The Wahlund Effect and F Statistics -- The Interaction of Drift and Gene Flow

The distribution of genetic variation within and between demes is primarily due to the balance between gene flow and drift. We can measure this balance by a set of "F statistics." This yields yet another interpretation of the phrase "inbreeding coefficient."

Consider first a simple model in which a species is subdivided into \( n \) discrete demes where \( N_i \) is the size of the \( i \)th deme. Suppose further that the species is polymorphic at a single autosomal locus with two alleles (A and a) and that each deme has a potentially different allele frequency (due to past drift in this neutral model). Let \( p_i \) be the frequency of allele A in deme \( i \). Let \( N \) be the total population size (\( N=\sum N_i \)) and \( w_i \) the proportion of the total population that is in deme \( i \) (\( w_i=N_i/N \)). For now, we assume random mating within each deme. Hence, the genotype frequencies in deme \( i \) are:

\[
\begin{align*}
\text{AA} & \quad \text{Aa} & \quad \text{aa} \\
p_i^2 & \quad 2p_iq_i & \quad q_i^2
\end{align*}
\]

The frequency of A in the total population is \( p = \sum w_ip_i \). If there were no genetic subdivision (i.e., all demes had identical gene pools), then with random mating the expected genotype frequencies in the total population would be:

\[
\begin{align*}
\text{AA} & \quad \text{Aa} & \quad \text{aa} \\
p^2 & \quad 2pq & \quad q^2
\end{align*}
\]

However, with subdivision, the actual genotype frequencies in the total population are:

\[
\begin{align*}
\text{Freq}(\text{AA}) &= \sum_{i=1}^{n} w_ip_i^2 \\
\text{Freq}(\text{Aa}) &= \sum_{i=1}^{n} 2w_ip_iq_i \\
\text{Freq}(\text{aa}) &= \sum_{i=1}^{n} w_iq_i^2
\end{align*}
\]

By definition, the variance in allele frequency across demes is:

\[
\text{Var}(p) = \sum w_i(p_i - p)^2 = \sum w_ip_i^2 - p^2
\]

\[
= \sum w_i(q_i - q)^2 = \sum w_iq_i^2 - q^2
\]

Using these formulae relating the \( w \)'s and \( p \)'s to \( \text{Var}(p)=\sigma_p^2 \), the genotype frequencies in the total population can be expressed as

\[
\text{Freq}(\text{AA}) = \sum_{i=1}^{n} w_ip_i^2 - p^2 + p^2 = p^2 + \sigma_p^2
\]
\[ \text{Freq}(aa) = q^2 + \sigma_p^2 \]
\[ \text{Freq}(Aa) = 1 - \text{Freq}(AA) - \text{Freq}(aa) = 1 - p^2 - q^2 - 2\sigma_p^2 \]
\[ = 2pq - 2\sigma_p^2 = 2pq(1 - \sigma_p^2/pq) \]

Define \( F_{st} = \sigma_p^2/pq \). (Note, the \( F_{st} \) from the previous handout was defined in terms of ibd -- and is not necessarily the same as this one. As is traditional in population genetics, most papers do not tell you which \( F_{st} \) is being used, but the variance-one defined here is by far the more common one.) Hence, the genotype frequencies are:

\[ \text{Freq}(AA) = p^2 + pqF_{st} \]
\[ \text{Freq}(Aa) = 2pq(1 - F_{st}) \]
\[ \text{Freq}(aa) = q^2 + pqF_{st} \]

Hence the subdivision of the population into genetically distinct demes causes deviations from Hardy-Weinberg that are identical in form to those caused by an inbreeding system of mating within demes. This “inbreeding coefficient” is called \( F_{st} \) because it refers to the deviation from Hardy-Weinberg caused by allele frequency deviations in the subdivided demes from the total population allele frequency. This \( F_{st} \) is simply a standardized variance of allele frequencies across demes. In general, the more important drift is relative to gene flow, the larger the value of \( F_{st} \). For example, the Yanomama Indians are very war-like, and new villages are frequently formed from a group of related individuals that leave an old village due to a dispute. This “lineal fissioning” of villages accentuates founder effects (because the founding individuals are related). \( F_{st} \) among the Yanomama villages is 0.073. The nearby Xavante Indians are more peaceful and do not have lineal fissioning, and their \( F_{st} \) is 0.0091. On a worldwide scale, the \( F_{st} \) for the 3 major human races is about 0.15, only about twice as much differentiation as seen among Yanomama villages. Other species show much more subdivision than humans. E.g., kangaroo rats have an \( F_{st} \) of 0.676 throughout their range, and the \( F_{st} \) between blocks on the same street for the snail *Rumina* (which has mixed random-mating and selfing as well as limited dispersal capabilities) is 0.538.

What happens when you have inbreeding within demes as well? Let \( F_{is} \) be the inbreeding for individuals within a subdivision such that within each deme the genotype frequencies are:

\[ AA \quad Aa \quad aa \]
\[ p_i^2 + p_iq_iF_{is} \quad 2p_iq_i(1-F_{is}) \quad q_i^2 + p_iq_iF_{is} \]

Note, \( F_{is} \) measures deviations from random mating expectations within a local deme. Hence, it is the same as “\( f \)” in earlier lectures.

With respect to the total population, the genotype frequencies are now:
Freq(AA) = \sum_{i=1}^{n} w_{i} p_{i}^{2} + F_{is} \sum_{i=1}^{n} w_{i} p_{i} q_{i}

= p^{2} + \sigma_{p}^{2} + F_{is} \sum_{i=1}^{n} (p_{i} - p_{i}^{2})

= p^{2} + \sigma_{p}^{2} + F_{is} (p - p^{2} - \sigma_{p}^{2})

= p^{2} + pq[F_{it} + F_{is}(1-F_{st})]

Let F_{it} F_{st} +F_{is}(1-F_{st}) \quad \text{[Note, } 1-F_{it} = (1-F_{is})(1-F_{st})].

Then, the genotype frequencies are

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Aa</td>
<td>aa</td>
</tr>
<tr>
<td>p^{2} +pqF_{it}</td>
<td>2pq(1-F_{it})</td>
<td>q^{2} +pqF_{it}</td>
</tr>
</tbody>
</table>

E.g., Yanomama tend to avoid inbreeding within villages; and their F_{is} is -0.01 (recall F_{st}=0.073). For the Snail Rumina, F_{is}=0.77 and F_{st}=0.538.

These F-statistics are designed to help us distinguish between deviations from HW expectations due to two confounding factors, nonrandom mating within subpopulations and subdivision between subpopulations. By observing how allele frequencies vary within and between subpopulations relative to those for the entire population, we can further discriminate between these forces. F_{it} is a not-very-useful term that represents the confounded deviation from HW expectation.

**Relationship of F_{st} to N (drift) and m (gene flow).**

Although F_{st} is most commonly estimated in terms of variances of allele frequencies, most of the theory relating F_{st} to N (drift) and m (gene flow) is in terms of ibd. E.g., island model. The species is subdivided into discrete demes, each of inbreeding effective size N and with random mating within. Each generation, m individuals leave a deme and are randomly dispersed over all other demes. If an infinite number of demes is assumed, it is effectively impossible for two migrant alleles to be ibd with each other or with an allele from the deme into which they immigrate. Hence, the probability that two alleles are ibd at generation t within a deme is

\[ F_{t} = \left[ 1/(2N) + \{1 - 1/(2N)\}F_{t-1} \right] (1-m)^{2} \]

And at equilibrium (F_{t}=F_{t-1}=F_{eq})

\[ F_{eq} \approx 1/(4Nm+1) \]
and with mutation (\(\mu\)), \(F_{eq} \approx 1/([4N(m+\mu)+1])\) (recall previous handout).

Wright’s equations describing the above relationships depends on an assumption of the “Island Model” of population structure where there are an infinite number of subpopulations and migration from any subpopulation to any other is equally likely (no geographic isolation). Of course, this is a theoretical construct with little meaning in the real world. However, changing the model upon which the relationships are based also changes the relationships. More realistic models are available, and some will be considered now.

E.g. isolation by distance model. Suppose species subdivided into discrete demes, and as with island model, let \(m_\infty\) of individuals leave deme and disperse at random over entire species. However, also assume the demes are arranged along a 1-dimensional habitat (e.g., a river, shore-line, etc.) with \(m_1\) of the individuals being exchanged between adjacent demes only. Then:

\[
F_{st}(eq) \approx \frac{1}{1+4N\sqrt{m_\infty^2 + 2m_1m_\infty}}
\]

For \(m_1 = 0\), \(F_{st} = 1/[1+4Nm_\infty]\) (the island model).

For \(m_\infty\) very small relative to \(m_1\)

\[
F_{st} \approx \frac{1}{1+4\sqrt{2m_1m_\infty}}
\]

Note that from the above equation that even when \(m_\infty\) is very small relative to \(m_1\), that \(m_\infty\) has a major impact on genetic subdivision. This means that gene flow is very difficult to measure accurately from dispersal data because very rare, but long distance, dispersal events have a disproportionate impact.

It is also difficult to measure effective size from actual size. To see this, consider the continuous analog of the above models. In these models, the species is not subdivided into discrete demes, but is rather continuous. Nevertheless, there is isolation by distance because of limited dispersal. Let \(\delta\) be the density (replaces \(N\) in the discrete models), and \(\sigma\) the standard deviation of distance between birthplace of parent and offspring (replaces \(\sqrt{m_1}\)) for "normal dispersal", and retain \(m_\infty\) as the long-distance dispersal parameter. Then, at equilibrium

\[
F_{st} \approx \frac{1}{1+4\delta\sigma\sqrt{2m_\infty}} \quad \text{for 1 dimensional habitats}
\]

\[
F_{st} = \frac{1}{1+ 8\pi\delta\sigma^2 / [-ln(2m_\infty)]} \quad \text{for 2 dimensions}
\]

In analogy to discrete effective size, can use the above equations to define a "neighborhood size". E.g., for 2 dimensions, the neighborhood size is \(4\pi\delta\sigma^2\). Levin
and Kerster calculated neighborhood size in the plant *Liatris*, taking into account both seed and pollen dispersal. Did this for several years during which there were dramatic fluctuations in population density. Their results are:

<table>
<thead>
<tr>
<th>Year of Study</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in original</td>
<td>30</td>
<td>98</td>
<td>150</td>
<td>330</td>
</tr>
<tr>
<td>Neighborhood area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neighborhood size:</td>
<td>30</td>
<td>75</td>
<td>97</td>
<td>191</td>
</tr>
</tbody>
</table>

Note that the neighborhood size did not fluctuate as much as the actual population size. The reason was that as density increased, bees (the main pollinator) had to fly less distance between plants. Thus, although $\delta$ increased in some years, $\sigma^2$ decreased those same years. Hence, actual population sizes or densities are not necessarily reliable indicators of effective or neighborhood sizes. Effective sizes are a function of both population size and dispersal and the interactions between them.

**Multiple genetic systems and population structure.**

In several of our previous examples, we noted that sometimes we had to change a $4N_e$ to a $2N_e$ because a genetic system was haploid and not diploid, and sometimes we had to note that an $N_e$ was referring to only 1 sex because the genetic system was uniparental in inheritance. All of this seems bothersome, but by simultaneously studying several genetic systems we can gain far more insight into underlying population structure than from just one system alone.

**EX)** Hawaiian *Drosophila mercatorum*.

Flies were samples from two nearby regions in the Kohala Mountains on the Island of Hawaii (one in the higher rainforest region and one in the lower desert region):

Then performed nuclear isozyme studies (electrophoresis), mtDNA (RFLP analyses), and Y-chromosome studies (polymorphisms in the Y-linked rDNA complex). Note that the isozymes were biparental diploid systems, mtDNA is a maternally inherited haploid system, and the Y-DNA is a paternally inherited haploid system. Here we use variance definitions of $F_{st}$ and $N_e$.

Results:

\[
F_{st} \quad (N_2 m_2)
\]
isozymes not sign diff from zero any value > 2 N_e m_e gives this F_{st} result
mtDNA 0.17 N_e m_e of approx 1 gives this F_{st} result
Y-chrom. 0.08 N_e m_e of approx 3 gives this F_{st} result

These are all very different genetic systems that refer to how genes are transferred through space and time. Can these apparently different results be reconciled? For mtDNA, the effective size (N_e) is multiplied by 2 not 4 because it is haploid rather than diploids and the effective size is expected to be only \(1/2\) that of males and females considered together because mtDNA is maternally inherited. In all, the (ploidy coeff.) times (effective size) of mtDNA should be only 1/4 that of autosomal DNA (isozymes). Likewise, the Y-chromosome is also 1/4 relative to the diploid autosomal system.

Taking this information into consideration we can revise the N_e m_e estimates to reflect the differences in transmission of genes. As a result, N_e m_e for autosomes remains the same but mtDNA and Y-chromosome increase to 4 and 12, respectively, when put into biparental, diploid equivalents. Thus, both haploid results (with significant F_{st}'s) are consistent with the low nuclear F_{st}. This extends to range of effective migrants to 2 < N_e m_e < 12.

m_e may also be different for the different genetic systems. Studies have shown that D. mercatorum females are commonly inseminated by multiple males and that the females retain the sperm for extended periods of time. Therefore, dispersing females carry not only their own genes but those of 1-3 males as well. So, even with similar sex ratios and dispersal rates for males and females, we see much more gene flow of male gametes than female gametes. This is consistent with the lower F_{st}'s for Y-DNA than for mtDNA. Thus, by combining the results of several genetic systems, can gain much insight into the system, including sex-specific differences in gene flow.

**Final point.** F_{st} can be measured from genetic survey data. In contrast, measuring gene flow directly dispersal studies is very difficult and unreliable. The genetic theory shows that even very little exchange between populations can result in much effective gene flow. Hence, “rare” dispersal events have great evolutionary importance but are difficult to study. Moreover, even when you detect dispersal, it doesn’t mean the migrants you see will successfully mate in their new population. In addition, such direct observation nearly always misses the occasional or rare long-distance dispersal events which we have shown are quite effective in keeping populations from diverging from one another. Overall, direct measurement of gene flow is tedious, ineffective and tends to underestimate the true values. On the other hand, estimation of gene flow via F_{st} gives an “evolutionary” picture which automatically takes into account all these various possibilities. However, F_{st} is an effective measure of gene flow over evolutionary time if your underlying model of subdivision is accurate. That is a big IF.