Notes for Lecture 5 Population subdivision Part 2 Gene flow

First a little bit about spatial patterns within populations. Your book covers this on pages 161-163, but I think it’s worth mentioning in your notes as well. As you can imagine, genotypes are not necessarily arranged at random through the landscape. Within a species a range, there may be differences in diversity but also differences in genotypes, especially for mobile species. While we were looking at a large geographic scale, we consider small geographic patterns even at the point of the local population. This is true for both mobile and non-mobile (plants, some marine invertebrates) species. Of course, this is of particular concern when thinking about sampling a local population.

Important parts of this puzzle are the dispersal distance, dispersion, and density of the study species. Typically, we use spatial autocorrelation among genotypes at varying distances to quantify the spatial genetic structure within populations. We will talk about this more when we look at landscape genetics.

Gene Flow

First consider the case were we have two populations of equal size that exchange individuals in both directions at a rate $m$.

$$q'_A = (1 - m)q_A + mq_B$$

Here $q'$ is the frequency of the $q$ allele in Population A after one generation. These two populations will reach equilibrium when $q_A' = q_B'$. We referred to this as $q^*$ in the past. In this case $q^* = \text{average of the initial allele frequencies in the two populations}$. The example from your book has a $m = 0.10$ (10% exchange per generation), and $q$ is fixed in population A and 0 in population B. Here $q^* = (1 + 0)/2 = 0.5$: 
**Gene flow, the great homogenizer:** In the absence of any other evolutionary force gene flow, even a wee bit around 1 migrant per generation, will make two populations homogeneous, given enough time. In addition, as you can see from the example above, it will also increase the diversity of the subpopulations unless the frequency of alleles is the same between the two populations. Gene flow is the main evolutionary mechanism that homogenizes subpopulations while genetic drift increases the genetic divergence between them. So too, in cases where there is divergent directional selection between subpopulations, gene flow will counteract this divergence. Let’s consider the relationships between gene flow and drift and gene flow and natural selection. In the process let’s look at that in the context of different gene flow models: island, stepping-stone, and isolation by distance. In a few lectures we will look at these questions in the context of metapopulation dynamics.

**Island model of migration:** In the island model each population is spatially arranged such that exchange of individuals is equal to all of the populations, including back into the population of origin. All of the populations are also of equal size ($N$). To correct the diagram below you need to add arrows that indicate migration back into the subpopulation:

Each subpopulation contributes an equal number of genes to the migrant pool and the proportion of migrants is $m$ while the ones staying is $m-1$. Now we can look at the relationship between drift and gene flow. While genetic drift increases the homozygosity and therefore divergence between the subpopulations, gene flow will counteract this by decreasing the divergence among the subpopulations as long as $m > 0$. Eventually you’ll reach a steady-state equilibrium between these two evolutionary forces, drift and gene flow. Wright’s equation for an island model of gene flow and with an infinite number of subpopulations is:
POPULATION SUBDIVISION: GENE FLOW

\[ F_{ST}^* = \frac{(1 - m)^2}{[2N - (2N - 1)(1 - m)^2]} \]

If \( m \) is small enough and \( N \) large enough the terms \( m^2 \) and \( m/N \) can be ignored, with that and some rearranging you can use this equation:

\[ F_{ST}^* = \frac{1}{1 + 4Nm} \]

You’ll see that this is the identical equation that is used for mutation rate as well (coming shortly). What I like about this equation is that the \( Nm \) product makes biological sense. First, the number of alleles in a subpopulation of size \( N \) diploid organisms is \( 2N \). In a generation the proportion of alleles that are replaced by alleles from migrant organisms is \( m \), therefore the number of migrant alleles is \( 2m \), giving you \( 4Nm \). Therefore, \( Nm \) can be interpreted as the absolute number of migrant organisms that come into each subpopulation in each generation. If \( m = 0 \) then \( F_{ST} = 1 \).

- \( Nm = 0.25 \) (one migrant every 4\textsuperscript{th} generation), \( F_{ST} = 0.50 \)
- \( Nm = 0.50 \) (one migrant every 2\textsuperscript{nd} generation), \( F_{ST} = 0.33 \)
- \( Nm = 1 \) (one migrant every generation), \( F_{ST} = 0.20 \)
- \( Nm = 2 \) (2 migrants every generation), \( F_{ST} = 0.11 \)

The take home message is that migration is a serious force against divergence among subpopulations. Now the equations above assume an infinite number of subpopulations, which can be corrected to the following:

\[ F_{ST}^* = \frac{1}{1 + 4Nam} \]

Where \( a \) is:

\[ a = \left( \frac{n}{n-1} \right)^2 \]

As you book points out, the amount of divergence among the subpopulations depends on the number of migrant organisms (\( Nm \)) and not on the proportion of exchange (\( m \)), only on the \# of migrants irrespective of population size. If you think about it, this is an equilibrium equation between drift and migration, and the larger the population the fewer migrants are needed to counteract the effects of drift.

**Stepping-stone model & Isolation by distance:** In this model migration is greater between nearby subpopulations that are near each other.
The idea is that there is more frequent short-range migration and fewer events of long-range migration. This can be applied to many populations that have a linear arrangement such as along a river, along a coastline, up a mountain ridge-line, etc. While in the island model genetic divergence is independent of geographic location, here there will be greater divergence as you increase distance between populations. Typically we demonstrate this in a graph that has distance along the x-axis and pairwise \((F_{ST})/(1-F_{ST})\) as genetic distance.

We can take this concept to continuously distributed populations, in which we no longer able to define discrete populations but only the neighborhood size. Wright developed an equation for the effective number of parents in a neighborhood or Neighborhood Size:

\[
NS = 4\pi\sigma^2 D
\]

In this equation, \(D\) is the density, and \(\sigma^2\) is the mean squared parent-offspring axial distances. NS is equivalent to then number of reproducing parents in a circle of radius \(2\sigma\) that includes approximately 88% of the parents in the center.
What about cytoplasmic DNA?
mtDNA has a smaller effective size, therefore it will show signs of greater genetic differentiation than autosomal genes. What is the effective size of mtDNA? If you assume transmission is exclusively from mothers (which is not the case for some marine inverts), the chance of drawing two mtDNAs from a population that are IBD is \(1/N_f\). \(N_f\) is the number of females. While the probability of two randomly chosen autosomal genes are IBD is \(1/(2N_E)\). If we equate the two equations we see that \(1/(2N_E) = 1/N_f\) resolves to \(N_E = N_f/2\). This is the effective size for the population of mtDNA molecules. Because \(N_f = N/2\) where \(N\) is the total number of individuals in an ideal population, the effective size for mtDNA, relative to an autosomal gene in an ideal population is \(N/4\). What does this mean for genetic divergence? The greater the genetic drift with the smaller effective size means there should be greater genetic divergence among subpopulations for the same amount of gene flow. Given equal gene flow in males and females, which is an assumption with interesting violations, and with an island model of migration we can use the following equation:

\[
F_{ST} \approx \frac{1}{Nm + 1}
\]

Usually this equation is written to consider only females (note that in an ideal population \(N_f = N_m\), hence the \(2N_f\)):

\[
F_{ST} \approx \frac{1}{2N_f m + 1}
\]

Gene Flow and Natural Selection:
As we saw above the effectiveness of natural selection depends in part on the strength of selection against random change in allele frequencies (drift) and in the homogenizing effects of migration. In the case of heterozygote advantage or overdominance in which both homozygous genotypes have a similar selection coefficient \((1-s)\), we would expect an equilibrium \(p^* = 0.5\). Even in the face of several migrants per generation, selection will keep the subpopulations equivalent, especially in subpopulations of large size (excluding the effects of drift). As the selection coefficient increases this is more the case. Similarly, in a different situation where there are differences among subpopulations in directional selection (divergent directional selection), the differences in allele frequencies among the subpopulations will be maintained even with migration, which as above depends on the strength of selection and the size of the subpopulations.

Couple of things to consider:
Drift and gene flow affect all loci, while selection only affects loci that are influenced by selection. This means that among loci there will be differences in \(F_{ST}\) among populations. We use this to help locate outlier loci, those loci that have higher \(F_{ST}\) against background \(F_{ST}\). We want to identify these outlier loci for two reasons (1) they may be influenced by selection and may help us understand adaptation, and (2)
they may bias the degree of divergence among populations if they are used and make it harder for us to infer the historical relationships among populations. For example for estimates of time since divergence and migration rates, we want to make sure we are looking at neutral, not selected loci.

**Limitations of measures of population subdivision:**

$F_{ST}$ by itself does not consider the genealogical degree of relatedness. For microsatellite loci, an allele with a one-step mutation (the loss of gain of a repeat unit) is more related to the allele that is one repeat unit away than the one that is 10 repeat units away. There are other measures such as $R_{ST}$ and $\phi_{ST}$ that use the degree of differentiation between alleles as a weighting factor that increases the metric proportional to the number of mutational differences between alleles. For measures of $R_{ST}$ you need to know how many repeat units you have and this may be why you don’t see it as often in publications. More recently, a population specific measure of $F_{ST}$, $F'_{ST}$ gives a better estimate of the IBD of alleles than a global measure of $F_{ST}$ and shows great promise.

For those markers that produce a lot of alleles, such as microsatellite markers or AFLPs, there is a trend to see much lower $F_{ST}$ values than if you looked at the same populations with nucDNA or mtDNA markers with only a few alleles. If you have high within population variation, the proportion of the total variation that is distributed between populations can never be very high. One time at a meeting I had a senior faculty member (okay an old famous guy) question my $F_{ST}$ values in my cryptic species. “How can you say these are two species with an $F_{ST}$ of 0.15?” I was so ready for the question that I could hardly contain my excitement as I explained the situation. He nodded his head, and that’s when I realized he was baiting me! There are different ways to correct for this, your book mentions quite a few, but proceed with caution because there are issues with the assumptions of all these measures. Typically, researchers approach this question by using several different tools to analyze differences among populations. For example, I used sequence data and microsatellite data when describing the cryptic species. I also compared adults and juveniles to look for hybrids. Instead of using traditional methods, I opted to used maximum likelihood and Bayesian methods to look at divergence and gene flow as well.

Your book does a great job of explaining all the scenarios that may affect genetic divergence among populations (pgs 174 – 178).

**Estimations of Gene Flow:**

Indirect measures of gene flow.

$$Nm \approx 1 - \frac{F_{ST}}{4F_{ST}}$$

There are certain assumptions for this measure. This assumes the island model of migration, infinite number of subpopulations, migration is symmetrical and the populations are the same size, selectively neutral, no mutation, that the population
POPULATION SUBDIVISION: GENE FLOW

is at a drift-migration equilibrium, and demographic equality of migrants and residents. You can imagine that all these assumptions are unlikely to hold true. However, this may give a reasonable estimate even when certain assumptions are violated. Typically you will see this indirect measure used more for a magnitude of migration rates (high versus moderate versus low) than for an absolute measure of migration.

There are other measures that are used, including a private allele method. The idea here is that there is a linear relationship between the number of private alleles and the degree of migration. This is likely better for highly polymorphic loci, like microsatellites.

Typically maximum likelihood, Bayesian, and coalescent methods, are better at making estimates of the amount of migration because many of the assumptions above can be violated. For example we can relax the assumption that migration is symmetric or that population sizes are identical. They use all the data provided for each population, and not summary statistics like $F_{ST}$. Several different programs are used including FLUCTUATE, MIGRATE, and GENETREE, but there are many more.

Direct measures of migration can come from mark-recapture and other related methods. A benefit of this method is that it gives real-time estimates of migration and does not make assumptions like those above. This pro is also a con. What we want is to understand is both short-term and long-term (migration that may only happen every 5 to 10 years) patterns of connectivity between populations. Moreover, these direct measures do not measure whether these individuals came and bred with individuals in the population. There may be lower or higher reproductive success of the migrants compared to the residents. One way you can get a better handle on the migrants is through assignment tests. We remove one individual, calculate the frequency of that particular genotype in each of the populations (usually done with many loci to be effective) and if that individual’s genotype is found at a higher frequency in a different population, then it is assumed to be a migrant. There are many different programs that use this method. The one I typically use is STRUCTURE because I can play around with different evolutionary models. We will talk about this program and others above later in the semester.