Gene Flow

- The Hardy-Weinberg Model Considered Only A Single Population
- Most Species Have Many Local Populations
- Although Most Matings Usually Occur Within A Local Population, Sometimes Individuals Mate Outside Their Local Deme
- Reproductive Connections Between Demes Allow Genes To Move From One Local Gene Pool To Another.

- The Movement of Genes Between Demes Is Called Gene Flow.

Gene Flow Between Two Demes
Gene Flow Between Two Demes

Gene Flow Between Two Demes

\[ p_1' = (1-m)p_1 + mp_2 = p_1 - m(p_1-p_2) \]
\[ p_2' = (1-m)p_2 + mp_1 = p_2 + m(p_1-p_2) \]

Let \( d_0 = (p_1-p_2) \) = the difference in allele frequencies between the demes at generation 0.

Then:
\[ p_1' = p_1 - md_0 \]
\[ p_2' = p_2 + md_0 \]

**Gene Flow (m > 0) Will Be An Evolutionary Force When \( d_0 \neq 0 \);
That is, When The Initial Gene Pools Are Different (\( p_1 \neq p_2 \)).**
Gene Flow Between Two Demes

- $m$ is defined in terms of the gene pools, and therefore $m$ represents the amount of exchange of gametes between the local populations and not necessarily individuals.
- In some species, gametes are exchanged directly without the diploid individuals moving at all. For example, most trees are wind pollinated
- Because gene flow requires both movement and reproduction, $m$ represents a complex interaction between the pattern of dispersal and the system of mating.
- In general, disassortative mating enhances gene flow (recall Yanomama & Makiritare) and inbreeding & assortative mating reduce gene flow for a given amount of dispersal of individuals.

The impact of system of mating on patterns and amount of gene flow is illustrated by the admixture of Europeans and Africans in the Americas.
Gene Flow Between Two Demes

Ancestral Gene Pools

<table>
<thead>
<tr>
<th></th>
<th>European Population</th>
<th>West African Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( p_E )</td>
<td>( p_W )</td>
</tr>
<tr>
<td>( A )</td>
<td>( a )</td>
<td>( a )</td>
</tr>
<tr>
<td>( q_E )</td>
<td>( q_W )</td>
<td>( q_W )</td>
</tr>
</tbody>
</table>

Gene Pools in Present North America

<table>
<thead>
<tr>
<th></th>
<th>European Americans</th>
<th>African Americans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( p_E )</td>
<td>( p_W )</td>
</tr>
<tr>
<td>( A )</td>
<td>( a )</td>
<td>( a )</td>
</tr>
<tr>
<td>( q_E )</td>
<td>( q_W )</td>
<td>( q_W )</td>
</tr>
</tbody>
</table>

Where \( M \) is the effective amount of gene flow since admixture began (in contrast to \( m \), a per generation gene flow parameter)

Gene Flow Between Two Demes

\[
M = \frac{p_A - p_W}{p_E - p_W} = \frac{\text{Change in Allele Freq. in African Americans from West Africans}}{\text{Initial Diff. in Allele Freq. Between Europeans and West Africans}}
\]

For Example, The Frequencies of the Rh\(^+\) Allele at the Rh Blood Group Locus Are:

- Western Europeans: 0.0279
- Western Africans: 0.5512
- African-Americans: 0.4381

So, \( M = \frac{(0.4381 - 0.5512)}{(0.0279 - 0.5512)} = 0.216 \).
Gene Flow Between Two Demes

In Northeastern Brazil, there were many categories for individuals of mixed ancestry, and because of these social definitions, there was much more gene flow between both parental populations.

Gene Flow Between Two Demes

E.g., self-identified “whites” in NE Brazil:

- 67% European
- 20% African
- 13% Native American

E.g., self-identified “non-whites” in NE Brazil:

- 58% European
- 25% African
- 17% Native American
The Evolutionary Impacts of Gene Flow

Gene Flow Between Two Demes

Gene Pools At Generation 0
Local Population 1
- \( A \)
- \( p_1 \)
- \( q_1 \)

Gene Pools At Generation 1
Local Population 1
- \( A \)
- \( p_1' = p_1 \cdot m \)
- \( q_1' \)

Gene Pools At Generation 2
Local Population 1
- \( A \)
- \( p_1'' = p_1' \cdot m \)
- \( q_1'' \)

Gene Pools At Generation 0
Local Population 2
- \( A \)
- \( p_2 \)
- \( q_2 \)

Gene Pools At Generation 1
Local Population 2
- \( A \)
- \( p_2' = p_2 + m \)
- \( q_2' \)

Gene Pools At Generation 2
Local Population 2
- \( A \)
- \( p_2'' = p_2' + m \)
- \( q_2'' \)
Gene Flow Between Two Demes

After 1 Generation of Gene Flow:
\[ p_1' = p_1 - md_o \quad \quad p_2' = p_2 + md_o \]

\[ d_1 = p_1' - p_2' = p_1 - md_o - p_2 - md_o \]
\[ = p_1 - p_2 - 2md_o \]
\[ = d_o - 2md_o \]
\[ = d_o(1-2m) \]

Can show that after n generations of gene flow:
\[ d_n = d_o(1-2m)^n \Rightarrow 0 \text{ as } n \Rightarrow \infty \]

**Gene Flow Is An Evolutionary Force That Reduces Genetic Differences Among Local Demes**

Gene Flow Between Two Demes

**Gene Flow Is An Evolutionary Force That Reduces Genetic Differences Among Local Demes**

For Example, The Frequencies of the Rh⁺ Allele at the Rh Blood Group Locus Are:

- Western Europeans: 0.0279
- Western Africans: 0.5512
- African-Americans: 0.4381
Recall that genetic drift increases genetic variation between demes.

4 isolated demes started from one ancestral deme with p = 0.5.

Gene flow between two demes:

Gene flow is an evolutionary force that increases genetic variation within local populations.
Recall that genetic drift decreases genetic variation within demes. Gene flow and genetic drift have opposite effects on genetic variation within and between demes. The balance between these two evolutionary forces is the major determinant of the relative amounts of genetic variation within versus between local populations of a species.
Measuring the Balance of Gene Flow and Genetic Drift

Sewall Wright
Wright quantified the balance of gene flow to drift as measured by $F_{st}$ for the Island Model.

**Impact of Drift and Gene Flow On Average $F$ In The Island Model**

Effect of Drift Alone On The Prob. Two Randomly Chosen Genes are I.B.D:

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

With gene flow, two genes can only be I.B.D. If they are from the same deme (by assumption), so:

$$F(t) = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F(t-1)\right](1-m)^2$$

At equilibrium, $F(t) = F(t-1) = F$, and the above equation yields:

$$F = F_{st} \approx \frac{1}{4Nn+1}$$
Impact of Drift and Gene Flow On Average F In The Island Model

$$F_{st} \approx \frac{1}{4Nm+1}$$

- $F_{st}$ does not measure ibd in the usual pedigree sense, but rather the ratio of drift to gene flow at the population (multiple deme) level.
- This is yet another “inbreeding coefficient”, but this one measures the proportion of genetic variation among individuals drawn from all demes that is due to genetic differences between demes.

Impact of Drift and Gene Flow On Average F In The Island Model

- In many cases cannot distinguish ibd from identity by state.
- Let $F_s$ be the probability of identity by state of two genes drawn from within a subpopulation (local deme).
- Let $F_t$ be the probability of identity by state of two genes drawn from the total population.
- Now, $F_{st}$ is additional probability of identity-by-state within a deme that occurs because of between deme differentiation:
  $\quad F_s = F_t + (1 - F_t) F_{st}$
  $\quad F_{st} = \frac{F_s - F_t}{1 - F_t}$
Gene Flow and Genetic Drift:

\[ F_{st} \approx \frac{1}{4Nm+1} \]

An effective 1 migrant per generation \((N_{ef}, m)\) will keep \(F_{st}\) low for neutral loci.

Coalescence
Before Gene Flow of Two Randomly Sampled Genes From Region 1

Region 1 Region 2
Gene Flow and Coalescence:

\[
\text{Prob. (gene flow before coalescence | gene flow or coalescence)} \approx \frac{4N_0m}{4N_0m + 1}
\]

\[= 1 - F_{st}\]

Also, Slatkin (1991) showed:

\[
\frac{\tilde{t}_0}{\tilde{t}} = \frac{4N_0m}{1 + 4N_0m}
\]

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.
The Wahlund Effect

Named after the person who showed that population stratification can cause deviations from Hardy-Weinberg Genotype Frequencies.

This provides yet another measure of the ratio of genetic drift to gene flow in terms of variances of allele frequencies (not identities by descent or state) that is called “\( f_{st} \)”, which is often confused with \( F_{st} \).

Let \( p_i \) be the frequency of allele \( A \) in deme \( i \).

Let \( N \) be the total population size (\( N = \sum N_i \)).

Let \( w_i \) the proportion of the total population that is in deme \( i \) (\( w_i = N_i / N \)).

Assume (for now) random mating within each deme.

Then within deme \( i \):

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

For the total population:

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

The Wahlund Effect

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

\[ \text{Var}(p) = \sigma_p^2 = \sum_{i=1}^{n} w_i (p_i - \bar{p})^2 = \sum_{i=1}^{n} w_i p_i^2 - \bar{p}^2 = \sum_{i=1}^{n} w_i q_i^2 - \bar{q}^2 \]

So:

\[ \text{Freq.}(AA) = \sum_{i=1}^{n} w_i p_i^2 - \bar{p}^2 + \bar{p}^2 = \bar{p}^2 + \sigma_p^2 \]
\[ \text{Freq.}(aa) = \sum_{i=1}^{n} w_i q_i^2 - \bar{q}^2 + \bar{q}^2 = \bar{q}^2 + \sigma_p^2 \]
\[ \text{Freq.}(Aa) = 1 - \text{Freq.}(AA) - \text{Freq.}(aa) = 2\bar{p}\bar{q} - 2\sigma_p^2 \]
\[ \text{Freq.}(Aa) = 2\bar{p}\bar{q}(1 - \frac{\sigma_p^2}{\bar{p}\bar{q}}) = 2\bar{p}\bar{q}(1 - f_{st}) \]

\[ f_{st} = \frac{\sigma_p^2}{\bar{p}\bar{q}} \]
The Wahlund Effect

The Genotype Frequencies in the Total Population Are:

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

Freq. \( (Aa) \) = 2\( pq(1 - f_{st}) \)

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

\( f_{st} \) measures the deviation from Hardy-Weinberg Genotype Frequencies in the total population due to differences among demes. Specifically, looks at deviations from expected heterozygosities:

\[
f_{st} = 1 - \frac{\text{Freq.}(Aa)}{2pq} = \frac{2pq - \text{Freq.}(Aa)}{2pq} = \frac{H_t - H_s}{H_t}
\]

The Average “Heterozygosity” Expected by Randomly Choosing Two Genes From Within a Subpopulation = \( H_s = \frac{[140(0.36)+372(0.29)]}{512} = 0.309 \)

\( H_s \) = 2\( pq = 0.36 \)

140 Pueblo Indians

<table>
<thead>
<tr>
<th>( M )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.76</td>
<td>0.24</td>
</tr>
</tbody>
</table>

372 Australian Aborigines

<table>
<thead>
<tr>
<th>( M )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.176</td>
<td>0.824</td>
</tr>
</tbody>
</table>

The “Heterozygosity” Expected by Randomly Choosing Two Genes From The Total Population = \( H_T = 2pq = 2(0.34)(0.66) = 0.449 \)

\( f_{ST} = (H_T - H_s)/H_T = (0.449 - 0.309)/0.449 = 0.312 \)
**f_{ST}** Measures The Balance of Gene Flow to Drift on a 0-1 Scale

\[ f_{ST} = \frac{H_T - H_S}{H_T} \]

- \( H_T = H_S \Rightarrow \) All Demes Have Identical Gene Pools; All Variation Shared Equally Throughout The Species
- \( H_S = 0 \Rightarrow \) No Variation Within Demes; All Variation Exists As Differences Between Demes’ Gene Pools

Gene Flow dominates Genetic Drift dominates

---

**Human Population Structure**

The Low **f_{ST}** Value For Humans Indicates Much Gene Flow Among Human Populations

[Diagram showing percentage of total genetic variation, among individuals from the same population, among populations on the same continent, between continents]
F \approx \frac{1}{4Nm+1}

Is defined in terms of identity by descent, so N is really \( N_{\text{ef}} \).

But \( f_{st} \) is defined as deviation from HW, and Li (1955) showed that the equilibrium variance under the island model is:

\[
\sigma^2 = \frac{\bar{pq}}{2N_{ev} - (2N_{ev} - 1)(1-m)^2}
\]

\[
f_{st} = \frac{\sigma^2}{\bar{pq}} \rightarrow f_{is} = \frac{1}{2N_{ev} - (2N_{ev} - 1)(1-m)^2} = \frac{1}{4N_{ev}m+1}
\]

---

The Wahlund Effect and System of Mating

Let \( p_i \) be the frequency of allele A in deme i.

Let \( N \) be the total population size (\( N=\sum N_i \)).

Let \( w_i \) the proportion of the total population that is in deme i (\( w_i=N_i/N \)).

Assume non-random mating within each deme with \( f_{is} \).

Then within deme j:

Freq.(\( AA \) in deme j) = \( p_j^2 + p_jq_jf_{is} \)
Freq.(\( Aa \) in deme j) = \( 2p_jq_j(1 - f_{is}) \)
Freq.(\( aa \) in deme j) = \( q_j^2 + p_jq_jf_{is} \)
The Wahlund Effect and System of Mating

With respect to the total population, the genotype frequencies are:

Freq.(AA) = \sum_{j=1}^{n} w_j p_j^2 + w_j p_j q_j f_{is}

= \sum_{j=1}^{n} w_j p_j^2 + \sum_{j=1}^{n} w_j (p_j - p_j^2) f_{is}

= \overline{p}^2 + \sigma_p^2 + f_{is} (\overline{p} - \overline{p}^2 - \sigma_p^2)

= \overline{p}^2 + \overline{p}q(f_{st} + f_{is}(1 - f_{st}))

Let \( f_{it} = f_{st} + f_{is}(1 - f_{st}) \)

Freq.(AA) = \overline{p}^2 + \overline{p}q f_{it}

The Wahlund Effect and System of Mating

With respect to the total population, the genotype frequencies are:

Freq.(AA) = \overline{p}^2 + \overline{p}q f_{it}

Freq.(Aa) = 2\overline{p}q(1 - f_{it})

Freq.(aa) = \overline{q}^2 + \overline{p}q f_{it}

(1 - f_{it}) = 1 - f_{st} - f_{is}(1 - f_{st}) = (1 - f_{st})(1 - f_{is}), so the deviation of heterozygote genotype frequency from Hardy-Weinberg at the total population level (1 - f_{it}) is partitioned into a component due to the local system of mating (1 - f_{is}) and a component due to differences in allele frequencies across local demes (1 - f_{st})
The Wahlund Effect and System of Mating

E.g., Yanomama:

\[ f_{is} = -0.01, \text{ indicating an avoidance of system–of–mating inbreeding (incest taboo) and a slight excess of observed heterozygosity within villages} \]

\[ f_{st} = 0.073 \text{ (between villages)} \]

\[ f_{it} = f_{st} + f_{is} (1 - f_{st}) = 0.073 - 0.01(0.927) = 0.064 \]

Even though the Yanomama avoid inbreeding within villages \((f_{is} = -0.01 < 0)\), the Wahlund effect creates an overall heterozygosity deficiency \((f_{it} = 0.064 > 0)\) at the tribal level.

The equation: \( F/f \approx \frac{1}{4Nm+1} \) is NOT the universal relationship

Between gene flow and genetic drift, as often presented. E.g., consider the one-dimensional stepping stone model (isolation by distance):

Common Gene Pool From All Demes
Impact of Drift and Gene Flow On $f_{st}$ In The Stepping Stone Model

$$f_{st} = \frac{1}{1 + 4N_{ev} \sqrt{2m_1 m_\infty}} \quad \text{When } m_1 >> m_\infty$$

Because the two migration parameters appear as the product $m_1 m_\infty$, this means that even small amounts of long distance gene flow have a major impact on $f_{st}$.

The reason is that the evolutionary impact of gene flow depends both on the amount of gene flow and the difference in allele frequency. The farther the distance, the greater the difference in allele frequency in general, so long distance dispersal has a disproportionate evolutionary impact.

E.g., let $N_{ev} = 100$ and $m_1 = 0.1$. Then

$$f_{st} = 0.053 \text{ if } m_\infty = 0.01$$  $$f_{st} = 0.276 \text{ if } m_\infty = 0.001$$

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.
Malecot showed that a general model of isolation by distance yields:

\[ f_{st}(x) = ae^{-bx}c \]

where \( f_{st}(x) \) is the pairwise \( f_{st} \) between two populations separated by a distance \( x \), where \( a \) & \( b \) are parameters estimated from the data and \( c = \frac{1}{d} \) (dimensionality of the habitat - 1).

Humans at global level

Genetic Distance

- The pairwise \( f_{st} \) is a type of population genetic distance that quantifies the differences between the gene pools of two populations.
- Many other population genetic distances are available, but all measure the degree of difference between two gene pools.
- Another type of genetic distance is a molecule genetic distance that measures the difference between two molecules of DNA; e.g., the number or percent of nucleotide differences.
f_{st} and Molecule Genetic Distance

- When you survey for genetic variation at the DNA sequence level, there is often so much variation that the probability of two randomly chosen genes being identical, even within the same deme, is very small and therefore hard to estimate reliably. “Heterozygosity” within demes often approaches one even when the demes’ gene pools are very different, allowing little discrimination with f_{st}.
- Instead of saying two genes are identical versus not identical, we can use a molecule genetic distance to measure the degree of non-identity.
- Then you can perform a standard f_{st} analysis using not identity/non-identity, but rather a quantitative measure of identity and non-identity. Such an analysis is called AMOVA (Analysis of MOlecular VAriation).

Genetic Survey of Lipoprotein Lipase

LPL Has 10 Exons Over 30 kb of DNA on Chromosome 8p22
Sequenced 9,734 bp from the 3' End of Intron 3 to the 5' End of Intron 9
Sequenced:
  24 Individuals from North Karelia, Finland (World’s Highest Frequency of CAD)
  23 European-Americans from Rochester, Minnesota
  24 African-Americans from Jackson, Mississippi
Found 88 Variable Sites
Ignored Singleton and Doubleton Sites and Variation Due to a Tetranucleotide Repeat, but Phased the Remaining 69 Polymorphic Sites by a Combination of Using Allele Specific Primer Pairs and “Haplotype Subtraction”
The Phased Site Data Identified 88 Distinct Haplotypes, treated as alleles; ie., 71 individuals with 88 alleles, with no allele being very common.
Genetic Survey of Lipoprotein Lipase

With 88 alleles in 71 individuals, with no allele being very common, almost every individual was a heterozygote in all three populations. Hence, $H_s \approx 1$, so $f_{st} = 0.02$, not significantly different from 0. Therefore, found no significant evidence for genetic differentiation among these three populations.

Instead of using heterozygosity, now use $\Phi$=the Molecule Genetic Distance of the number of nucleotides by which the pair of genes differed instead of identical vs. non-identical.

Obtain $\Phi_{st} = (\Phi_t - \Phi_s)/\Phi_t = 0.07$, 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 and heterozygosity as a qualitative measure of genetic differentiation.

In general, when high levels of genetic variation are encountered, a quantitative scale of differences between alleles is preferable to a qualitative one.

Like $f_{st}$, $\Phi_{st}$ can be broken down into hierarchical components.