Chapter 6

Gene Flow and Population Subdivision
In deriving the Hardy-Weinberg Law in Chapter 2, we assumed that the population was completely isolated. Isolation means that all individuals that contribute to the next generation come from the same population with no input at all from individuals from other populations. However, most species consist of not just one deme but rather many local populations or subpopulations consisting of the individuals inhabiting a geographic area from which most mating pairs are drawn that is generally small relative to the species' total geographic distribution. Although most matings may occur within a local population, in many species there is at least some interbreeding between individuals born into different local populations. Genetic interchange between local populations is called gene flow. In Chapter 1 we noted that DNA replication implies that genes have an existence in space and time that transcends the individuals that temporally bear them. Up to now, we have been primarily focused upon a gene’s temporal existence, but with gene flow we begin to study a gene’s spatial existence. In this chapter, we will study the evolutionary implications of gene flow and investigate how a species can become subdivided into genetically distinct local populations when gene flow is restricted. Restricted gene flow leads to variation in the frequency of a gene over space.

**The Balance of Gene Flow and Drift**

Recall from Chapter 4 that to measure the impact of genetic drift upon identity–by–descent, we started with equation 4.3:

\[
\bar{F}(t) = \frac{1}{2N} + \frac{1}{2N} \bar{F}(t-1)
\]

where \(N\) is replaced by the inbreeding effective size for non-ideal populations. To examine the balance between drift and mutation, we modified the above equation to yield equation 5.4:
Because gene flow and mutation behave in an analogous manner with respect to genetic variation
within a local deme, a similar modification of equation 4.3 can be used to address the following
question: suppose two local populations each of inbreeding effective size $N_{ef}$ are experiencing gene
flow at a rate of $m$ per generation. What is the probability that two randomly drawn genes from the
same subpopulation are identical by descent AND came from parents from that same
subpopulation? That is, if one of the genes came from the other subpopulation’s gene pool, we no
longer regard it as “identical.” The equation for this probability is then:

$$F(t) = \frac{1}{2N_{ef}} + \frac{1}{2N_{ef}} F(t-1) (1-m)^2$$

6.9

Equation 6.9 is the probability of identity by descent as a function of genetic drift in the local deme
(equation 4.3) times the probability that both of the randomly chosen gametes came from the local
deme. At equilibrium, equation 6.9 yields (analogous to equation 5.6):

$$F_{eq} = \frac{1}{4N_{ef}m + 1}$$

6.10

if $m$ is small such that $m$ is much greater than $m^2$ and $m$ is on the order of magnitude of $1/N_{ef}$ or
smaller.

Recall from equation 5.7 that the balance of mutation to genetic drift was measured by
$\frac{q}{N_{ef}}$. The balance of gene flow to genetic drift is measured in equation 6.10 by a similar
parameter: $4N_{ef}m$. The similarity between gene flow and mutation can also be framed in terms of a
coalescent process. For example, we can determine the conditional probability that two genes
randomly drawn from the same subpopulation coalesce back to a common ancestor before either lineage experienced a gene flow event given than either coalescence or gene flow has occurred. In analogy to equation 5.13,

\[
\text{Prob.}(\text{gene flow before coalescence} \mid \text{gene flow or coalescence}) = \frac{4N_{ef}m}{4N_{ef}m + 1} \quad 6.11
\]

Since the probability of identity in this model is the probability of coalescence before gene flow given that either gene flow or coalescence has occurred, the equilibrium probability of identity in the gene flow coalescent model is simply one minus equation 6.11, which yields equation 6.10. Whether we look backward or forward in time, we obtain the same equilibrium balance of gene flow (proportional to \(m\)) to drift (proportional to \(1/N_{ef}\)) as measured proportional to their ratio \([m/(1/N_{ef})=N_{ef}m]\).

Note that our concept of “identity” has been altered once again from what it was in Chapter 4. Wright (1931), who first derived equation 6.10, defined the “\(F\)” in equation 6.10 as “\(F_{st}\)” where the “\(st\)” designates this as identity–by–descent in the subpopulation relative to the total population. Therefore, we have yet another “inbreeding coefficient” in the population genetic literature that is distinct mathematically and biologically from the “inbreeding coefficients” previously used in this book. In particular, \(F_{st}\) does not measure identity-by-descent in the pedigree inbreeding sense (\(F\)), nor system of mating inbreeding (\(f\)), nor the impact of genetic drift within a single deme upon average identity-by-descent inbreeding (\(F\)); rather, the “inbreeding coefficient” \(F_{st}\) measures the ratio of gene flow to drift and how this ratio influences population structure in a process–oriented sense. In terms of patterns, a high value of \(F_{st}\) indicates that there is little genetic variation in a local population relative to the total population, whereas a small value indicates much local variation relative to the total. Hence, in terms of the pattern definition of population structure, \(F_{st}\) measures the apportionment of genetic variation within and among local demes. \(F_{st}\) is one of the most commonly used measures of population structure in the evolutionary genetic literature.
Equation 6.10 shows us how processes can generate patterns of population structure. As \( m \) increases (gene flow becomes more powerful), \( F_{st} \) decreases (more variation within local demes and less genetic differences between). As \( 1/N_{ef} \) increases (drift becomes more powerful), \( F_{st} \) increases (less variation within local demes and more genetic differences between). These properties are exactly as expected from the impact of genetic drift and gene flow on variation within and between local demes when considered separately. What is surprising from equation 6.10 is how little gene flow is needed to cause two populations to behave effectively as a single evolutionary lineage. For example, let \( N_{ef} \cdot m = 1 \), that is, one “effective” migrant per generation. \( N_{ef} \cdot m \) is not the actual number of migrating individuals per generation but rather is an effective number of migrants because it depends upon the local inbreeding effective size \( N_{ef} \) and not the census number of individuals.

Then, \( F_{st} = 1/5 = 0.20 \), as shown in Figure 6.3. This means that 80% of the gene pairs drawn from the same subpopulation will show gene flow before coalescence. Hence, the genealogical histories of the local demes are extensively intertwined when \( N_{ef} \cdot m \geq 1 \). Note also that \( N_{ef} \cdot m = 1 \) defines an inflection point in the plot of equation 6.10 against the effective number of migrants (Figure 6.3). \( F_{st} \) declines only very slowly with increasing effective number of migrants when \( N_{ef} \cdot m \geq 1 \), but \( F_{st} \) rises very rapidly with decreasing effective number of migrants when \( N_{ef} \cdot m \leq 1 \). Because \( N_{ef} \cdot m = 1 \) is at the inflection point of equation 6.9, an effective number of migrants of one marks a biologically significant transition in the relative evolutionary importance of gene flow to drift. It is impressive that very few effective migrants are needed (only one or more per generation on average) to cause gene flow to dominate over genetic drift, leading to subpopulations that display great genetic homogeneity with one another.

It is also surprising that the extent of this genealogical mixing depends only upon the effective number of migrants \( (N_{ef} \cdot m) \) and not upon the rate of gene flow \( (m) \). For example, two subpopulations of a billion each would share 80% of their genes by exchanging only 1 effective individual per generation from equation 6.10, but so would two subpopulations of size 100. However, the rates of gene flow would be greatly different in these two cases: \( m = 0.000000001 \) for the first case and \( m = 0.01 \) for the second. Thus, very different rates of gene flow can have
similar impacts upon population structure. Alternatively, identical rates of gene flow can have very different impacts on population structure. Suppose, for example, that \( m = 0.01 \) in both the cases considered above. For the two subpopulations with inbreeding effective sizes of one billion, the resulting \( F_{st} \) is effectively zero \((2.5 \times 10^{-8})\) from equation 6.10, whereas for the local populations of inbreeding effective size 100, \( F_{st} = 0.20 \). The reason why the same number of effective migrants is needed to yield a specific value of \( F_{st} \) and not the same rate of gene flow is that \( F_{st} \) represents a balance between the rate at which genetic drift causes subpopulations to diverge versus the rate at which gene flow makes them more similar. In large populations, divergence is slow, so small amounts of gene flow are effective in counterbalancing drift-induced divergence; as populations because smaller, larger and larger rates of gene flow are needed to counterbalance the increasing rate of drift-induced divergence.

Similarly, it is the product \( N_{ef}m \) (the ratio of gene flow to drift) and not \( m \) alone that determines the relative coalescence times of genes within and among local populations. If there is restricted gene flow among demes, it makes sense that the average time to coalescence for two genes sampled within a deme will be less than that for two genes sampled at random for the entire species. In particular, Slatkin (1991) has shown that these relative times are determined by \( N_{ef}m \). The exact relationship of coalescence times within and among local populations depends upon the pattern of gene flow. Consider the simple case of a species subdivided into a large number of local demes each of size \( N_{ef} \) and each receiving a fraction \( m \) of its genes per generation from the species at large. This “island model” of gene flow among multiple local demes of identical inbreeding effective size (Figure 6.4) also leads to equation 6.10, and hence is a straightforward multi-deme extension of the two deme model illustrated in Figure 6.1. Slatkin (1991) has shown that

\[
N_{ef}m = \frac{t_0}{4(t_t t_0)} \tag{6.12}
\]

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
where \( \tilde{t}_0 \) is the average time to coalescence of two genes sampled from the same subpopulation and \( \tilde{t} \) is the average time to coalescence of two genes sampled from the entire species. Hence, the ratio of within deme coalescence time to entire species coalescence time is:

\[
\frac{\tilde{t}_0}{\tilde{t}} = \frac{4N_e m}{1 + 4N_e m} \tag{6.13}
\]

Therefore, \( F_{st} \) now has a simple interpretation in terms of coalescence times:

\[
F_{st} = \frac{\tilde{t} \cdot \tilde{t}_0}{\tilde{t}} \quad \text{or} \quad \frac{\tilde{t}_0}{\tilde{t}} = 1 - F_{st} \tag{6.14}
\]

For example, \( F_{st} = 0.156 \) when averaged over 109 loci for local populations of humans scattered throughout the world (Barbujani et al., 1997). From equation 6.10, this \( F_{st} \) value yields 1.4 effective migrants per generation. Thus, although there is seemingly little gene flow among human populations at a global scale in an absolute sense (a little over one effective migrant on average per generation among human subpopulations), this is sufficient to place the human species in the domain where gene flow dominates over drift, resulting in human populations at the global level showing little genetic differentiation by the standard criterion used in population genetics (Figure 6.3). An \( F_{st} \) value of 0.156 also means that \((1-F_{st}) \cdot 100\% = 84.4\%\) of the time even when your parents are both from the same human subpopulation, the gene you inherited from your mother and the gene you inherited from your father trace back to different human subpopulations (from equations 6.13 and 6.14). Therefore, human subpopulations even at the global level are extensively intertwined genetically. However, this interpretation is made under the assumption that the human \( F_{st} \) value is actually due to the balance of gene flow and drift and not some other factors. This assumption will be examined in detail in Chapter 7.
Equation 6.9 can be extended to include mutation. If we assume that “identity” can be destroyed by both mutation and gene flow, then the appropriate analogue to equation 6.8 is:

\[
F(t) = \frac{1}{2N_{ef}} + \frac{1}{2N_{ef}} \overline{F(t-1)} \left[ (1 - m)(1 - m) \right]^2
\]

6.15

If both \( \mu \) and \( m \) are small, then using a Taylor’s series we have

\[
\overline{F_{eq}} = \overline{F_{st}} = \frac{1}{4N_{ef}(1 + m) + 1}
\]

6.16

Equation 6.16 shows that the joint impact of mutation and gene flow is described by the sum of \( m \) and \( m \), once again emphasizing the similar role that the disparate forces of mutation and gene flow have upon genetic variation and identity–by–descent. Equation 6.16 also shows that when \( m \) is much larger than \( m \) (frequently a realistic assumption), then gene flow dominates over mutation in interacting with genetic drift to determine population structure.

In all of the above equations, \( F_{st} \) was measured in terms of identity–by–descent, but in many cases all we can really observe is identity–by–state. In practice, \( F_{st} \) is actually measured by the proportional increase in identity–by–state that occurs when sampling within versus between subpopulations (Cockerham and Weir, 1987). Let \( F_s \) be the probability of identity–by–state of two genes randomly sampled within a deme, and let \( F_t \) be the probability of identity–by–state of two genes randomly sampled from the total species. If all subpopulations had identical gene pools, then \( F_s = F_t \). But with population subdivision, we expect an increase in identity–by–state for genes sampled within the same subpopulation beyond the random background value of \( F_t \). Note that \( 1 - F_t \) is the probability that two genes are not identical–by–state due to random background sampling of the total population. If \( F_{st} \) is now regarded as the additional probability of identity–by–state that occurs beyond random background sampling, we have:

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
Equation 6.16 provides another way of estimating $F_{st}$ by randomly sampling pairs of genes drawn either from the the total population ($F_t$) or a subpopulation ($F_s$) (Davis et al. 1990).

Equation 6.16 is also used to extend the models so far discussed to multiple hierarchies. All our models of gene flow so far have assumed a total population subdivided into a series of local demes or subpopulations, which are then all treated equally with respect to gene flow. However, this is often an unrealistic model. For example, most gene flow among human subpopulations occurs between subpopulations living on the same continent (Cavalli-Sforza et al. 1994). As a result, instead of subdividing the global human population simply into local demes, it is more biologically realistic to subdivide the global human population first into continental subpopulations, and then subdivide each continental subpopulation into local intracontinental populations. To deal with this hierarchy of three levels, we need three levels of sampling. Therefore, let $F_i$ be the probability of identity-by-state of two genes randomly sampled from the total human species, $F_c$ be the probability of identity-by-state of two genes randomly sampled from humans living on the same continent, and $F_s$ be the probability of identity-by-state of two genes randomly sampled from the same intracontinental subpopulation. Then, in analogy to equation 6.17, we have:

\[
F_{ct} = \frac{F_c \cdot F_t}{1 - F_t} \quad 6.17
\]

\[
F_{st} = \frac{F_t \cdot F_s}{1 - F_s}
\]

\[
F_{sc} = \frac{F_s \cdot F_c}{1 - F_c}
\]

where $F_{ct}$ measures the increase in identity-by-descent due to sampling within continental subpopulations relative to the total human species, and $F_{sc}$ measures the increase in identity-by-
descent due to sampling within intracontinental local populations relative to the continental subpopulations. $F_{st}$ can be recovered in this three-hierarchy model from (Wright 1969):

\[
(1 - F_{sc})(1 - F_{ct}) = 1 - F_{st}
\]

Equation 6.19 indicates the total $F_{st}$ can be partitioned into two components: $F_{ct}$ that measures genetic differentiation among continental subpopulations of humans, and $F_{sc}$ that measures genetic differentiation among local populations living in the same continent. Both contribute to $F_{st}$ as shown by equation 6.19. In particular, for humans, the $F_{st}$ of 0.156 can be partitioned into a component of 0.047 that measures the relative proportion of genetic variation among local populations within continental groups and 0.108 among continental groups (Barbujani et al. 1997).

Also note that equation 6.19 defines a chain rule that allows $F$ statistics for measuring population structure to be extended to an arbitrary number of levels (Wright 1969).

**The Wahlund Effect and F Statistics**

We have just seen how yet another “inbreeding coefficient” based upon the concept of identity–by–descent enters into the population genetic literature; but this time as a measure of how the balance of drift and gene flow influences identity–by–descent and coalescent times within and between demes in a subdivided species. We also saw in Chapter 4 that genetic drift influences many genetic parameters besides identity–by–descent, including the variance of allele frequencies across isolated replicate demes. This aspect of drift motivates an alternative definition of $F_{st}$ in terms of variances of allele frequencies across the local demes (Cockerham and Weir 1987).

Consider first a 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:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>( p_i^2 )</td>
</tr>
<tr>
<td>Aa</td>
<td>( 2pq_i )</td>
</tr>
<tr>
<td>aa</td>
<td>( q_i^2 )</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>( p^2 )</td>
</tr>
<tr>
<td>Aa</td>
<td>( 2pq )</td>
</tr>
<tr>
<td>aa</td>
<td>( q^2 )</td>
</tr>
</tbody>
</table>

However, in the general case where the demes can have different allele frequencies, the actual genotype frequencies in the total population are:

\[
\text{Freq.(AA)} = \prod_{i=1}^{n} w_i p_i^2 \\
\text{Freq.(Aa)} = 2 \prod_{i=1}^{n} w_i p_i q_i \\
\text{Freq.(aa)} = \prod_{i=1}^{n} w_i q_i^2
\]

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

\[
\text{Var}(p) = \left[ \prod_{i=1}^{n} w_i (p_i - p)^2 \right] = \prod_{i=1}^{n} w_i p_i^2 \prod_{i=1}^{n} p^2 = \prod_{i=1}^{n} w_i q_i^2 \prod_{i=1}^{n} q^2
\]

Substituting 6.22 into 6.21, the genotype frequencies in the total population can be expressed as
Freq.\((AA)\) = \(\sum_{i=1}^{n} w_i p_i^2 \equiv p^2 + \bar{p}^2 \)

Freq.\((aa)\) = \(\sum_{i=1}^{n} w_i q_i^2 \equiv q^2 + \bar{q}^2 \)  \hspace{1cm} 6.23

Freq.\((Aa)\) = 1 \(\Bigg[\)Freq.\((AA)\) \(\Bigg[\)Freq.\((aa)\) \(\Bigg] = 2 \bar{p} \bar{q} \equiv 2 \bar{p}^2 \)

By factoring out the term \(2p\bar{q}\) from the heterozygote frequency in equation 6.23, the observed frequency of heterozygotes in the total population can be expressed as

Freq.\((Aa)\) = \(2 \bar{p} \bar{q} \bigg(1 \equiv \frac{\bar{p}^2}{\bar{q}^2}\bigg) = 2 \bar{p} \bar{q} \bigg(1 \equiv f_{st}\bigg) \)  \hspace{1cm} 6.24

where \(f_{st} = \equiv \frac{\bar{p}^2}{\bar{q}^2}\). Hence, the genotype frequencies can now be expressed as:

Freq.\((AA)\) = \(\bar{p}^2 + \bar{p} \bar{q} f_{st}\)

Freq.\((Aa)\) = \(2 \bar{p} \bar{q} \bigg(1 \equiv f_{st}\bigg)\)  \hspace{1cm} 6.25

Freq.\((aa)\) = \(\bar{q}^2 + \bar{p} \bar{q} f_{st}\)

Note the resemblance between equation 6.25 and equation 3.1 from Chapter 3. Equation 3.1 describes the deviations from Hardy-Weinberg genotype frequencies induced by system of mating inbreeding \((f)\). Because a variance can only be positive, \(f_{st} \geq 0\), which implies that the subdivision of the population into genetically distinct demes causes deviations from Hardy-Weinberg that are identical in form to those caused by system-of-mating inbreeding within demes \((f < 0)\). This “inbreeding coefficient” is called \(f_{st}\), because it refers to the deviation from Hardy-Weinberg at the total population level caused by allele frequency deviations in the subdivided demes from the total population allele frequency. This deviation from Hardy-Weinberg genotype frequencies in the
species as a whole that is caused by population subdivision is called the Wahlund effect after the man who first identified this phenomenon.

The parameter $f_{st}$ is a standardized variance of allele frequencies across demes. In the extreme case where there is no gene flow at all ($m=0$), we know from Chapter 4 that drift will eventually cause all populations to either lose or fix the $A$ allele. Since drift has no direction, $\bar{p}$ of the populations will be fixed for $A$, $\bar{q}$ fixed for $a$, and the variance (equation 6.22) becomes $\bar{p}(1 - \bar{p}) + \bar{q}(0 - \bar{p}) = \bar{p}\bar{q}$. Therefore, $f_{st}$ is the ratio of the actual variance in allele frequencies across demes to the theoretical maximum when there is no gene flow at all.

From equation 6.24, an alternative expression for $f_{st}$ is given by:

$$f_{st} = 1 - \frac{\text{Freq.}(Aa)}{2\bar{p}\bar{q}} = \frac{2\bar{p}\bar{q} \cdot \text{Freq.}(Aa)}{2\bar{p}\bar{q}} = \frac{H_t - H_s}{H_t}$$

where $H_t = 2\bar{p}\bar{q}$ is the expected heterozygosity if the total population were mating at random, and $H_s$ is the observed frequency of heterozygotes in the total population, which is the same as the average heterozygosity in the subpopulations (recall that random mating is assumed within each subpopulation). The definition of $f_{st}$ given by equation 6.26 is useful in extending the concept of $f_{st}$ to the case with multiple alleles, as expected and observed heterozygosities are easily calculated or measured regardless of the number of alleles per locus. Equation 6.26 is used more commonly in the literature to measure population structure than equation 6.17, but readers need to be wary as many papers do not explicitly state which of the two definitions of $F_{st}$ is being used. This is unfortunate, because the distinction can sometimes be important.

Equation 6.10 shows that the $F_{st}$ defined in terms of probability of identity by descent can be related to the amount of gene flow, $m$, under the island model of gene flow in an equilibrium population. In the island model, a species is subdivided into a large number of local demes of equal size and with each local deme receiving a fraction $m$ of its genes per generation from the
species at large (Figure 6.4). Under this model, the variance in allele frequency across demes for an
autosomal locus with two alleles reaches an equilibrium between drift and gene flow of (Li 1955):

$$\sigma^2 = \frac{\bar{pq}}{2N_{ev} \cdot (2N_{ev} \cdot 1)(1 \cdot m)^2}$$  \hspace{1cm} \text{6.27}

Because \( f_{st} = \sigma^2 \) for a two allele system, we have

$$f_{st} = \frac{1}{2N_{ev} \cdot (2N_{ev} \cdot 1)(1 \cdot m)^2} \cdot \frac{1}{4N_{ev} \cdot m + 1}$$  \hspace{1cm} \text{6.28}

when \( m \) is small.

Note that equation 6.28 is almost identical to equation 6.10. The only difference is that in
equation 6.10 for \( F_{st} \), we have \( N_{ef} \) and in equation 6.28 for \( f_{st} \) we have \( N_{ev} \). This seemingly minor
difference between equations 6.10 and 6.28 points out that there are two qualitatively different ways
of defining \( F_{st} \):

- \( F_{st} \) is measured through identity by descent (equations 6.9, 6.10, and 6.17, and the related
coalescent equation 6.14)
- \( f_{st} \) is measured through variances of allele frequencies or heterozygosities (equations 6.25,
6.26, and 6.28).

Both \( f_{st} \) and \( F_{st} \) ideally measure the balance of genetic drift versus gene flow and in many cases are
similar in value. However, as we saw in Chapter 4, \( N_{ef} \) and \( N_{ev} \) can sometimes differ by orders of
magnitude under biologically realistic conditions. Accordingly, the two alternative definitions of \( F_{st} \)
and \( f_{st} \) can also differ substantially because they focus upon different impacts of genetic drift. In
particular, the two definitions can differ substantially when recent events have occurred that disturb
any equilibrium. The identity–by–descent definitions depend upon long-term coalescent properties

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published
by Wiley & Sons, 2003. Copyright © by Alan Templeton
(e.g., equation 6.13) and are not as sensitive to recent disturbances as are the allele frequency variance definitions, which respond rapidly to altered conditions.

For example, using an estimation procedure based upon equation 6.17 for $F_{st}$, Georgiadis et al. (1994) obtained an $F_{st}$ value not significantly different from zero for mtDNA for African elephants sampled either within eastern Africa or within southern Africa. Using the same data with an estimation procedure based upon equation 6.26, Siegismund and Arctander (1995) found significant $f_{st}$ values of 0.16 and 0.30 for eastern and southern African elephant populations respectively. This apparent discrepancy is expected given that human use of the habitat in these regions has fragmented and reduced the elephant populations over the last century in these two regions of Africa. Recall from Chapter 5 that $N_{ef}$ is generally larger than $N_{ev}$ when population size is decreasing. Hence, we expect $4N_{ef}m > 4N_{ev}m$ for these elephant populations, which means we expect $F_{st} < f_{st}$, as observed. $F_{st}$ and $f_{st}$ are not two alternative ways of estimating a population structure parameter; as the elephant example shows, they are biologically different parameters that measure different aspects of population structure in non-ideal (sensu effective size) populations.

Both $F_{st}$ and $f_{st}$ measure the relative balance of drift to gene flow, and in both cases this balance appears as a product of an effective size and $m$, the amount of gene flow. Because the balance in either equation 6.10 or 6.28 depends only upon the product $N/m$, you cannot predict the population structure of a species just by knowing either the effective sizes alone or the rate of gene flow alone. Thus, a species with large local population sizes could still show extreme genetic subdivision and a large Wahlund effect if it had low rates of gene flow, and similarly a species subdivided into small local populations could show little subdivision if it had high rates of gene flow. For example, the snail *Rumina decollata* is particularly abundant in the parks and gardens along the Boulevard des Arceaux in the city of Montpellier, France. However, like most snails, it has limited dispersal capabilities, even over this area of about 2.5 acres. Hence, despite thousands of snails living in this small area, the $f_{st}$ among 24 colonies along the Boulevard des Arceaux is 0.294 (Selander and Hudson, 1976), nearly twice as large as the $f_{st}$ value of around 0.15 for humans on a global scale. In contrast, the mouse, *Mus musculus*, tends to live in small populations within
barns in Texas, but it has much better dispersal capabilities than the snail *Ruminia*. The *f*$_{st}$ among barn subpopulations within the same farm (the “total” population in this case) is 0.021 (Selander et al., 1969), a value an order of magnitude smaller than the snail’s *f*$_{st}$ even though the geographical scales are comparable and the total population size of mice is much smaller than that of the snails. These examples show that the amount of genetic subdivision in a species depends upon the balance of drift and gene flow, and *not* upon either evolutionary factor considered by itself.

**Other Relationships of *f*$_{st}$ to *N* (Drift) and *m* (Gene Flow)**

Equations 6.10 and 6.28 relate the balance of drift and gene flow (as measured by *F*$_{st}$ or *f*$_{st}$) to underlying quantitative measures of genetic drift (*N*$_{ef}$ or *N*$_{ev}$) and gene flow (*m*). However, we derived equations 6.10 and 6.28 only for specific models of gene flow: either symmetrical gene flow between two demes, or the island model of many demes. The two–deme model is of limited generality, and the island model depends upon several specific and biologically implausible assumptions, such as a portion *m* of the gametes being extracted from each deme and distributed at random over all other demes regardless of their locations relative to one another. The island model was chosen primarily for its mathematical convenience rather than its biological realism. Changing the underlying assumptions of the model can change the balance between drift and gene flow. Consequently, equations 6.10 and 6.28 do *not* represent the general quantitative relationship between drift and gene flow as forces causing genetic subdivision. Instead, these equations represent only special and highly unrealistic cases. We now consider some alternate models.

One aspect of the island model of gene flow and drift that is unrealistic for many species is the assumption that all dispersing individuals (or gametes) are equally likely to end up in any local deme. In most real species, some pairs of local demes experience much more genetic interchange than others. One common type of deviation from the island model is *isolation by distance* in which local demes living nearby to one another interchange gametes more frequently than do geographically distant demes. We need look no further than our own species to see isolation–by–distance. For example, of 2022 marriages recorded in the upper Ina Valley of Japan from “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
(Seikiguchi and Sekiguchi 1951), nearly half occurred between people from the same buraku
(hamlet), over two thirds from the same village (including the burakus within the village), and less
than a third from more remote locations (Table 6.1). As geographical distance between spouses
increases, the percentage of marriages decreases; exactly the pattern expected under
isolation–by–distance.

There are many models of isolation–by–distance in the population genetic literature, and we
will only consider a few simple ones. We start with a one-dimensional stepping-stone model in
which a species is subdivided into discrete local demes, as with the island model. These local
demes are arrayed along a one dimensional habitat such as a river, a valley, or shore-line. One
version of this model allows two types of gene flow (Figure 6.5). First, a fraction \( m \) of the
gametes leave each deme and disperse at random over the entire species, just as in the island model.
Second, a fraction \( m_1 \) of the gametes from each deme disperse only to the adjacent demes. Because
this is a one-dimensional model, each deme has just two neighbors (Figure 6.5, ignoring the demes
at the two ends of the habitat), and we assume symmetrical gene flow at this local geographic level;
that is \( m_1/2 \) go to one of the neighboring demes, and the other \( m_1/2 \) go to the other (Figure 6.5).
Then (Weiss and Kimura, 1965):

\[
\hat{f}_{st} = \frac{1}{2N_{ev}} \frac{\frac{1}{2N_{ev}}}{2R_1R_2 \left( \frac{R_1}{R_1 + R_2} \right)}
\]

where the correlations between allele frequencies of demes one \((R_1)\) and two \((R_2)\) steps apart are:

\[
R_1 = \sqrt{\left[ 1 + (1 \square m_1)(1 \square m) \right]^2 \square \left[ m_1(1 \square m) \right]^2}
\]

\[
R_1 = \sqrt{\left[ 1 \square (1 \square m_1)(1 \square m) \right]^2 \square \left[ m_1(1 \square m) \right]^2}
\]
When $m_1 = 0$, all dispersal is at random over the entire species’ geographical distribution (the island model) and equation 6.32 simplifies to $f_{st} = 1/[1 + 4N_e m_\infty]$. Hence, equation 6.28 is a special case of 6.32 when there is no additional gene flow between adjacent demes.

In most real populations (for example, Table 6.1) there is much more dispersal between neighboring demes than between distant demes. In terms of our stepping stone model, this means that $m_\infty$ is very small relative to $m_1$, and under these conditions equation 6.32 is approximately:

$$f_{st} = \frac{1}{1 + 4N_e \sqrt{2m_1 m_\infty}}$$  \hspace{1cm} 6.34

Note that from equation 6.34 that even when $m_\infty$ is very small relative to $m_1$, the long-distance dispersal parameter $m_\infty$ still has a major impact on genetic subdivision because the impact of gene flow is given by the product of $m_1$ and $m_\infty$. Consequently, equation 6.34 is extremely sensitive to long-distance dispersal even when such dispersal is very rare compared to short distance dispersal.

For example, let $N_e = 100$ and $m_1 = 0.1$. Then $f_{st} = 0.053$ if $m_\infty = 0.01$, and $f_{st} = 0.276$ if $m_\infty = 0.001$. Note in this example that large differences in $f_{st}$ are invoked by changes in long-distance dispersal even though long-distance dispersal is ten to a hundred times less common than short-distance dispersal. At first this sensitivity to rare, long-distance dispersal may seem counterintuitive, but the reason for it can be found in equations 6.2. From those equations, we saw that the evolutionary impact of gene flow upon allele frequencies depends upon two factors: 1) how much interchange is actually occurring between two demes and 2) how genetically distinct the demes are in their allele frequencies. When much dispersal occurs between neighboring demes, the allele frequencies in those neighboring demes are going to be very similar. Hence, even though there is a large amount of exchange between neighbors relative to long-distance dispersal, exchange between neighboring demes has only a minor evolutionary impact upon allele frequencies. On the other hand, when long-distance dispersal occurs, it generally brings in gametes to the local deme that come from a distant deme with very different allele frequencies. As a consequence, long-distance dispersal has a major impact on genetic subdivision.
dispersal has a large evolutionary impact when it occurs. This trade-off between frequency of genetic interchange and the magnitude of the evolutionary impact given genetic interchange explains why both \( m_1 \) and \( m_\infty \) contribute in a symmetrical fashion to \( f_{st} \).

The importance of long-distance dispersal (\( m_\infty \)) upon overall gene flow even when it is rare relative to short-distance dispersal means that gene flow is difficult to measure accurately from dispersal data. Many methods of monitoring dispersal directly are limited to short geographical scales, and it is usually impossible to quantify the amount of long-distance dispersal. Yet these rare, long-distance dispersal events have a major impact on a species’ genetic population structure.

Another interesting implication of the isolation–by–distance models is that the degree of genetic differentiation between two demes or two points on a geographical continuum should increase with increasing separation — either the number of “steps” (in the stepping stone model) or the geographical distance (in the continuous distribution models). *The degree of genetic differentiation between two populations is measured by a population genetic distance.* There are several types of population genetic distances, and Box 6.1 shows Nei’s population genetic distance, one of the more commonly used measures. [Box 6.1 about here.] However, the distance measure most convenient for isolation–by–distance models is the pairwise \( f_{st} \). A pairwise \( f_{st} \) is an \( f_{st} \) calculated from equation 6.26 and applied to just two populations at a time. The total population now used to calculate \( H_t \) refers just to the two populations of interest and all other populations in the species are ignored. Note that this population genetic distance (and all others as well) is biologically quite distinct from the molecule genetic distances discussed in Chapter 5. A molecule genetic distance ideally measures the number of mutations that occurred between two DNA molecules during their evolution from a common ancestral molecule. A pairwise \( f_{st} \) is a function of allele frequencies between two demes, and mutation is not even necessary for this population genetic distance to take on its maximum value. For example, suppose an ancestral population were polymorphic for two alleles, \( A \) and \( a \), at a locus. Now assume that the ancestral population split into two isolates, with one isolate becoming fixed for \( A \) and the other for \( a \). Then,
the pairwise $f_{st}$ for these two isolates would be 1 (and the Nei’s distance in Box 6.1 would be infinite), the maximum value possible, even though not a single mutation occurred. Population genetic distances should never be confused with molecule genetic distances.

In the isolation–by–distance models, let $f_{st}(x)$ be the pairwise $f_{st}$ between two populations $x$ steps apart or $x$ geographical units apart. Malécot (1950) has shown that:

$$f_{st}(x) = \frac{e^{\frac{x^2}{4mN_{ev}}} / m_1}{1 + 4N_{ev} \sqrt{2mN_{ev}} \cdot m \cdot m_1} \quad \text{when} \quad m \ll m_1 \quad 6.38$$

for the discrete, one-dimensional stepping stone model.

In general, the isolation–by–distance models can be approximated by an equation of form (Malécot 1950):

$$f_{st} = ae^{\left(\frac{b}{c}\right)x} \quad 6.40$$

where $c = \frac{1}{2}$ (dimensionality of the habitat -1), and where $a$ and $b$ are estimated from the $f_{st}(x)$ data. The parameter $c$ can also be estimated from the observed pairwise population genetic distances when it is not obvious what the dimensionality of the habitat may be. For example, a species may be distributed over a two-dimensional habitat, but more constraints on movement may occur in one direction than another. Hence, non-integer dimensions between one and two are biologically meaningful.

Equation 6.40 has been applied to human data. Figure 6.7 (redrawn from Cavalli-Sforza et al. 1994) shows the human pairwise $f_{st}$ on a global scale along with the isolation–by–distance curve fitted from equation 6.40. As can be seen, the patterns of genetic differentiation between human populations fit well to the isolation–by–distance model. This good fit is not unexpected. Table 6.1 shows that humans in the Ina Valley in Japan display an isolation–by–distance pattern, and other
studies have shown that isolation–by–distance holds for human gene flow on many different geographical scales and virtually all studied populations (Lasker and Crews 1996; Santos et al 1997).

Using Sequence or Restriction Site Data to Measure Population Subdivision

In our previous models of population subdivision, any pair of homologous genes was classified into one of two mutually exclusive categories: the pair was identical–by–descent or not (for defining \( F_{st} \)); or the pair was heterozygous or not (for defining \( f_{st} \)). However, with the advent of restriction site and DNA sequence data, we can refine this categorization by the use of a molecule genetic distance. Suppose, for example, that we have sequence data on a 10 kb locus, and a pair of genes at this locus is examined that differs by only a single nucleotide site. Now consider another pair of genes, but this time differing by 20 nucleotide sites. In the models used previously in this chapter, both of these pairs of genes would be placed into the same category: they are not identical–by–descent at this locus, or they are heterozygous at this locus. However, the second pair is much less identical than the first pair; or alternatively, the first pair has much less nucleotide heterozygosity than the second pair. These quantitative differences in non-identity or amount of heterozygosity are ignored in all of our previous formulations. However, taking into account these quantitative differences can increase power and sensitivity for detecting population subdivision and restricted gene flow (Hudson et al. 1992). Accordingly, there have been several proposed alternatives to \( f_{st} \) that incorporate the amount of difference at the molecular level between heterozygous pairs of alleles (that is, a molecule genetic distance). Among these are \( N_{st} \), which quantifies the amount of molecular heterozygosity by the average number of differences between sequences from different localities (Lynch and Crease, 1990), and \( K_{st} \), which uses the average number of differences between sequences randomly drawn from all localities (Hudson et al. 1992). In general, we will let \( \Delta_{st} \) designate any \( f_{st} \)-like statistic that uses a molecule genetic distance instead of heterozygosity as its underlying measure of genetic diversity (Excoffier et al., 1992).
Recall that we can define multiple $f$ statistics to partition genetic variation at different biological levels in a hierarchy of individuals and populations. For example, recall that by using the relationship $(1-f_{st}) = (1-f_{is})(1-f_{it})$, we could partition genetic diversity as measured by heterozygosity in the Yanomama Indians into a portion due to heterozygosity within local demes ($1-f_{is}$) and a portion due to heterozygosity between local demes ($1-f_{st}$). The same types of hierarchical partitions can be made using $F$ statistics instead of $f$ statistics. A partition of genetic variation that substitutes a molecule genetic distance for heterozygosity is called an Analysis of MOlecular VAriance (AMOVA).

These newer measures of population structure are essential when dealing with data sets with extremely high levels of heterozygosity, as are becoming increasingly common with DNA sequencing. When there are high levels of allelic or haplotypic heterozygosity in all populations, both $H_r$ (the expected heterozygosity if the total population were mating at random) and $H_s$ (the actual heterozygosity in the total population) are close to one in value. When all heterozygosities are high, the $f_{st}$ value calculated from equation 6.26 approaches zero regardless of the values of the underlying evolutionary parameters (such as $N_{ev}$, $m$, or $\theta$). Thus, the all or nothing nature of traditional heterozygosity measures can induce serious difficulties when dealing with highly variable genetic systems. For example, recall from Chapter 1 the survey of 9.7 kb of the human lipoprotein lipase locus ($LPL$) in which 88 haplotypes were found in a sample of 142 chromosomes coming from three human populations; a population in Rochester, Minnesota; a population of African-Americans from Jackson, Mississippi; and a population of Finns from North Karelia, Finland. Regarding each haplotype as an allele, the value of $f_{st}$ was 0.02 (Clark et al. 1998), which was not significantly different from zero. This seemingly indicates that these populations display no significant genetic differentiation. However, using the nucleotide differences at the sequence level as the molecular genetic distance, then $\theta_{st}$ was calculated to be 0.07, a value significantly different from zero. Therefore, these populations did indeed show significant genetic differentiation, but these differences were not detected by the traditional $f_{st}$ using.
haplotypes as alleles. In general, when high levels of genetic variation are encountered, a quantitative scale of differences between alleles is far preferable to a qualitative one.

**Multiple modes of inheritance and population structure**

Most of the theory presented in this chapter up to now has focused upon autosomal loci. However, we can study DNA regions with different modes of inheritance that are found in the same individuals. For example, many animals, including humans and fruit flies in the genus *Drosophila*, have DNA regions with four basic modes of inheritance:

- autosomal nuclear DNA regions, with diploid, biparental inheritance
- X-linked nuclear DNA regions, with haplo-diploid, biparental inheritance
- Y-linked nuclear DNA regions, with haploid, uniparental (paternal) inheritance
- mitochondrial DNA, with haploid, uniparental (maternal) inheritance.

These different modes of inheritance interact strongly with the balance between genetic drift and gene flow that shapes genetic population structure. In our previous models, we assumed an autosomal system with diploid, biparental inheritance. As a consequence, an effective population size (of whatever type) of \( N_e \) individuals corresponds to an effective sample of \( 2N_e \) autosomal genes. As we saw earlier, the force of drift in this case is given by \( 1/(2N_e) \). In general, the force of drift is given by \( 1/[\text{the effective number of gene copies in the population}] \). An autosomal nuclear gene is only a special case in which the effective number of gene copies is \( 2N_e \).

The effective number of gene copies is not the same when we examine other modes of inheritance. For example, in a species with a 50:50 sex ratio (such as humans or *Drosophila*), \( N_e \) individuals corresponds to an effective sample of \( \frac{3}{2}N_e \) X-linked genes. Because different modes of inheritance have different numbers of genes per individual, the force of drift is expected to vary across the different genetic systems that are coexisting in the same individuals. Variation in the force of drift directly induces variation in the *balance* of drift and gene flow. Therefore, we do not expect different DNA regions imbedded within the same populations to display the same degree of
genetic subdivision, even when the rate of gene flow is identical for all DNA regions. For example, if dealing with a haploid genetic system in a population of size $N_{ev}$, equation 6.28 becomes:

$$f_{st} = \frac{1}{1 + 2N_{ev}m} \quad 6.45$$

Note that the term “$4N_{ev}$” in equation 6.28 has been replaced by “$2N_{ev}$” in equation 6.45. This change by a factor of two is due to the fact that a diploid system in a population with variance effective size $N_{ev}$ actually is based upon an effective sample of $2N_{ev}$ genes, whereas the haploid system only has an effective sample of $N_{ev}$ genes. If we also stipulate that the haploid system has uniparental inheritance in a species with a 50:50 sex ratio, only half the individuals sampled can actually pass on the haploid genes, reducing the effective size by another factor of two (assuming the sexes have equal variance effective sizes). Thus, for haploid, uniparental systems like mitochondrial or Y-chromosomal DNA in humans or *Drosophila*, the effective sample size is actually $\frac{1}{2}N_{ev}$, so equation 6.45 becomes:

$$f_{st} = \frac{1}{1 + N_{ev}m} \quad 6.46$$

Similarly, adjusting for the number of genes per individual, the analogue of equation 6.28 for an X-linked locus in a species with a 50:50 sex ratio is

$$f_{st} = \frac{1}{1 + 3N_{ev}m} \quad 6.47$$

Hence, as we go from autosomal to X-linked to haploid uniparental systems, we expect to see increasing amounts of genetic subdivision for the same rate of gene flow ($m$). This expected

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
variation in the balance of drift and gene flow presents an opportunity to gain more insight into population structure by simultaneously studying several inheritance systems rather than just one.

For example, populations of the fruit fly *Drosophila mercatorum* were studied that live in the Kohala Mountains on the northern end of the Island of Hawaii (DeSalle et al. 1987). Samples were taken from several sites in the Kohalas (Figure 6.9), but the bulk of the collection came from site IV and the three nearby sites B, C and D. We will consider this as a two-population system by pooling the nearby sites B, C and D and contrasting them with site IV. An isozyme survey of nuclear autosomal genes yields an $f_{st}$ of 0.0002 for this contrast, which is not significantly different from zero. Given the sample sizes and using equation 6.10, this non-significant $f_{st}$ is incompatible with any $N_e m$ value less than 2. Hence, the nuclear autosomal system tells us that $N_e m > 2$. This is not surprising, as sites IV and B-C-D are less than three kilometers apart, with the intervening area consisting of inhabited area for this species. Although the nuclear $f_{st}$ tells us that there is significant gene flow between these two localities, we cannot distinguish between $N_e m$ of 3, or 30, or 300, or 3,000 with the nuclear autosomal isozyme data alone.

The same individual flies were also scored for restriction site polymorphisms for mtDNA, yielding a statistically significant $f_{st}$ of 0.17. Using equation 6.46, this significant $f_{st}$ implies $N_e m \approx 5$. This value of $N_e m \approx 5$ from the mtDNA data is consistent with the inference that $N_e m > 2$ from the isozyme data. Thus, there is no biological contradiction for these samples to yield a non-significant $f_{st}$ for isozymes and a significant $f_{st}$ for mtDNA. This population of flies is also polymorphic for a deletion of ribosomal DNA on the Y-chromosome (Hollocher et al. 1994). The Y-DNA $f_{st}$ was 0.08 and significantly different from 0. Using equation 6.46 for the Y-DNA data yields an estimate of $N_e m \approx 12$, a result also compatible with the isozyme data. Hence, the nuclear autosomal system told us very little about the possible values for $N_e m$ other than that the value is greater than two, but by combining all the data we know that $N_e m$ in this population is around 5 to 12 and not in the hundreds or thousands or more.

Although both of the haploid elements were individually consistent with the isozyme results, note that the $f_{st}$ calculated from the Y-DNA is less than half that from the mtDNA. This difference
was statistically significant even though both Y-DNA and mtDNA are haploid, uniparental systems. To understand why these two haploid, uniparental systems could yield significantly different results, recall that equation 6.46 was applied to both of them under the assumptions that the variance effective sizes of both sexes were equal and that gene flow was equal for both sexes. In most non-monogamous species, the variance in offspring number is greater in males than in females, resulting in males having a smaller variance effective size than females (see equations 4.31 in Chapter 4). This in turn, through equation 6.46, would imply that the $f_{st}$ for the Y-DNA should be larger than that for mtDNA — exactly the opposite of the observed pattern. Therefore, the typical pattern in variance of male versus female reproductive success cannot explain the observed results. We therefore turn our attention to sex-specific influences on the rate of gene flow.

Direct studies on dispersal in this species reveal identical dispersal behaviors in males and females (Johnston and Templeton 1982), but this does not mean that $m$ is identical in males and females. All dispersal in this species occurs during the adult phase, and almost all adult $D. mercatorum$ females are inseminated, often by one to three males. The females have a special organ for storing sperm and can retain viable sperm for several days. Recall that $m$ measures the rate of exchange of gametes, not individuals. Therefore, when a male disperses, only male gametes are potentially being dispersed. However, dispersing females carry not only their own gametes but those of one to three males as well. So, even with equal dispersal rates for males and females, there is actually much more gene flow of male gametes than female gametes. This male biased gene flow predicts a smaller $f_{st}$’s for Y-DNA versus mtDNA, as is observed. Hence, we need two gene flow rates, $m_f$ for female gametes, and $m_m$ for male gametes. The overall gene flow rate for an autosomal locus is just the average of the two sexes as they are in a 50:50 ratio; that is, $m = (m_f + m_m)/2$.

Assuming the variance effective sizes are the same for both sexes, the nuclear isozyme results imply that $N_{ev}m = N_{ev}(m_f + m_m)/2 > 2$; the mtDNA implies that $N_{ev}m_f = 5$; and the Y-DNA implies that $N_{ev}m_m = 12$. All systems together therefore yield $N_{ev}m = N_{ev}(m_f + m_m)/2 = (5 + 12)/2 = 8.5$. Thus, by combining the results of several genetic systems with different modes of inheritance, we can gain more insight into population structure, including sex-specific differences.

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
A Final Warning

$F_{st}$ and $f_{st}$ (and related measures such as $N_{st}$, $K_{st}$, or $\theta_{st}$) can be measured from genetic survey data. In contrast, measuring gene flow directly from dispersal studies is often difficult and unreliable. The theory developed in this chapter shows that even rare exchanges between populations can result in much effective gene flow and have major consequences for population structure. Long distance dispersal is the most difficult type of dispersal to study, but as we learned, its genetic impact can be great even when exceedingly rare. Moreover, dispersal of individuals is not the same as gene flow, as shown by the *Drosophila mercatorum* example. Therefore, gene flow, as opposed to dispersal, is more accurately measured from genetic survey data than from direct observations on the movements of plants and animals. However, estimating gene flow from $F$ or $f$ and related statistics *assumes* that the underlying cause of differentiation among demes or localities is due to the balance of genetic drift and recurrent gene flow. Unfortunately, the level of differentiation among populations can be influenced by forces other than recurrent gene flow.

For example, suppose a species is split into two large subpopulations that have no genetic interchange whatsoever (that is, $m = 0$). Equation 6.28 then tells us that $f_{st} = 1$. However, this is an equilibrium prediction, but historical events could have occurred that placed the populations far out of equilibrium or that created large temporal fluctuations in the amount of gene flow or drift. When equilibrium is disrupted, it takes time for the relationship shown in equation 6.28 to become reestablished. If the subpopulations are large, it would take many generations after the cessation of gene flow before we would actually expect to see $f_{st} = 1$. Until that equilibrium is achieved, $f_{st} < 1$ so that using the equilibrium equation in this case would incorrectly indicate that $m > 0$ when in fact $m = 0$.

On the other hand, suppose a species recently expanded its geographical range into a new area from a small and genetically homogeneous founder population. The local demes formed in the new geographical range of the species would display much genetic homogeneity for many generations because of their common ancestry regardless of what $N_{e} m$ value is established in the
newly colonized region (Larson 1984). Once again, if we sampled shortly after such a range expansion event, we would mistakenly infer high values of $m$ regardless of what the current values of $m$ were. Suppose now that an event occurred that restored gene flow among previously long–isolated populations. If $m$ were small, equation 6.48 implies that $f_{st}$ would decline only slowly, so the observation of a non-equilibrium $f_{st}$ in this case would imply less gene flow than is actually occurring at present. These hypothetical examples show that $f_{st}$ is an effective indicator of gene flow over evolutionary time only if your population is in equilibrium and has not been influenced by recent historical events. This is a big IF, so we will address how to separate current population structure from historical events in the next chapter.
Table 6.1. Isolation–by–distance in the Ina Valley of Japan as measured by the location of the spouse’s birthplace for 2022 marriages (Seikiguchi and Sekiguchi 1951).

<table>
<thead>
<tr>
<th>Spouse’s Birthplace</th>
<th>Percentage of Marriages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Buraku (hamlet)</td>
<td>49.6%</td>
</tr>
<tr>
<td>Within Village but Outside Buraku</td>
<td>19.5%</td>
</tr>
<tr>
<td>Neighboring Villages</td>
<td>19.1%</td>
</tr>
<tr>
<td>Within Gun (county)</td>
<td>6.4%</td>
</tr>
<tr>
<td>Within Prefecture (state) but Outside Gun</td>
<td>2.9%</td>
</tr>
<tr>
<td>Outside Prefecture</td>
<td>2.5%</td>
</tr>
</tbody>
</table>
Box 6.1. Nei’s (1972) Population Genetic Distance

Consider two populations, 1 and 2, scored for allelic variation at a locus. Let $p_{1i}$ and $p_{2i}$ be the frequencies of the $i$-th allele in populations 1 and 2 respectively. Then the probability of identity by state of two genes chosen at random from population 1 is $j_1 = \sum p_{1i}^2$ where the summation is taken over all alleles. Similarly, the probability of identity by state of two genes chosen at random from population 2 is $j_2 = \sum p_{2i}^2$. The probability of identity by state between two genes, one chosen at random from population 1 and one chosen at random from population 2, is $j_{12} = \sum p_{1i} p_{2i}$. Nei (1972) defined the normalized genetic identity by state between these two populations as:

$$I_{12} = \frac{j_{12}}{\sqrt{jj_{12}}}$$

Note that this measure of identity between the two populations ranges from zero (when the two populations share no alleles in common, thereby making $j_{12} = 0$) to one (when the two populations share all alleles in common and at the same allele frequencies, thereby making $j_{12} = j_1 = j_2$).

In contrast to an identity measure, a distance measure should get larger as the two populations share less and less in common in terms of alleles and their frequencies. Nei mathematically transformed $I_{12}$ into a distance measure by taking the negative of the natural logarithm of identity:

$$D_{12} = \ell n(I_{12})$$

This population genetic distance ranges from zero when the populations share all alleles in common and at the same allele frequencies ($I_{12} = 1$) to infinity when the populations share no alleles in common ($I_{12} = 0$).

When data from multiple loci exists, Nei (1972) recommended that the $j$’s be averaged over all loci, including monomorphic loci (loci with only one allele, thereby ensuring that all $j$’s are one at monomorphic loci). These average $j$’s are then used to calculate an overall identity, which is then transformed to yield an overall genetic distance. Hillis (1984) pointed out that averaging the $j$’s across loci can sometimes lead to distances that make little sense biologically. For example, suppose that two populations are fixed for the same allele at one locus (and hence all the $j$’s are 1

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
and $I_{12} = 1$), and at a second locus they share no alleles, but each population is polymorphic for two alleles each with a frequency of 0.5. At this second locus, $j_{12} = 0$ and $I_{12} = 0$ because they share no alleles, but $j_1 = j_2 = 0.5$. Hence, across both loci, the average $j_{12} = \frac{1}{2}(1+0) = 0.5$, the average $j_1 =$ the average $j_2 = \frac{1}{2}(1+0.5) = 0.75$. Using these average values of the $j$’s, $I_{12}$ is calculated to be $0.5/0.75 = 0.667$ and $D_{12} = 0.41$.

Now consider another case in which the first locus is polymorphic with both populations sharing two alleles and with each allele in each population having a frequency of 0.5. In this case, all the $j$’s are 0.5 in value, and $I_{12}$ for this locus has a value of 1. At the second locus in this case, each population is fixed for a different allele, so $j_{12} = 0$ and $j_1 = j_2 = 0.5$. In this second case, the average $j_{12} = \frac{1}{2}(0.5+0) = 0.25$ and the average $j_1$ and average $j_2$ are both $\frac{1}{2}(0.5+1) = 0.75$. Using these average $j$’s, $I_{12} = 0.25/0.75 = 0.333$ and $D_{12} = 1.1$. In this case and in the case given in the previous paragraph, the two populations are both completely identical at the first locus and completely different at the second, yet they have very different genetic distances (0.41 versus 1.1) as originally defined by Nei (1972). Hillis (1984) points out that situations like this are likely to occur when different loci have different overall rates of evolution. To make the population genetic distance measure robust to this heterogeneity in evolutionary rates across loci, Hillis (1984) recommends that the $I$’s be averaged across loci, not the $j$’s. For example, in the first case where the populations are identical at the monomorphic locus and different at the polymorphic one, the average $I_{12} = \frac{1}{2}(1+0) = 0.5$, yielding a population genetic distance of 0.69. In the second case where the populations are identical at the polymorphic locus and different at the monomorphic one, the average $I_{12} = \frac{1}{2}(1+0) = 0.5$, yielding the same population genetic distance of 0.69. Therefore, it is better to average the $I$’s and not the $j$’s in calculating this type of population genetic distance from multi-locus data when there is rate heterogeneity across loci.
Figure 6.3. A plot of $F_{st}$ in equation 6.9 versus the effective number of migrants per generation, $N_{ef}m$. The plot shows the inflection point at $N_{ef}m = 1$, corresponding to $F_{st} = 0.2$.

Figure 6.4. The island model of gene flow among multiple local demes, each of idealized size $N$, each contributing a fraction $m$ of its gametes to a common gene pool which is then distributed at random over all local demes in the same proportion.

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
Figure 6.5. The one-dimensional stepping stone model of gene flow between discrete demes. Each
deme is of idealized size $N$ and is represented as a circle arrayed on a line. A portion $m_i$ of the
gametes from any one population are exchanged with the two neighboring populations, half going
to each neighbor. Moreover, each population contributes a fraction $m\circ$ of its gametes to a common
gene pool which is then distributed at random over all demes in the same proportion.
Figure 6.7. Isolation by distance in human populations on a global scale. The x-axis gives the geographical distance in miles between pairs of human populations. The y-axis gives the pairwise $f_{st}(x)$ where $x$ is the geographical distance in miles. The solid line gives the predicted curve under a model of isolation–by–distance.

FROM “Population Genetics and Microevolutionary Theory” by A. Templeton. To be published by Wiley & Sons, 2003. Copyright © by Alan Templeton
Figure 6.9. A topographical map of collecting sites for *Drosophila mercatorum* in the Kohala Mountains near the town of Kamuela on the Island of Hawaii. A transect of collecting sites on the slopes of the Kohalas are indicated by the letters A, B, C, D and F, and a site in the saddle at the base of the Kohalas is indicated by IV.