 cM per 1 Mb, we would expect to observe strong allelic association ($r^2 > 0.5$) commonly at physical distances up to 2.5 kb, and to observe useful allelic association ($r^2 > 0.1$) commonly at distances up to 20 kb.

Pritchard and Przeworski review studies that have examined the relation of allelic association to physical distance, and compare the empirical findings to the theoretical predictions. They note that some studies have reported allelic associations at distances greater than 1 cM, which is not predicted by the population genetic model. On the other hand, at short distances (<10 kb), as when multiple SNPs are typed within a single gene, the allelic associations are weaker than predicted by the model.

They suggest several possible explanations:-

1. $The relationship of recombination rate to physical distance varies across the genome$
2. $Several studies are based on the X chromosome, where higher levels of allelic association are expected because the effective population$
3. $Haplotype frequencies may be estimated inaccurately, especially where populations show pronounced departure from H-W equilibr$
4. $Predictions of allelic association based on a model of constant population size may be underestimates: but Pritchard and Przeworski a$
5. $Allelic association between microsatellite loci may be stronger than between SNP loci$
6. $Some of the populations studied have been formed by recent admixture, or are inbred.$
7. ^{4} $Some of the populations studied may have undergone a pronounced bottleneck in the past, so that the estimate of Ne=10 may be far too hi$
8. $Selection may have generated allelic associations over relatively long distances where a rare favourable mutation has been rapidly swe$
9. $Gene conversion may break up allelic associations between closely-linked markers$
10. $Inversion polymorphisms may allow strong allelic associations to develop over long distances$

1.11 Time since coalescence

(i) infinite sites model For a sample of n alleles drawn from a population of alleles of constant size 2N, the expected number t of generations since coalescence of the descent tree of these n alleles is given by

$$E(t) = 4N\left(1 - \frac{1}{n}\right)$$

For instance, in an isolated population of constant size 5000 individuals, there will be 3000 copies of an allelic variant that has frequency 0.3. If we assume that the allele frequency has been constant, the copies of this allelic variant can be considered as a population of constant size 3000. For all copies of the allele that exist in the present-day population, the expected number of generations back to coalescence is 12000: about 250 000 years. This is the "age" of the allele

– note that this coalescence time is not necessarily the same as the time back to a common ancestor.

As discussed later, there is evidence that the effective size of the population ancestral to modern humans may have been only about 10 000 individuals(.)(Zietkiewicz et al. 1998), and that populations that were ancestral to non-African populations had even smaller effective size.

Where the allele frequency is lower, and the population has been small for many generations, the expected time back to coalescence is much shorter. For instance, in an isolated population of constant effective size 1000 individuals, the expected time back to coalescence of an allele with frequency 1% (20 copies) is only 80 generations: about 2000 years.

The same argument applies when we consider all alleles of a given type as the population under study, assuming that no new copies of this variant arise by mutation of other allelic variants.

The time since coalescence is likely to be shorter for alleles at microsatellite loci, which have a high mutation rate and are highly polymorphic so that no single allele is common, than at SNP loci which have a low mutation rate and where both alleles may be common. , as at microsatellite loci. Both these factors will tend to shorten the time since coalescence for a randomly-chosen allele at one or other loci.

Association of a disease-causing mutation with haplotypes, or alleles at a highly polymorphic locus may be detectable over relatively long distances even if not all the copies of the mutation coalesce on a recent common ancestor. For instance, where a mutation has arisen more than once in the history of the population, it may be that some copies of the mutation existing in the present-day population coalesce on a recent common ancestor, while the other copies coalesce much further back in time. In this situation, there will not be a single ancestral haplotype, but the distribution of haplotypes between chromosomes bearing the disease-causing mutation and chromosomes not bearing the mutation will be non-random. than one point.

(ii) infinite alleles model Hudson ()(Hudson 1985) undertook extensive simulations to examine the sampling distribution of allelic association under the infinite alleles model. He concluded that if $4N\theta$ was less than 10, strong allelic association would commonly occur, and that allelic association might be detectable if $4N\theta$ was as large as 50.

Thus in a population where (until recently) the effective size has been about 5000 individuals, we might expect to see allelic association commonly when θ $<10/(4 \text{ x } 5000)$, or when $x < 0.05$ cM, equivalent to about 50 kilobases. For strong allelic association to occur frequently at $x = 1$ cM, the effective population size must be only about 250 individuals.

1.11.1 *Averageageofaneutralmutation*

The statistical theory of drift can be used to estimate the average age of neutral mutations from the allele frequency and effective population size. The average

time required for the mutant allele to reach for the first time a frequency p from its initial frequency of 1/2N is the average first arrival time. The average age of the allele will be greater than the average first arrival time, as it is possible that for the allele frequency to arrive at a value of p more than once. If we can neglect the possibility that the allele that is less common has previously been fixed and has subsequently declined as a result of new mutations, there is a simple formula (Kimura and Ohta 1973) for the average age t of a neutral allele that has reached an allele frequency of p in a population of effective size N

$$t = -4N \frac{p}{1-p} \log_e p$$

where t is the number of generations since the mutation arose.

The average age of the mutation is thus independent of the mutation rate, and of the same order as the effective population size N. If the effective population size through most of human evolution has been about 10^5 individuals, the average age of a neutral allele that has frequency 0.2 is between 20 000 generations - equivalent to 4 x 10^5 years. This is greater than the estimated 2 x 10^5 years since coalescence of mitochondrial ancestry for all humans now alive. As the mutation rate at the nucleotide level is very low, most neutral SNPs are very old. As shown earlier, the average age of alleles that have frequency of more than 0.20 can be estimated to be of the same order as the age of our species, and far greater than the 5000 generations since the deepest split between human subpopulations: that between Africans and non-Africans.

1.12 Fitness and selection

1.12.1 *Tests of selective neutrality*

The infinite alleles model can be used to construct tests of selective neutrality. Selection will generally lead to lower heterozygosity than expected under the neutral theory. In general the commonest allele will be more common than expected, and the less common alleles will be rarer than expected, because selection acts to eliminate rare and usually deleterious alleles as they enter the population by mutation.

Suppose that we type the locus under study in a sample of n chromosomes from the population, and identify k allelic variants in this sample. We can summarize the distribution of these distinct alleles in a table as an **allelic partition**. This table has two rows. In each column, the cell in the first row contains a value i, and the cell in the second row contains the number of alleles that are present i times in the sample.

For example, suppose we type a sample of 150 chromosomes from the population and identify eight allelic variants. Three of these alleles occur only once, one occurs twice, two occur 5 times, another 20 times and the commonest allele occurs 115 times

The allelic partition is

						Total
i	1	2	5	20	115	
a_i	3	1	2	1	1	8

Note that Σa_i = k and $\Sigma i.a_i$ = n. For an alleles that occurs i times, the allele frequency is i/n. The heterozygosity is therefore

$$1 - \sum_i a_i \left(\frac{i}{n}\right)^2 = 1 - \frac{1}{n^2}\sum_i a_i i^2$$

Given k distinct alleles (k $\geq$ 2) observed in a sample of size n (n > k), we can imagine two extreme situations:-

1. an lopsided allelic partition, such that one distinct allele type occurs [n - (k - 1)] times, and the remaining (k-1) distinct allele types occur once only. The average value of a_i is then $\frac{1}{2}[1 + (k-1)] = \frac{1}{2}k$, the maximum possible value . For example, if eight allelic variants are present in a sample of 150 chromosomes, the lowest possible heterozygosity would be when one allele occurs 143 times and the other seven alleles occur once only. The average value of a_i is then $\frac{1}{2}(1 + 7)$

(ii) a more even allelic partition, such that the k $\times$ 1 table of allele frequencies in the sample contains k unique values. Thus $a_i = 1$ for all i, and the average value of a_i is 1.

This suggests that we can use the average value of a_i as the basis of a test for whether the commonest alleles are more common than expected, and the rarest alleles are rarer than expected under a model for equilibrium between mutation and drift. If selection pressure has acted to eliminate rare alleles, the mean value of a_i will be higher than expected. This is the basis of the **Ewens-Watterson test** of selective neutrality.

Ewens's sampling formula – quote it in full?

Ewens (1972) showed that the expected value of a_i under the infinite alleles model, given that k distinct alleles have been observed in a sample of n chromosomes, is given by:-

E(a_i | k distinct alleles in sample size n)= $\frac{n!\left|S_{n-1}^{k-1}\right|}{i(n-i\left|S_n^k\right|)}$

where $\left|S_n^k\right|$ is a Stirling number of the first kind: the coefficient of θ^k in the expansion of $\theta(\theta + 1)(\theta + 2)$. . .$(\theta$ + n -1) as a polynomial, where $\theta = 4N\mu$.

To test for departure from the infinite alleles model of drift, given that k alleles have been observed in a sample of n gametes, the sum of the squared allele frequencies in the sample (the homozygosity) can be used as a test statistic. In a large sample, this is equal to the homozygosity in the population. The sampling distribution of this test statistic under the null hypothesis of selective neutrality can be generated by simulation, given the sample size and the number of distinct alleles. An algorithm to generate this distribution, based on the Ewens sampling distribution, was described by Stewart (1977).

As this test depends upon the infinite alleles model, it can be used to test for selective neutrality only where alleles are classified at the level of the entire gene and the infinite alleles model is therefore a reasonable approximation.

Example of a test of selective neutrality: Ros Harding's paper on the MC1R gene

1.12.2 $Effectofselectionpressurewhenthereisheterozygoteadvantage$

For some polymorphisms, allele frequencies have been determined by the point at which the advantages of the heterozygous genotype are balanced by the disadvantages of the homozygous genotype. The best-understood examples are haemoglobinopathies where the heterozygote has increased resistance to malaria.

Fitness of an individual is defined as the expected number of that individual's offspring that survive to reproduce in the next generation

Effect of selection pressure on allele frequencies for a biallelic polymorphism				
	A_1A_1	A_1A_2	A_2A_2	Total
Fitness	w_1	w_2	w_3	
Frequency before selection	p^2	2pq	q^2	1
Frequency after selection	w_1p^2	$2w_2pq$	w_3q^2	$w_1p^2 + 2w_2pq + w_3q^2$

Frequency $p^/$ of allele A_1 after selection is
$w_1p^2 + {}^1\!/_2.2w_2pq$
$= p(w_1p + w_2q)$
At equilibrium $\frac{p'}{q'} = \frac{p(w_1p+w_2q)}{q(w_2p+w_3q)} = \frac{p}{q}$

$$\frac{p_e}{q_e} = \frac{w_2 - w_3}{w_2 - w_1} = \frac{1 - \frac{w_3}{w_2}}{1 - \frac{w_1}{w_2}}$$

Thus at equilibrium, the ratio of the frequencies of allele A_1 and A_2 is the ratio of the proportionate reduction in fitness for genotype A_2A_2 compared with the heterozygote to the proportionate reduction in fitness for genotype A_1A_1 compared with the heterozygote.

We can use this to estimate what reduction in fitness has selected for allele frequencies of 10% for beta-thalassaemia in some eastern Mediterranean populations. As beta-thalassaemia homozygotes cannot survive to reproductive age without blood transfusions, it can be assumed that over the period in which selection has been operating the fitness of beta-thalassaemia homozygotes has been zero.

We have $p_e = 0.1$, $q_e = 0.9$, and $w_1/w_2 = 0$

Substitution into the equation above gives $w_3/w_2 = 0.89$

Thus we can estimate that if the reduction in fitness for AA homozygotes compared with beta-thalassaemia heterozygotes was entirely attributable to increased risk of dying from malaria before reproductive age, this excess risk was about 11% in eastern Mediterranean populations.

For the sickle cell trait (Hb S) allele, allele frequencies in some west African populations are as high as 15%. The fitness of SS homozygotes is higher than zero however, as some survive to reproductive age. If the fitness of SS homozygotes is assumed to be 30%, we have $p_e = 0.15$, $q_e = 0.85$, and $w_1/w_2 = 0.3$ Substitution into the equation above gives $w_3/w_2 =$. This implies that the risk of death from falciparum malaria before reproductive age in AA homozygotes in some areas of west Africa may have been be as high as XX%. This is consistent with what is known of the magnitude and causes of childhood mortality in west Africa: randomized trials of measures to prevent exposure to mosquitoes have shown that childhood mortality can be reduced by up to one-third (check rates).

Time taken to reach equilibrium is about 100 generations. This is compatible with estimates that the spread of malaria in Africa was associated with the development of agriculture within the last 3000 years.

Effect of change in selection pressure

1.12.3 *Geneticload*

Genetic load is defined as the proportion by which average fitness of a population is decreased in relation to the fitness that the population would have if all individuals had the genotype that has maximum fitness (Crow 1958).

Comparison with the population attributable risk fraction in epidemiology.

Mutational and segregational load Mutations that have an observable effect are almost always deleterious. Experimental studies in Drosophila suggest that most mutations which have observable effects cause only a slight reduction in fitness, and that the average fly carries about one new mutation causing slight reduction in fitness{Crow 1997 4798 /id}.

Estimation of the genetic load from mortality data

Number of lethal equivalents

Number of genes contributing to severe mental handicap

1.13 Exercises

1.13.1 *Calculationofeffectivepopulationsize*

Suppose that we are able to observe the fluctuation of allele frequencies between successive generations at a locus where the two alleles have frequencies p and q, and estimate the variance V of allele frequencies at this locus. In an idealized population of size N, we could calculate this variance as the variance of the mean of 2N observations from a binomial distribution, given by pq/2N.

N_e can therefore be defined as $\frac{pq}{2V}$

1.13.2 *Effectivepopulationsizewhenthenumbersofmalesandfemalesareunequal*

This formula can be derived by using the definition of population effective size as the inbreeding effective number.

The inbreeding effective number is defined as half the reciprocal of the probability that two alleles chosen at random from the population for any locus are identical by descent.

We consider a population of N_m males and N_f females in which population size is constant and mating is random. For two alleles chosen at random from the population in generation t, the probability that both alleles came from males in generation t -2 is $1/4$. The conditional probability that they are identical by descent, given that both alleles came from males in generation t - 2, is $1/2N_m$. The probability that these two alleles are identical by descent and came from a male in generation t is therefore $1/8N_m$. Similarly, the probability that these two alleles are identical by descent and came from a female in generation t -2 is $1/2N_f$.

The inbreeding effective number N_e is therefore half the reciprocal of $\frac{1}{8N_m} + \frac{1}{8N_f}$

This yields $N_e = \frac{4N_m N_f}{N_m + N_f}$

1.13.3 *RelationbetweenWahlundvarianceandfixationindex*

Suppose that there are s subpopulations derived from an ancestral total population. We write p_i, q_i for the allele frequencies in the ith subpopulation at a biallelic locus. Let J_0 be the mean homozygosity of the subpopulations: the probability that two alleles chosen at random from a randomly-chosen subpopulation are identical by state. Let J_1 be the homozygosity of the total population: the probability that when two subpopulations are chosen at random and one allele is chosen at random from each subpopulation, these two alleles are identical by state.

We have $J_0 = \frac{1}{s}\sum\left(p_i^2 + q_i^2\right)$ and $J_1 = \frac{1}{s^2}\sum_j\sum_j\left(p_i p_j + q_i q_j\right)$

We can derive a recurrence relation for $J_0^{(t)}$ as

$$J_0^{(t+1)} = \frac{1}{2N} + \left(1 - \frac{1}{2N} J_0^{(t)}\right)$$

We have $D_{ST} = \left(1 - \frac{1}{s}\right)(J_0 - J_1)$ and $H_T = 1 - J_0 + D_{ST} = 1 - J_0 - \frac{J_0 - J_1}{s}$

Therefore $f = \frac{\left(1-\frac{1}{s}\right)(J_0 - J_1)}{1 - J_0 - \frac{J_0 - J_1}{s}}$

If equal numbers of individuals from the two subpopulations were combined to form a total population, the heterozygosity H_t of this total population would be given by

$H_t = 2\bar{p}\,\bar{q}$ where $\bar{p} = \frac{1}{2}(p_1 + p_2)$ *and* $\bar{q} = 1 - \bar{p}$.

Instead, because there are two endogamous subpopulations with different allele frequencies, the mean heterozygosity (H_s) of the two subpopulations is given by

$H_s = {}^1\!/_2(2p_1q_1 + 2p_2q_2)$

The reduction in heterozygosity that results from partition of the total population into two subpopulations that have different allele frequencies is given by

H_s - H_t

$= 2\bar{p}\,\bar{q} - (p_1q_1 + p_2q_2) = 2 \times \frac{1}{2}(p_1 + p_2) . \frac{1}{2}(q_1 + q_2) - (p_1q_1 + p_2q_2) = \frac{1}{2}(-p_1q_1 + p_1q_2 + p_2q_1 - p_2q_2)$

$= \frac{1}{2}[p_1(p_1 - p_2) - p_2(p_1 - p_2)] = \frac{1}{2}(p_1 - p_2)^2$

$$f = \frac{H_t - H_s}{H_t} = \frac{\frac{1}{2}(p_1 - p_2)^2}{2\bar{p}\,\bar{q}} = \frac{\frac{1}{4}(p_1 - p_2)^2}{\bar{p}\,\bar{q}}$$

1.13.4 *Parent of origin for X−linked mutations*

Haldane (1947), showed that almost all males affected with classical X-linked haemophilia had inherited the mutation from a mother who was a heterozygous carrier of the mutation. He compared this to the theoretical prediction that if the mutation rate were the same in both sexes, two-thirds of affected males would have inherited the mutation from a carrier mother (rather than from a mutation in a germ cell inherited from the mother).

Subsequent work has supported Haldane's conclusion that most mutations for classical haemophilia are of paternal origin.

Almost half of the mutations causing classical haemophilia are caused by X chromosome inversions that for some reason occur in males far more commonly than in females.

Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K., & Cavalli-Sforza, L. L. 1991, "Drift, admixture, and selection in human evolution: a study with DNA polymorphisms", *Proceedings of the National Academy of Sciences of the USA*, vol. 88,no. 3, pp. 839-843.

Crow, J. F. 1997, "The high spontaneous mutation rate: is it a health risk?", *Proc.Natl.Acad.Sci.U.S.A*, vol. 94, pp. 8380-8386.

Erickson, J. D. & Cohen, M. M., Jr. 1974, "A study of parental age effects on the occurrence of fresh mutations for the Apert syndrome", *Annals of Human Genetics*, vol. 38, pp. 89-96.

Neel, J. V., Satoh, C., Goriki, K., Fujita, M., Takahashi, N., Asakawa, J., & Hazama, R. 1986, "The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides", *Proc.Natl.Acad.Sci.U.S.A*, vol. 83, pp. 389-393.