Lecture 7 Week 8 6th March

Population structure

Gil McVean

The evolutionary significance of population structure

Population structure is one of the most studied and least understood aspects of population genetics. Broadly speaking, structure refers to any deviation from random-mating, and includes phenomena such as inbreeding, associative mating (where reproduction is stratified among genotypes), and geographical subdivision. Geographical structure has received the most attention for two reasons. First, geographical structure is an inescapable fact of biology. Populations may be separated by oceans, mountains or deserts. Even when there are no barriers to gene flow, organisms do not disperse randomly across the species range – rather, they tend to remain close to where they were born. Under these circumstances, genetic and phenotypic differences can accumulate between populations. The second reason is that differentiation between local populations must represent the early stages of speciation. It is a fundamental aim of evolutionary biology to understand how and why partially isolated populations diverge at both the genetic and phenotypic levels, and when this can lead to reproductive isolation and ultimately speciation. Because geography is the most important scale of population structure, it will be the major focus of this lecture.

Geographical structure

Geographical structure is the non-random mating of individuals with respect to location. Among species, it is probably ubiquitous. In humans it is obvious – someone from England is more likely to mate with someone else from England than they are with someone from China. Even within England, you are more likely to pair up with someone who lives close to you, and because we tend to live close to where we grew up, this naturally leads to geographical stratification. Historically, the scale of migration was much less than in contemporary Britain and the legacy of such behaviour is the geographical structure we see in things like surnames and dialects. This slide shows the density of the Scottish surname Hannah in contemporary Britain (excluding N. Ireland). This was kindly lent by Sara Goodacre who is working on large-scale project to map names across Britain with Brian Sykes at the IMM. Not surprisingly, there is a strong excess of the name in Scotland, though it clearly shows some migration down the Pennines and throughout England.

Why is this pattern important? Because surnames are transmitted through the paternal line, and many, particularly Scottish names, have probably arisen only a few times in history, the non-random location of surnames is indicative of non-random distribution of Y-chromosomes. Consequently, any genetic variability on the Y chromosome will also show geographical structure. Genetic variability at other loci may show less structure (for example if there is greater female dispersal), but in general we do not expect genetic variation to be evenly distributed across Britain. This is important for two reasons. First, local mating within locally structured populations may have different population dynamics from the classical Fisher-Wright model. Second, geographical structure of genetic variability may imply geographical structure of phenotypically important traits. If we are

prepared to call surname a phenotype, then the Y-chromosome represents a locus of major effect. More important geographical structure in phenotype is seen in things like the incidence of red hair, skin pigmentation and certain diseases (e.g. Tay Sachs disease in Ashkenazi Jews). Consequently we would expect geographical structure in genetic variability at genes influencing such traits.

The non-random localisation of the surname Hannah is largely due to chance – chance that Hannah became a surname in Scotland, chance that the first Hannahs lived near Glasgow, Chance that they migrated they way they have. There is nothing about possessing the surname Hannah that causes such people to be better off in one place rather than another (excepting perhaps the possibility of group selection – Hannahs help each other). This is not generally true for phenotypic variation, and the geographical structure of traits such as hair and skin colour, not just within Britain, but across Europe and the rest of the World, is probably indicative of geographical variation in selection pressures (perhaps UV irradiation). So genetic differentiation can come about by chance and by selection. As we will see later in this lecture, genetic differentiation at loci influencing traits of selective importance can also come about through the interaction of chance and selection. Distinguishing between these possibilities is an exciting challenge for population genetics.

Detecting and describing genetic structure

In previous lectures I have talked about ways of describing patterns of genetic variability, and using such patterns to infer things or test hypotheses about the underlying evolutionary processes. How does incorporating population structure affect this procedure?

The first tool we need is a way of describing structure within genetic data. The most commonly used methods of summarising structure within genetic variability are the *F* statistics developed by Sewall Wright (1951). *F* statistics partition genetic variability as measured by levels of heterozygosity into components of within population and between population variation. For example, suppose you have collected data on genetic variability within your favourite species, from samples spread across the country. Although the population may actually be continuous across the country, it is natural to divide your sample into different populations, and to ask how much variation there is within each level of structure relative to other levels. The most cited statistic is the proportion of total heterozygosity (H_T) that is explained by within population heterozygosity (H_S).

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

Where the line over H_S indicates that it is the average heterozygosity within populations. Other *F* statistics may measure the proportion of heterozygosity within populations that is explained by within individual heterozygosity (F_{IS}: a measure of inbreeding) or the proportion of variation explained by successively higher levels of population classification (e.g. sample site < region < country < continent).

F statistics describe the partitioning of variability within the sampled data. In themselves they do not tell us whether there is any significant structure within the data. Significance levels are best estimated by permutation. The null distribution of the statistic of interest (e.g. F_{ST}) under the hypothesis of no significant structure is obtained empirically by randomising alleles or genotypes with

respect to location. If the observed level of structure is greater than expected by chance, there is evidence for genetic differentiation.

Before looking at some estimates of F_{ST} from natural populations, it is worth mentioning a couple of things about *F* statistics. First, because it is a ratio, the statistic contains no information about absolute levels of genetic variability. In many ways this is good because we want to know about differentiation relative to other processes (e.g. inbreeding, mutation rate), but it also throws away much information, and is liable to have high sampling variance when levels of heterozygosity are low. Second, some *F* statistics can actually be negative. For example, suppose there is a tendency for individuals to actively avoid breeding with relatives. Levels of heterozygosity within individuals will therefore tend to be higher than levels of heterozygosity in the local population, and the statistic F_{IS} will be negative.

F_{ST} in natural populations

In the early days of molecular population genetics, calculating *F* statistics from patterns of allozyme variation was a growth industry. Naturally, the greatest interest was in the differentiation of human populations, and studies of the major races of humans (Caucasians, Africans, Chinese) put F_{ST} in the region of 0.07. In other words, 93% of all allozyme variation is within populations and only 7% is between. Remarkably, similar levels of differentiation can be observed at much finer scales. For example, about 8% of the variation among Yanomama American Indians is between villages and 92% is within (though the total level of heterozygosity among the villages is much less than the worldwide level).

Is this a lot or a little differentiation? The answer is really only meaningful in relation to other species. Human commensals, such as house mice and *Drosophila melanogaster* show similar levels of differentiation (perhaps not surprisingly), though *D. melanogaster* is less differentiated. Certainly humans are on the low end of the spectrum for levels of differentiation. Some organisms, for example the Jumping Rodent have an F_{ST} of over 0.5, suggesting strong racial differentiation, and maybe even the presence of reproductively isolated sub-species.

More recently, in the era of DNA sequencing studies, the habit of calculating F_{ST} has gone out of fashion, but it is of interest to compare the results of allozyme and nucleotide studies. Using the data from recent surveys of nucleotide diversity from SNPs in humans (Goddard *et al.* 2000) and *D. melanogaster* (Andolfatto, unpublished) the levels of differentiation for DNA sequences seem very similar to those from allozymes.

F statistics can be used to describe genetic differentiation between any groups of organisms, whether they are spatially separated or not. For example, a study of the tapeworm *Ascaris* in Guatemala found strong differentiation between samples from humans and samples from pigs kept in the same villages (Anderson *et al.* 1993). Host preference, or low migration rates between the two populations might explain why populations differentiate even when in sympatry (without geographic separation).

The inbreeding effect of population structure

F statistics provide a way of summarising information on geographical structure to genetic variability, but what is it they are actually measuring? If we just consider a single locus, genetic differentiation between populations means nothing more than differences in allele frequency between populations (with the extreme of different alleles being fixed in different populations). Suppose we have just two populations in which just two alleles are segregating, but at different frequencies (p_1 and p_2 respectively). If each population is in Hardy-Weinberg equilibrium, the expected homozygosity in each population is given by

$$E[F_i] = p_i^2 + q_i^2$$

Where q = 1- p. However, suppose we did not know that we were actually sampling from different populations. In this case, the expected frequency of homozygotes is

$$E[F_T] = \bar{p}^2 + \bar{q}^2 \text{ , where } p = (p_1 + p_2)/2$$

With a bit of algebraic rearrangement, it follows that the observed frequency of homozygotes in the combined populations is inflated relative to that expected by the variance in allele frequency over populations

$$F_T = E[F_T] + 2\sigma_p^2$$

Consequently, a naive analysis that did not account for population structure would find an excess of homozygotes – exactly the same result as would occur if individuals within a single population have a tend to mate with relatives (inbreed). Deviations from Hardy-Weinberg equilibrium in the direction of an excess of homozyogotes may be indicative of unaccounted for levels of local population structure.

What is the relationship between the inbreeding effect of structure and population differentiation as measured by *F* statistcs? From the relationship $H = 1 - F$ (heterozygosity = 1 – homozygosity) it follows that we can write F_{ST} in terms of the inflation of levels of homozygosity

$$F_{ST} = \frac{\bar{F}_S - F_T}{1 - F_T} = \frac{\sigma_p^2}{\bar{p}\bar{q}}$$

In other words, the degree of population differentiation as measured by Wright's F_{ST} statistic is directly proportional to the variance in allele frequency over populations. This relationship generalises in the case of multiple alleles at many loci

$$F_{ST} = \frac{\sum_{i,j} \sigma_i^2}{1 - \sum_{i,j} \bar{x}_i^2}$$

Where the summation is over alleles *i* at loci *j*.

The Wahlund effect

Population structure creates effective inbreeding, because local fluctuations in allele frequency tend to inflate the frequency of homozyogotes. The opposite side of the coin is that if two differentiated populations are brought into contact and allowed to mate, the frequency of heterozygotes will increase relative to their frequency in the individual populations. The Wahlund effect, as this process is known,

has an important medical implication. Due to genetic drift and founder effects, the frequency of recessive diseases, or abnormal phenotypes varies considerably between populations. For example, the combined frequency of mutations that cause cystic fibrosis is about 0.07 in Caucasian populations but is considerably lower in other races (e.g. Arab and African populations). Other recessive disorders at high frequency in particular populations include albinism in the South American Indian Hopi tribe and Tay Sachs disease in Ashkenazi Jews. Consequently, offspring where one parent is from a different race will tend to have a lower risk of inheriting a disease-causing mutation.

Unusual patterns of F_{ST}

Summaries of patterns of genetic variability at many loci paint an overall picture of genetic differentiation within a species. Yet some of the most interesting aspects of differentiation can only be seen by looking at a finer scale. The general picture for humans and *D. melanogaster* is that patterns of allozyme and DNA variability tell the same story about levels of genetic differentiation. However, this is not always the case. In the American Oytser (*Crssostrea virginica*) allozyme variation shows no differentiation between Atlantic populations and those from the Gulf of Mexico. However, looking at DNA variation, there is a sharp discontinuity in allele frequencies between the two populations, which is particularly pronounced for mtDNA. Very similar sharp discontinuities are also seen in mtDNA from a diverse array of organisms including Sea Bass and the Seaside Sparrow. The difference between DNA and allozyme studies suggests the influence of natural selection on protein variability, but there is no clear understanding of how selection might be acting.

Variation between loci in levels of differentiation also provides a fascinating window into the processes creating genetic differentiation. A study of eight allozyme loci in the Checkerspot butterfly (*Euphydrya editha*: McKenchie *et al.* 1975) found similar, low levels of differentiation for seven of them, but one locus, *hexokinase* has a much higher F_{ST}. One possibility is that ecological differences between the population studies have driven local adaptation at this gene in different directions in different populations. However, testing this hypothesis is not a straightforward process.

Finally, it is worth reiterating some of the problems with using *F* statistics as a measure of population differentiation. First, delineating populations, or geographic levels over which to test is arbitrary, and has the potential to be influenced by the data in such a manner that testing by permutation is not appropriate. Second, *F* statistics have large sampling variance, particularly when polymorphism is low. Finally, and perhaps most importantly, by focusing on a single summary statistic, a huge amount of information is thrown away.

Population genetic models of structure

The aim of population genetics is to understand the forces that shape patterns of genetic variability within and between species. To understand how different evolutionary forces can create genetic differentiation between populations it is natural to analyse simple models that extract the key elements of the process we are interested in. In previous lectures I have introduced the Fisher Wright model as the standard for understanding patterns of genetic variability within populations. However, the Fisher-

Wright model assumes random-mating between all individuals. How can we introduce population structure?

There are two simple models that are widely used as caricatures of population structure. The island model was first introduced by Haldane and considers a single island that receives a constant proportion of migrants, m each generation from an infinitely large mainland population. There is no migration from the island back to the mainland. A subtle variant of this model is the n-island model, in which n identical populations exchange migrants each generation such that each population receives a proportion m/n of migrants from every other population. As the number of islands gets very large, the properties of the n-island model become very similar to those of the island model.

Identity by descent in the island model

As with the standard Fisher-Wright model, the natural place to start analysing the properties of the island model is to consider identity-by-descent (ibd) for alleles sampled from within a population (symbolised by f). That is, we wish to look at the build up of ibd within the island, starting from the current time and looking back to previous generations. Suppose we choose two chromosomes at random from within the island population. Looking backwards in time, there are three possible events that might have occurred in the previous generation. As in the standard model, both chromosomes may have come from the same parent, with probability $1/2N_e$ in a diploid population (where N_e is the effective population size of the island). If so, the alleles are identical by descent. Another possibility is that the chromosomes are derived from different parents, both of which were on the island. In this case, the identity-by-descent is f_{t-1}. Finally, we have the possibility that one parent was an immigrant from the mainland population. For each chromosome this has probability m, so ignoring the possibility that both parents were immigrants, the probability of migration is $2m$. What is the identity-by-descent in this case? What we are really interested in is the build up of identity within the population due to the local structure. So the ibd for chromosomes in this configuration is zero. Putting this together

$$f_t = \frac{1}{2N_e} \times 1 + 2m \times 0 + \left(1 - \frac{1}{2N_e} - 2m\right) f_{t-1}$$

Solving for the equilibrium

$$f = \frac{1}{1 + 4N_e m}$$

What does this mean? There are two important points raised by this result. First, the critical value for determining the build up of ibd within the island population relative to the mainland population is the product of the island effective population size and the migration rate. This should not be unexpected by now – we have seen in previous lectures how mutation, selection and recombination typically influence genetic diversity only through their product with the effective population size. This is because the effects of deterministic forces are only important relative to genetic drift (which occurs at the rate of $1/2N_e$).

The second point is that remarkably little migration is required to prevent the build up of ibd within the island population. The product $2N_e$ x m is the (effective) number of migrants (assumed to be

diploids) that appear in the island population each generation. So even a handful of migrants per generation are sufficient to prevent extensive ibd from accumulating within the island.

The relationship between the population migration rate and F_{ST}

We can use the result concerning ibd to tell us about the relationship between the migration rate and the level of genetic differentiation as measure by F_{ST}. A heuristic approach is to say that ibd is closely related to identity in state if the mutation rate is low relative to the migration rate and mainland population size. Under these circumstances the build up of identity in state within the island population relative to the mainland population is almost equivalent to the build up of ibd. In other words

$$\mathrm{f}_{ST} = \frac{1}{1+4N_e m}$$

Where f_{ST} is used, rather than F_{ST} to indicate that this is differentiation in ibd rather than heterozygosity. In the *n*-island model, it has been shown that

$$\mathrm{f}_{ST} = \frac{1}{1+4N_e m\left(\frac{n}{n-1}\right)^2}$$

Which converges on the result for the island model when the number of populations gets very large. These results suggest that a simple means of estimating the compound parameter $N_e m$ from empirical data is to use the moment estimate

$$N_e m = \frac{1-F_{ST}}{F_{ST}}$$

For example, if we take F_{ST} in humans to be 0.067, $N_e m$ is estimated to be 3.5. What should we make of this number? In truth, not much. .First, as I have said before, F_{ST} has large sampling variance, so the estimate of $N_e m$ will also have large variance. Second, if we plot the relationship between F_{ST} and $N_e m$, it is clear that for F_{ST} values less than 0.1 (the usual situation) there is very little power to accurately estimate $N_e m$. In short, do not trust moment estimates of $N_e m$ from F_{ST}.

Wright's diffusion model for allele frequency differentiation

The relationship between identity-by-descent and f_{ST} is just one of many possible ways of looking at the effects of population structure on genetic differentiation. Wright (1931, 1951) took a different approach, by extending his diffusion theory method for looking at the effects of mutation and selection on the distribution of allele frequencies within populations. Consider the island model in which migrants from the mainland population replace a fraction *m* of the population each generation. Wright wanted to ask how genetic drift within the island population may lead the frequency of an allele on the island to vary relative to the mainland. If the mainland population is very large relative to the island, the frequency of an allele among migrants, x_I will be constant over time. Using the usual diffusion theory notation, we can describe the mean and variance in change in allele frequency within the island, *x*, over a single generation

$$M_{\delta x} = m(x_I - x)$$
$$V_{\delta x} = \frac{x(1-x)}{2N_e}$$

Over an infinite collection of identical populations, the allele frequency distribution at equilibrium is

$$\phi(x \mid x_I) = Cx^{4N_e m x_I - 1}(1-x)^{4N_e m(1-x_I)-1}$$

Where C is a normalising constant. Two examples are shown, where $4N_em$ = 0.2 and $4N_em$ = 10. When migration is low, populations can substantially diverge in allele frequency from the mainland, but when migration is high, island populations are fairly tightly clustered around the mainland frequency.

Wright's allele frequency distribution gives us much more information about the process of differentiation. We might therefore hope to get much better estimates of the important parameters if we could apply these ideas to empirical data. A natural possibility is to use allele frequency distributions in different populations as an estimator of gene flow between them. For example, in the SNP data collected by Goddard *et al*. (2000), we can compare the estimated allele frequencies in populations relative to the estimated worldwide allele frequency. The African Americans show moderate differentiation in allele frequency from worldwide frequencies, with a tendency for intermediate frequency SNPs to show greater divergence. Although the analysis is not rigorous (for example it assumes independence between SNPs and the worldwide allele frequencies as having no sampling error), we can ask about the likelihood of seeing this much divergence in an island model for different values of the parameter N_em. Combining Wright's allele frequency distribution with Ewens' sampling theory, gives the likelihood surface in the lower part of the slide. The maximum likelihood estimate of N_em is 5.0 and we say something about how likely the data is under different values.

While this analysis uses much more of the information in the genetic data, it suffers from two very serious limitations. First, the island model is clearly inappropriate for the data, but there is no coherent theory for allele frequency distributions in non-equilibrium models. Second, diffusion theory is not tractable for more than one locus. There is simply no way of incorporating information about linkage disequilibrium to give greater power. Fortunately, both these problems are relatively easy to deal with under a coalescent model. (The one situation concerning population structure where diffusion theory is currently more powerful than coalescent theory is in the case of continuous population models – as opposed to the discrete populations imposed here).

The coalescent in structured populations

The coalescent is a statistical description of the genealogical history of a sample taken from a population. Looking backwards in time, we can trace the line of ancestry from a chromosome in the current sample until the point where it coalesces with the ancestral lineage leading to another chromosome in the sample. In previous lectures we have discussed the coalescent process in standard Fisher-Wright population models, and how it can be adapted to incorporate recombination, population growth, and even types of natural selection. It is a simple manner to adapt the coalescent to describe ancestral processes under population structure.

Suppose you have sampled k chromosomes from a single population within the n-island model. As previously, we look back in time to the first event. Every generation the probability of one of the pairs of lineages coalescing is

$$\Pr\{Co\} = \frac{k(k-1)}{4N_e}$$

The probability per generation that one of the lineages was the offspring of an immigrant to the population is

$$\Pr\{Migration\} = km$$

As before, we rescale time as a continuous variable in units of $2N_e$ generations, and write $4N_em = M$. Looking backwards, the time until the first event is exponentially distributed with rate

$$\lambda = \binom{k}{2} + \frac{kM}{2}$$

And the probability that the first event is a coalescence is

$$\Pr\{1^{st} \text{ event coalescence}\} = \frac{k-1}{k-1+M}$$

If the first event is a migration (which occurs with probability)

$$\Pr\{1^{st} \text{ event migration}\} = \frac{M}{k-1+M}$$

Then we choose one lineage at random and assign the location of the parental chromosome at random from the n-1 other populations in the species range (under the n-island model). For more complex migration patterns, the ancestral population for the immigrant is picked from the backwards migration matrix. The process can begin again, with the proviso that the total rate of coalescence when ancestral chromosomes are spread across two or more populations is

$$\Pr\{Co\} = \sum_i \binom{k_i}{2}$$

This scheme allows rapid simulation of genealogical histories for a wide range of demographic scenarios. Furthermore, it can be adapted to include variable population sizes, migration rates, recombination, mutation, non-stationary migration patterns. However, for most cases, there are no simple analytical results – and extensive simulations over parameter ranges are necessary to investigate the full dynamics.

Pairwise coalescence time in the structured coalescent

There is one analytical result of importance that arises directly from the structured coalescent. Consider the history of a pair of chromosomes sampled at random from within one population. What is the expected time to coalescence for this pair of chromosomes? Looking backwards in time, the waiting time until the first event is exponentially distributed with rate

$$\lambda = 1 + M$$

And the probability that the first event is a coalescent is

$$\Pr\{1^{st} \text{ event coalescent}\} = \frac{1}{1+M}$$

Let T_W be the time to coalescence for a pair of chromosomes currently in the same population, and T_B be the time to coalescence for a pair of sequences in different populations. We can write

$$E[T_W] = \Pr\{1^{st} \text{ event = coalescent}\} \times E[\text{time to first event}]$$
$$+ \Pr\{1^{st} \text{ event = migration}\} \times (E[\text{time to first event}] + E[T_B])$$

Algebraically

$$E[T_W] = \left(\frac{1}{1+M}\right)^2 + \frac{M}{1+M}\left(\frac{1}{1+M} + E[T_B]\right)$$

For two chromosomes in different populations, coalescence cannot occur. The probability of either chromosome migrating to the same population as the other allele per generation is

$$\Pr\{\text{migration to same population}\} = \frac{2m}{n-1}$$

So we can write

$$E[T_B] = E[\text{time until in same population}] + E[T_W]$$
$$= \frac{n-1}{M} + E[T_W]$$

This gives a pair of simultaneous equations that can be solved to give

$$E[T_W] = n$$
$$E[T_B] = \frac{n-1}{M} + n$$

In other words, when sampling within a population, the expected time to coalescence (hence also the expected pairwise differences in the infinite sites model) for a pair of chromosomes is equivalent to that expected if the entire ensemble of populations were a single panmictic unit. In contrast, the expected pairwise differences for a pair of chromosomes sampled from between populations can be much greater. However, for $M >> 1$, the effect of subdivision on total diversity will be small.

While subdivision does not affect the expected value pairwise differences, it greatly affects the distribution. When migration between populations is low, most chromosome pairs will coalesce rapidly within the population, while a few will have much longer coalescence times as chromosomes. Consequently, by looking at the distribution of pairwise differences for chromosomes sampled within a population, it should be clear whether there is overdispersion relative to the single population expectation.

The effect of population structure on allele frequency

The coalescent within a structured population can almost be divided up into two separate phases that operate on different time scales. When migration rates are low, for chromosomes sampled from a single population we expect a rapid phase during which there are multiple coalescent events, but during which some lineages 'migrate' to other populations. When there is only a single ancestral lineage

remaining in the sampled population, the second phase begins, during which ancestral lineages in different populations slowly migrate around the species range, with occasional coalescent events.

Because the second phase occurs on a much longer time-scale than the first phase, most mutations segregating in a sample will have occurred during this phase. If mutations occur on a branch that is the ancestor of multiple chromosomes within the sample, such mutation will be at high frequency. Because of the rapid coalescence during the first phase, this is much more likely in the structured coalescent than in the standard coalescent in a panmictic population. Consequently, if we look at the expected frequency distribution of derived mutations under a structured population model, we see an excess of high frequency mutations.

Can we use standard techniques for detecting departures from neutrality to detect this effect? In general, the answer is no. Standard statistics, such as Tajima's *D* statistic and Fu and Li's *D* statistic are hardly affected by population structure – either in terms of the mean value, or the distribution. There is a slight tendency for greater variance with population structure, but the effect is small. The reason is that neither test uses information on the relative frequency of ancestral and derived mutations. Other tests, for example that of Fu (1995), developed by Fay and Wu (2000) explicitly uses information on the frequency of derived mutations as inferred from outgroup sequences (e.g. chimps for humans and vice versa).

Patterns of variability between populations

A very different picture emerges if we look at patterns of genetic variability for sequences sampled from different populations when there is population structure. These charts show the distributions of the Tajima *D* statistic and derived allele frequency when 10 chromosomes have been sampled from each of two populations that exchange migrants at the rate $M = 0.2$. These patterns are quite different from the case of a sample within a population. First, Tajima's *D* statistic tends to be positive in this case. The reason for this is clear if we look at the allele frequency distribution at segregating sites. There is a peak of segregating sites at an intermediate allele frequency, caused by the rapid coalescence of lineages within both populations, but slow coalescence between them

Linkage disequilibrium in structured populations

So far we have only considered how structure affects patterns of variability at a single locus. One of the most interesting, and underdeveloped areas of population genetics is how population structure affects patterns of association between alleles at different loci – linkage disequilibrium. The classical definition of linkage disequilibrium for a pair of alleles (*A* and *B*) at two loci is

$$D_{AB} = f_{AB} - f_A f_B$$

Linkage disequilibrium (LD) is generated by the random processes of mutation and sampling in a finite population, and is broken down by recombination. Population structure affects patterns of LD in two ways. First, for chromosomes sampled from the same population, structure tends to increase LD relative to the case of no structure. This is because the rapid coalescence within a population generates high frequency derived mutations that are in complete association with each other – leading to an excess of variants in near total association. This is seen in the chart of the distribution of

$$r^2 = \frac{D_{AB}^2}{f_A(1-f_A)f_B(1-f_B)}$$

For the two-locus, two-allele case. The second effect of structure on LD occurs when chromosomes from different populations are compared. Suppose we have two isolated populations, both of which are in complete linkage equilibrium, but there are differences in allele frequency between the populations. If we did not know that the populations were separate, a naive analysis would detect linkage disequilibrium between alleles, even at unlinked loci. The magnitude of LD caused by this process is proportional to the difference in allele frequency between the populations. For a pair of populations, if we write

$$\delta_A = f_A^1 - f_A^2$$

And something similar for the *B* locus. The apparent LD in the naive, combined analysis is

$$D_{AB} = \tfrac{1}{4}\delta_A\delta_B$$

In general, if we are combining data from many populations, the apparent LD between unlinked loci (which are in linkage equilibrium within each population) is

$$D_{AB} = Cov(f_A, f_B)$$

One interesting feature is that while the covariance will be non-zero for any two-population comparison, as the number of populations combined increases, the apparent disequilibrium will tend to be smaller.

In this analysis we are pretending that there are two populations that are in fact separate, but that we are unaware of the distinction. There is another, biologically important situation in which very similar rules apply. The term admixture is used to describe the combination of two (or more) previously separate populations. Admixture is very common in humans, and probably also in human commensals, because of large-scale changes in migration patterns over human history. For example, interbreeding between American Indians and Europeans, between Africans and other races in South Africa, between the settlers of north and south Japan, brought together genetic material from previously differentiated peoples. Consequently, differences in allele frequency between these groups will tend to generate apparent LD between even unlinked loci. Recombination in subsequent generations will slowly erode LD over time, but significant levels of LD can persist for many generations following secondary contact. Admixture is a particularly important problem in applying population genetic methods to disease mapping – a topic that will be covered in much greater depth in the last lecture.

Selection in structured populations

Coalescent theory provides a powerful way of predicting patterns of genetic variability in structured populations for neutral mutations. Furthermore, the coalescent can be adapted to include features such as time-varying migration rates and changes in population size, which are common elements of biological reality. However, for many people, the goal of evolutionary biology is to understand how natural selection shapes variation, both within and between species. In the last part of this lecture I want to talk about how population structure influences the process of natural selection. In general, this

is not an area under which we can use standard coalescent theory, and other mathematical approaches are necessary.

There is a large and highly technical literature on selection in subdivided populations, yet the main results can be reduced to relatively few key concepts. The first question we might ask is how structure affects the fixation probability of beneficial mutations, and hence the rate of adaptive evolution. Suppose a new mutation appears that is beneficial to all individuals in all environments, and has a fitness advantage of s relative to the wild-type. Maruyama (1970) used a branching-process argument to show that for the n-island model, the fixation probability of such unconditionally beneficial mutations is essentially unaffected by population structure. That is, the fixation probability is given by Haldane's original approximation of $2s$. This result is important, because it means we don't need to worry about including structure into general models of species adaptation. It should be noted that while the fixation probability is unaffected by structure, elements of the fixation process such as the time to fixation and the allele frequency distribution en route to fixation are considerably affected by structure.

The second type of problem we may want to address is what happens when different genotypes are favoured in different places. That is, there is environmental heterogeneity across a species range and this creates different selection pressures in different places. Can spatially varying selection pressures maintain polymorphism within the population? Levene (1953) showed that under certain circumstances, environmental heterogeneity can, in fact, maintain polymorphism within a species. Suppose there are just two types of habitat, scattered across a species range, and just two types of genotype. One genotype is favoured in one habitat; the other genotype is favoured in the other habitat. If environmental heterogeneity is fine-grained, such that individuals experience both habitats during their lifetime, then the genotype with the highest mean (geometric) fitness will spread to fixation. However, if heterogeneity is coarse-grained, and individuals experience only a single habitat during their life, then polymorphism can be maintained, even if offspring disperse evenly over the species range (see Barton and Clark 1990 for mathematical details).

Levene's result is of considerable importance, but its generality has been questioned. A number of authors have pointed out that the conditions under which polymorphism is maintained in the Levene model are very narrow – selection has to be strong and finely balanced against habitat frequency. Modifications to the model, such as habitat choice and assortative mating make the conditions less restrictive, but it is clear that the Levene model is not a general explanation for genetic polymorphism. Perhaps the single most unrealistic assumption in the model is that offspring disperse evenly over the entire species range. In most species, dispersal is localised. This creates correlations in the environment experienced by parents and their offspring, and creates the potential for local adaptation.

Local adaptation can occur when migration (offspring dispersal) occurs on a shorter scale than heterogeneity in the environment. Mutations that are beneficial within a region, but deleterious outside, can reach high frequency in localised patches. There are many examples of local adaptation, which occur at vastly differing scales. At the smallest scale, some plants that live in or near waste dumps have become tolerant to the high concentrations of heavy metals in the soil, while those plants

just a few metres beyond show no tolerance. At a larger scale, melanism (black pigmentation) in the peppered moth (*Biston betularia*) in Britain varies over the scale of a hundred miles or so and is associated with the degree of polluation (black morphs occur are cryptic on polluted trees but visible on non-polluted trees). At the continental scale, variation in humans in hair colour and skin pigmentation suggests variation in selective pressures over thousands of miles.

Indirect evidence for local adaptation: clines

When one allele is favoured in one place and another in a different place, local adaptation can occur if migration rates are low. But migration, however slow, will ensure that genotypes from one place end up in the other. Consequently, local adaptation will result in relatively smooth gradients in allele frequency at selected loci over the scale of environmental heterogeneity.

Such gradients in allele frequency are known as clines. And the detection of clines is one way of indirectly detecting local adaptation. One of the most famous clines in population genetics is the gradient in the frequency of the fast and slow (electrophoretic) alleles of the enzyme Alcohol dehydrogenase (Adh) in *Drosophila melanogaster*. Adh breaks down alcohol (present in the flies' diet as they eat fruit), and the fast allele has a two-fold higher level of activity than the slow variant. The fast allele is at high frequency in northern Europe and the north of the USA, and the slow variant is at high frequency in southern Europe and Africa and in the southern USA. In order to look more closely at the cline in allele frequency across the USA, Berry and Kreitman (1993) carried out an RFLP analysis of variability within the *Adh* gene. From Louisiana to Maine the fast allele varies in frequency from 0.15 to 0.5, while most other polymorphisms within the gene show no such trend. However, they also found another polymorphism, an insertion-deletion polymorphism called *Ⅴ1*, which shows a more pronounced cline (frequency changes from 0.05 to 0.6) and is almost complete linkage disequilibrium with the fast/slow variant. It seems likely that in fact this polymorphism is the target of selection, and that the gradient in the fast/slow polymorphism is an indirect consequence of linkage disequilibrium (and maybe also epistatic selection).

That markers closely linked to sites experiencing selection (local adaptation) may show similar patterns of geographic variation as the selected mutations themselves provides a potential way of detecting local adaptation without full characterisation of all genetic variation. The most extreme example of this situation occurs when two partially reproductively isolated species are brought into secondary contact. As described earlier in this lecture, admixture between previously isolated populations creates strong linkage disequilibrium even between unlinked markers, simply due to allele frequency differences between populations. If the offspring of matings between the two species/populations suffer a strong fitness disadvantage due to incompatibilities at many loci across the genome, indirect selection on neutral markers due to linkage disequilibrium with the selected loci creates an effective barrier to gene flow across the entire genome. Regions where previously isolated species come into contact are called hybrid zones. For example, there is a hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata* in Poland. Within hybrid zones, there are steep, concordant clines in allele frequency at neutral markers across the genome, and also in phenotypic traits (Syzmura and Barton 1991). The few instances where genetic variants from one population have

introgressed (spread into) the other population may be indicative of the spread of unconditionally beneficial mutations.

Another way of using linkage between neutral markers and selected loci is to look for the traces of local selective sweeps. Local selective sweeps occur when a new mutation that is locally advantageous arises in a population and sweeps to a high local frequency, removing variation at linked, neutral loci. Locally reduced variability at a marker that is consistently variable in other populations may be indicative of local adaptation. For example, in a study of 10 microsatellite loci across worldwide populations of *D. melanogaster* Schlötterer *et al.* (1997) found a number of instances where one marker showed unusually low variability in a single population. However, it should be noted that interpretation of this data is not straightforward. As we saw earlier in this lecture, when there is population structure we expect considerable variability between loci in the depth of the genealogy, hence the number of mutations in a sample. Whether the *Drosophila* data can be explained simply through population structure remains to be explored.

Sewall Wright's shifting-balance theory

Finally, it is worth discussing one of the most important and contentious theories relating to population structure. Sewall Wright's overwhelming passion was population structure – much of theory in this lecture is due to him – and his great ambition was to combine his work on drift and selection in subdivided population into a single, general theory of evolution. This theory has become known as the shifting-balance theory. The shifting-balance argues that the majority of adaptation in species occurs not through the mass selection principles expounded by Fisher and Haldane, but in a manner that can only work in subdivided populations. The key feature of the shifting-balance theory is that alleles at different loci in a genome interact such that there is no simple relationship between genotype and fitness. This is element, called epistasis, formally states that the fitness effects of alleles at different loci are not multiplicative. For example, suppose we have two loci, and two alleles at each. Suppose the fitnesses of different genotypes are

	AA	*Aa*	*aa*
BB	1	$1-h$	$1-2h$
Bb	$1-h$	$1-2h$	$1-h$
bb	$1-2h$	$1-h$	$1+s$

If the population is initially fixed for alleles *A* and *B*, then while the genotype *aabb* is fitter than the genotype *AABB*, in order to reach the state *aabb* the population has to decrease in fitness. By way of analogy, cycling is faster than walking, but only if you have two wheels – just one wheel would slow you down.

Epistasis between alleles creates a complex surface of population fitness (a function of allele frequency) that is known as the adaptive landscape. Earlier Wright had shown that the expected change in allele frequency due to selection is

$$E_S[\Delta x] = \frac{x(1-x)}{2\bar{w}} \frac{d\bar{w}}{dx}$$

Where x is allele frequency and w with a line above is the mean population fitness. So an allele will only increase in fitness by selection if it increases mean population fitness. Consequently, under the mass-selection rules of Fisher and Haldane, the population will never go from *AABB* to *aabb*. However, things are different in a subdivided population. Actually, the important thing is the subdivided populations consist of multiple finite populations. As we discussed in previous lectures, when the population is small, genetic drift can lead to deleterious mutations reaching high frequency. Consequently, in a small, finite population, there is some chance that the population will drift down the adaptive landscape to a point of lower fitness, before going up the other side (through selection) and reaching the higher peak. In the language of the shifting-balance, the population can cross the adaptive valley. Partially isolated populations can therefore be thought of as natural experiments, allowing a species to try out different regions of the adaptive landscape.

The ideas of the shifting-balance are intuitive and visually powerful. Furthermore, there is considerable evidence for epistasis in natural populations. F2 hybrid breakdown (the low fitness of second generation hybrids) can be explained by the breakdown of coadapted gene complexes (Fenster *et al.* 1997), and some coadapted gene complexes are well known (e.g. genes controlling mimicry in the butterfly *Heliconius*). However, there is a good theoretical reason to suppose that the shifting-balance is not the general explanation of adaptation that Wright wished for. The main problem is the last phase of the process. Once a subpopulation has reached the new, higher peak, this genotype then has to spread throughout the rest of the species (see Coyne *et al.* 1997). The problem is that *aabb* genotypes spreading throughout the rest of the species will tend to mate with *AABB* genotypes and consequently will produce offspring with low fitness (for exactly the same reasons we get F2 breakdown). Adaptation will tend to be restricted to the local population.

Although Wright's theory may not be a general explanation for adaptation within species, it seems quite plausible that it is an important feature of local adaptation. Or at least that local adaptation can create epistatic interactions between alleles that are then exposed when populations are brought into secondary contact. For example, Haldane's rule of unisexual sterility and inviability in species crosses is probably explained by epistasis. Haldane's rule states that when only one sex of hybrids between two species is sterile, it is heterogametic sex (the XY sex or equivalent). In mammals the heterogametic sex is male, but in birds and butterflies, it is the female. Sterility and inviability in these cases seems to be caused by a breakdown of recessive epistatic interactions between alleles at loci on the X chromosome (Z in birds) and autosomes. Epistasis is probably an important feature of evolution, but not in the way Wright supposed.

