1. Population Structure

2. Wahlund Effect

3. Population subdivision and F coefficients

4. Migration

5. Migration and Differentiation

6. Methods to estimate population structure

7. Methods to estimate migration
Population Structure
Most populations are grouped into smaller subpopulations within which mating usually takes place.

- Ubiquitous across species.
- Genetic differentiation between subpopulations.
- i.e. allele frequencies among subpopulations maybe different.
- This genetic differentiation (change in allele frequency within each subpopulation) can be due to what?
- Extreme case of subpopulation structure?
Human population structure - isolation by distance
Chimpanzee population structure - isolation by geography
Convergent lady beetles - homogenization by augmentation
Wahlund Effect
Problems in sampling

What is the effect of sampling what appears to be a single population, but is actually two or more subpopulations with limited gene flow between them?

- Consider a population that is divided into two subpopulations of equal size, randomly mating within each subpopulation, but individuals from one subpopulation don’t mate with individuals from the other.
- Sample 100 individuals from each, obtaining:
  - Genotype $AA$ $Aa$ $aa$
  - Subpop 1 64 32 4
  - Subpop 2 4 64 64
  - Total 68 64 68
- What are the allele frequencies in each population?
- How about genotype frequencies? Are they in HWE?
- How about if you lumped them all together? What are allele frequencies and genotype frequencies?
The *perceived* deficiency of heterozygotes in a population compared to Hardy-Weinberg expectations due to population structure. Conversely, if population structure were eliminated, there would be an increase in heterozygosity, and corresponding decrease in homozygosity.

- Let the two subpopulations have allele frequencies \( p_1, q_1, p_2, q_2 \) respectively, such that \( p_1 + q_1 = 1 \), and \( p_2 + q_2 = 1 \).
- Expected genotype frequencies are then \( p_1^2, 2p_1q_1, q_1^2 \) and \( p_2^2, 2p_2q_2, q_2^2 \).
- Allele frequencies in the total (fused) population are \( \bar{p} = \frac{p_1 + p_2}{2} \) and \( \bar{q} = \frac{q_1 + q_2}{2} \).
- Genotype frequencies in the fused population are \( \frac{p_1^2 + p_2^2}{2} \), \( p_1q_1 + p_2q_2, \frac{q_1^2 + q_2^2}{2} \).
- Hence difference in heterozygosity due to population structure = ?
# Define observed and expected heterozygosities

hetexp <- function(x, y)
{x*(1-x)+y*(1-y)+(1-x)*y+(1-y)*x}/2
hetobs <- function(x, y)
{x*(1-x)+y*(1-y)}

# Simulate some frequencies
p1 <- sort(runif(20, 0, 1))
p2 <- sort(runif(20, 0, 1))

# Plot heterozygosity
plot(hetexp(p1, p2), type='l',
ylab="Heterozygosity", xlab="Allele Frequency")
points(hetobs(p1, p2), type='l', col="red")

# Add a legend
legend(8, 0.15, c("HetExp", "HetObs"),
lty=c(1, 1), col=c("black", "red"))

# Compare heterozygosities
hetexp(p1, p2) > hetobs(p1, p2)
Population subdivision and F coefficients
Extending Wahlund Effect

• Generalizing the Wahlund Effect to $K$ subpopulations, where $N_k$ is the size of the $k^{th}$ subpopulation, such that $N = \sum_k N_k$ is the size of the total population.

• Let $w_k = \frac{N_k}{N}$ be the proportion of individuals in the total population that are in subpopulation $k$.

• Expected genotype frequencies in subpopulation $k$ are then $p_k^2, 2p_kq_k, q_k^2$.

• Frequency of the $A$ allele in the total population is the weighted average of $p_k$ across all $K$ subpopulations, such that $\bar{p} = \sum w_k p_k$.

• So if there were no subdivision, the observed genotype frequencies in the fused population would be $\bar{p}^2, 2\bar{p}\bar{q}, \bar{q}^2$.

• But with subdivision, the expected genotype frequencies in the total fused population would be

$$P(AA) = \sum_{k=1}^{K} w_k p_k^2, \quad P(Aa) = \sum_{k=1}^{K} 2w_k p_k q_k, \quad P(aa) = \sum_{k=1}^{K} w_k q_k^2$$
Another definition of $F$

- By definition, the variance in allele frequency across subpopulations is \[ \sigma_p^2 = \text{Var}(p) = \sum_{k=1}^{K} w_k (p_k - \bar{p})^2 = \sum_{k=1}^{K} w_k p_k^2 - \bar{p}^2 = \sum_{k=1}^{K} w_k (\bar{q} - q_k)^2 = \sum_{k=1}^{K} w_k q_k^2 - \bar{q}^2 \]
- Hence \[ P(AA) = \bar{p}^2 + \sigma_p^2 \]
- \[ P(aa) = \bar{q}^2 + \sigma_p^2 \]
- \[ P(Aa) = 1 - P(AA) - P(aa) = 1 - \bar{p}^2 - \bar{q}^2 - 2\sigma_p^2 = 2\bar{p}\bar{q} - 2\sigma_p^2 = 2\bar{p}\bar{q}(1 - \frac{\sigma_p^2}{\bar{p}\bar{q}}) \]

$F_{ST}$

\[ F_{ST} = \frac{2\sigma_p^2}{2\bar{p}\bar{q}} = \frac{H_T - H_S}{F_T} \], the proportionate reduction in heterozygosity in the total population due to population structure, where $S$ represents the subpopulation, $T$ is the total population.

- As before, \[ P(AA) = \bar{p}^2 + \bar{p}\bar{q}F_{ST} \]
- \[ P(Aa) = 2\bar{p}\bar{q}(1 - F_{ST}) \]
- \[ P(aa) = \bar{q}^2 + \bar{p}\bar{q}F_{ST} \]
Inbreeding and Population Structure

• What if the $K$ subpopulations themselves were also inbreeding, with an inbreeding coefficient $F_{IS} = \frac{H_S - H_I}{H_S}$, where $I$ represents the individual, and $S$ is the subpopulation.

• Recall our previous derivation: $P(AA) = p_k^2 + p_k q_k F_{IS}$.

• $P(Aa) = 2p_k q_k (1 - F_{IS})$

• $P(aa) = q_k^2 + p_k q_k F_{IS}$

• Now in the total population, the genotype frequency of $AA$ becomes

$$P(AA) = \sum_{k=1}^{K} w_k \bar{p}^2 + F_{IS} \sum_{k=1}^{K} w_k p_k q_k$$

$$= \bar{p}^2 + \sigma_p^2 + F_{IS} \sum_{k=1}^{K} w_k (p_k - p_k^2)$$

$$= \bar{p}^2 + \sigma_p^2 + F_{IS} (\bar{p} - \bar{p}^2 - \sigma_p^2)$$

$$= \bar{p}^2 + \bar{p} \bar{q} (F_{ST} + F_{IS}(1 - F_{ST}))$$

• Set $F_{IT} = F_{ST} + F_{IS}(1 - F_{ST})$

$F_{IT}$

$F_{IT} = \frac{H_T - H_I}{H_T}$ such that $(1 - F_{IS})(1 - F_{ST}) = 1 - F_{IT}$, also defined as the proportionate reduction in heterozygosity due to inbreeding, relative to the total population.
Interpreting $F_{ST}$

- $F_{ST}$ defines the degree of differentiation between subpopulations, so the greater the $F_{ST}$, greater the differentiation.
- When is $F_{ST} = 0$?
- When is $F_{ST} = 1$?
- Caution - this does NOT make a statement about the cause of differentiation between subpopulations.
- What causes differentiation?
Consider a total population comprised of three subpopulations, with equal number of individuals in each, i.e. $w_k = 0.333$. Given the following genotype and allele frequencies, compute the $F$ coefficients. What can you say about the subpopulations?

<table>
<thead>
<tr>
<th></th>
<th>$P_{11}$</th>
<th>$P_{12}$</th>
<th>$P_{22}$</th>
<th>$p_j$</th>
<th>$q_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>0.42</td>
<td>0.48</td>
<td>0.31</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>0.50</td>
<td>0.23</td>
<td>0.52</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>0.42</td>
<td>0.09</td>
<td>0.70</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Migration
Migration refers to the movement of some organisms (or their gametes) among subpopulations and subsequent breeding, also referred to as gene flow, or admixture. Migration acts to homogenize populations, or reduce differences between populations.
- Migration occurs over space and time.
- Migration is expected to be homogenizing, unless there is selection against migration (or migrant alleles).
- How would $F_{ST}$ vary with increased or decreased migration?
- What does migration do to allele frequencies?
Continent-Island Model

• Consider the simplest model, where migration is unidirectional from a large mainland population, called the continent to a smaller population, called the island.

• Consider a single, biallelic genetic locus, such that $p_c$ is the frequency of an allele $A$ on the continent, and $p_t$ is the frequency of the same allele in the island in generation $t$.

• Let $m$ be the migration rate, or rate of gene flow, the proportion of alleles on the island that have come from the continent in each generation.

• Then in one generation, on the island, a proportion $m$ of individuals would have just come from the continent, and a proportion of $1 - m$ would be from the island.

• Allele frequency, $p_{t+1} = (m)p_c + (1 - m)p_t = p_t + m(p_c - p_t)$

• Therefore $\Delta p = p_{t+1} - p_t = m(p_c - p_t)$
Equilibrium frequencies

- At equilibrium, setting $\Delta p = 0$ we get $\hat{p} = p_c$.
- Recursively, $p_t = p_c + (1 - m)^t(p_0 - p_c)$, where $p_0$ is the initial allele frequency on the island.
- As $t$ increases, $(1 - m)^t$ goes to 0, and $p_t$ approaches $p_c$. 

Equilibrium frequencies

# Scenario 1
pc = 0.4
p0 = 0.1
m = 0.01
pt1 <- c(1:500)
pt1[1] = p0
for (i in 2:500) {
  pt1[i] = pc + (1 - m)^i * (p0 - pc)
}

# Scenario 2
p0 = 1.0
pc = 0.0
m = 0.01
pt2 <- c(1:500)
pt2[1] = p0
for (i in 2:500) {
  pt2[i] = pc + (1 - m)^i * (p0 - pc)
}

jpeg("contisland.jpeg")
plot(pt1, type = 'l', xlab = "Time in generations", ylab = "Allele frequency on island", ylim = c(0:1))
points(pt2, type = 'l', col = "red")
dev.off()
Island Model

- Consider a large population, split into many subpopulations dispersed geographically.
- Let each subpopulation comprise an equal proportion $m$ of migrants from each other.
- Consider a single island - with respect to this island, the remaining islands together form a continent with allele frequency $\bar{p}$.
- So allele frequency in this island after one generation is $p_{t+1} = m\bar{p} + (1 - m)p_t$, which is very similar to the continent-island model.
Consider four islands in an island model with gene flow, such that \( m = 0.1 \). If initial allele frequencies in these islands are 0, 0.3, 0.7, 1.0 respectively, plot the allele frequency trajectories under an island model. What is the equilibrium frequency?
Other models

Isolation by Distance

• Direct extension of stepping-stone models, where two subpopulations that are far apart will experience little of the homogenizing effects of gene flow, and thus will be more different than two subpopulations that are close to each other.

• What would you predict under this model if you were to plot genetic differentiation measured as $F_{ST}$ against geographical distance?
Isolation by Distance contd.

**Fig. 2** Plot of linear geographic distance between sampled populations of *Emys blandingii* vs. Slatkin’s linearized $F_{st}$ genetic distance between these same sampled populations to estimate the presence of isolation by distance. This plot was derived from a Mantel test performed in GenAlex v.6.2 with 1000 permutations.

Image from Sethuraman et al. 2014 Conservation Genetics DOI:10.1007/s10592-013-0521-8
Migration and Differentiation
Let’s redefine $F$ (again!)

- How much migration is required to prevent differentiation/divergence among finite subpopulations as measured by $F_{ST}$?
- Let two alleles be drawn at random from a subpopulation in generation $t$ from generation $t-1$.
- Probability of IBD from one generation to the next is simply $\frac{1}{2N}$.
- If $m$ is the rate of migration, the probability that both alleles are IBD, and from the same population is $\frac{1}{2N}(1 - m)^2$.
- Alternately, two alleles can be not IBD, and not be migrants, and IBD in some previous generation with probability $(1 - \frac{1}{2N})(1 - m)^2 F_{t-1}$.
- Hence $F_t = \frac{1}{2N}(1 - m)^2 + (1 - \frac{1}{2N})(1 - m)^2 F_{t-1}$
- Look familiar?
- Therefore, setting $F_t = F_{t-1}$, at equilibrium, $\hat{F} = \frac{1}{1+4Nm}$.
- $Nm$ can be interpreted as the total number of immigrant individuals per generation, or $2Nm$ is the total number of immigrant alleles per generation.
Equilibrium $F_{ST}$

```r
Nm <- sort(runif(100, 0, 10))
F <- 1/(1+4*Nm)
jpeg("eqfst.jpeg")
plot(Nm, F, type='l',
     xlab="Number of migrants per generation",
     ylab="Equilibrium Fst", ylim=c(0, 1))
dev.off()
```
Methods to estimate population structure
Summary Statistics

• Population differentiation, measured as $F_{ST}$, or variants of it - for example, use `F_ST.stats` function in `PopGenome` package in R.

• Also see `mmod` package in R, function `diff_stats`.

• AMOVA - Analysis of MOlecular VAriance - see function `amova` in `pegas` package in R.

• Isolation By Distance analyses - see function `mantel.randtest` in `adegenet` package to perform Mantel Tests between genetic distances, and geographic distances.

• DAPC - Discriminant Analyses of Principle Components - built into the `adegenet` package in R, function `dapc`.

Model-based methods

- STRUCTURE - Pritchard et al. 2000
  http://www.genetics.org/content/155/2/945
- MULTICLUST - Sethuraman 2013
  http://lib.dr.iastate.edu/etd/13332
Identify **origins** of individuals each with a **single** ancestry

Approaches: Based on Hardy-Weinberg genotype proportions

Adapted from John Novembre slide
Identify **ancestry proportions** for individuals with **admixed** ancestry

Approaches: Structure (MCMC, Bayesian)
Or ADMIXTURE (quadratic programming)

Adapted from John Novembre slide
Methods to estimate migration
Recall that equilibrium $F_{ST} = \frac{1}{1+4N_em}$, so in theory, if you were able to estimate $F_{ST}$, and the effective population size, $N_e$, you should be able to estimate $m$. 
Model-based methods

- **MIGRATE** - Beerli and Felsenstein 2001

- **IMa2, IMa2p** - Hey and Nielsen 2007
  DOI:10.1073/pnas.0611164104, Sethuraman and Hey 2016
  DOI:10.1111/1755-0998.12437