Gene Flow and Paternity Analysis

Oct 6, 2006
Last Time

◆ Variation among populations: F-statistics

◆ Indirect estimates of gene flow
Today

- Lab recap
- More about indirect measures of gene flow
- Direct measures of gene flow: paternity analysis
## Results of Lab Exercise

<table>
<thead>
<tr>
<th>Program</th>
<th>GenAlEx</th>
<th>GenAlEx</th>
<th>Fstat</th>
<th>Fstat</th>
<th>PopGene</th>
<th>GenAlEx</th>
<th>PopGene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>$\phi_{RT}$</td>
<td>$\phi_{FR}$</td>
<td>$F_{ST}$</td>
<td>$R_{ST}$</td>
<td>$G_{ST}$</td>
<td>$F_{IS}$</td>
<td>$F_{IS}$</td>
</tr>
<tr>
<td>SSR_P_trichocarpa</td>
<td>0.112</td>
<td>0.110</td>
<td>0.119</td>
<td>0.108</td>
<td>0.121$^a$</td>
<td>0.037</td>
<td>0.050$^a$</td>
</tr>
<tr>
<td>AFLP_blow_fly</td>
<td>0.000</td>
<td>0.187</td>
<td>NA</td>
<td>NA</td>
<td>0.192</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AFLP_P_hybr</td>
<td>0.300</td>
<td>0.191</td>
<td>NA</td>
<td>NA</td>
<td>0.684</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>RFLP_hybr</td>
<td>0.451</td>
<td>0.173</td>
<td>0.435</td>
<td>NA</td>
<td>0.599</td>
<td>0.027</td>
<td>-0.007</td>
</tr>
<tr>
<td>Alloz_pine</td>
<td>0.003</td>
<td>0.060</td>
<td>0.042</td>
<td>NA</td>
<td>0.05</td>
<td>0.222</td>
<td>0.241</td>
</tr>
</tbody>
</table>

*a*: calculated with FSTAT, Popgene can't handle that many alleles!!!

- Pine had lowest among population variation, then *P. trichocarpa*; *Populus* hybrid zone highest
- Blow flies had no regional variation, high interpopulation variation
- *P. trichocarpa* contrast between $F_{ST}$ and $R_{ST}$
- RFLP lower differentiation than AFLP for *Populus* hybrids
- GenAlEx estimates VERY different from PopGene for poplar hybrid zone
Populus trichocarpa

Why was differentiation so high compared to pine?

Why was $R_{ST}$ lower than $F_{ST}$?

<table>
<thead>
<tr>
<th>Program</th>
<th>Among Regions</th>
<th>Among Populations</th>
<th>Among Populations</th>
<th>Among Populations</th>
<th>Inbreeding</th>
<th>Inbreeding</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR_P_trichocarpa</td>
<td>$\Phi_{RT}$</td>
<td>$\Phi_{PR}$</td>
<td>$F_{ST}$</td>
<td>$R_{ST}$</td>
<td>$F_{IS}$</td>
<td>$F_{IS}$</td>
</tr>
<tr>
<td></td>
<td>0.112</td>
<td>0.110</td>
<td>0.119</td>
<td>0.108</td>
<td>0.037</td>
<td>0.050⁴</td>
</tr>
</tbody>
</table>
Why was differentiation so high?

Why was $G_{ST}$ so much higher for AFLP?

Why big difference between GenAlex and PopGene?
Pinus mugo (Swiss Mountain Pine, Mugo Pine)

- Occurs exclusively above timber line (>6000 ft)
- Supposed to be mostly outcrossing
- Some evidence for clonality (sprouts from branches: layering)
Pinus mugo

✦ Most measures agreed well
  ➢ Good, even sampling among populations

✦ Low population differentiation
  ➢ Why?

✦ High inbreeding
  ➢ Why?

<table>
<thead>
<tr>
<th>Region</th>
<th>Site</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 2</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>50</td>
</tr>
<tr>
<td>Region 2 Total</td>
<td></td>
<td>208</td>
</tr>
<tr>
<td>Region 3</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>Region 3 Total</td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>Region 4</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>Region 4 Total</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Region1</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Region1 Total</td>
<td></td>
<td>266</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>769</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Program</th>
<th>Among Regions</th>
<th>Among Populations</th>
<th>Among Populations</th>
<th>Among Populations</th>
<th>Inbreeding</th>
<th>Inbreeding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>GenAlEx</td>
<td>GenAlEx</td>
<td>Fstat</td>
<td>Fstat</td>
<td>PopGene</td>
<td>GenAlEx</td>
</tr>
<tr>
<td>Alloz_pine</td>
<td>( \Phi_{RT} )</td>
<td>( \Phi_{PR} )</td>
<td>( F_{ST} )</td>
<td>( R_{ST} )</td>
<td>( G_{ST} )</td>
<td>( F_{IS} )</td>
</tr>
</tbody>
</table>
Blow Fly

- Nice agreement between $G_{ST}$ and Phi-PR
  - Plenty of loci compensated for low sampling
- All variation was among populations, none among regions!
- Christine?

<table>
<thead>
<tr>
<th>Region</th>
<th>Pop</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>PA</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>VA</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>WV</td>
<td>10</td>
</tr>
<tr>
<td>East Total</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Mid</td>
<td>IA</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>IN</td>
<td>10</td>
</tr>
<tr>
<td>Mid Total</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>West</td>
<td>CA</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>10</td>
</tr>
<tr>
<td>West Total</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>90</td>
</tr>
</tbody>
</table>
Importance of Gene Flow

- **Homogenizing force**
  - Speciation requires isolation

- **Primary determinant of Effective Population Size**
  - Determines inbreeding and drift

- **Must understand gene flow to design genetic conservation plans**

- **Key component of invasion process**
  - Risk of exotic introductions and transgenic organisms
Using $F_{ST}$ to Measure Gene Flow

$$Nm = \frac{(1-F_{ST})}{(4F_{ST})}$$

where $N$ is effective population size,
$m$ is migration rate per generation

◆ **Assuming island model**

- Balance between migration and genetic drift: equilibrium
- Infinite number of populations of equal size
- No selection or mutation
- Equal probability of exchange of migrants among populations
- No population extinction
Parentage Analysis

◆ Directly estimate real-time gene flow

◆ Different approaches depending on goals and configuration of system

◆ Need increasingly polymorphic markers depending on approach:
  - unique allele < paternity < maternity < parentage

Lowe, Harris, and Ashton 2004
Paternity Analysis

- All flowering plants at a site are spatially mapped and their genotypes determined at a sufficient number of loci.

- Seeds from individual mother plants are similarly genotyped.

- Paternal contribution is obtained by “subtracting” maternal genotype from that of each offspring. Paternity is inferred by comparing the multilocus genotype of pollen gametes to those of all potential males in the populations.

- Paternity analysis is particularly effective if all potential males, except one, can be excluded on the basis of genetic incompatibility with the observed pollen gamete genotype.
First step is to determine paternal contribution based on seedling alleles that do not match mother.

Notice for locus 3 both alleles match mother, so there are two potential paternal contributions.

Male 3 is the putative father because he matches paternal contributions at all loci.
Paternity Exclusion Analysis

Possible outcomes:

- Only one male cannot be excluded
- More than one male cannot be excluded
- All males are excluded

Consequences:

- Paternity is assigned
- Analyze more loci
- Pollen migration from external sources
What information can we get from paternity analysis?

- Distribution of pollen migration distances
- Factors affecting male reproductive success and mating patterns: floral phenology, three size, location, wind direction, etc..
- Realized gene flow (not only pollen dispersal)
Characteristics of an ideal genetic marker for paternity analysis

- Highly polymorphic, (i.e. with many alleles)
- Codominant
- Reliable
- Low cost
- Easy to use for genotyping large numbers of individuals
- Mendelian or paternal inheritance
Characteristics of an ideal genetic marker for paternity analysis

- Highly polymorphic, (i.e. with many alleles)
- Codominant
- Reliable
- Low cost
- Easy to use for genotyping large numbers of individuals
- Mendelian or paternal inheritance
Most Likely Method

Calculate a score for each male, reflecting probability he is correct parent:

\[
L(H_1, H_2 | g_m g_a g_o) = \frac{T(g_o | g_m g_a) \cdot P(g_m) \cdot P(g_a)}{T(g_o | g_m) \cdot P(g_m) \cdot P(g_a)} = \frac{T(g_o | g_m g_a)}{T(g_o | g_m)}
\]

where

- \( H_1 \) is probability male a is father,
- \( H_2 \) is probability male a is not father
- \( T \) is transition probability
- \( P \) is probability of observing the genotype

and \( g_o, g_m \) and \( g_a \) are genotypes of offspring, mother, and father
### Transition Probabilities

Table 1: Likelihood ratios for all compatible mother–alleged father–offspring trios. X represents any allele other than B; Y represents any allele that is neither B nor C. The frequencies of alleles B and C are denoted b and c. The likelihood ratio, \( L(H_1,H_2) \), is the probability of the offspring’s genotype given the mother’s and alleged father’s genotypes, \( T(g_o|g_m,g_a) \), divided by the probability of the offspring’s genotype given the mother’s genotype, \( T(g_o|g_m) \). A similar table is shown in more condensed form in Brenner (1997).

| Offspring’s genotype (\( g_o \)) | Alleged father’s genotype (\( g_a \)) | Mother’s genotype (\( g_m \)) | \( T(g_o|g_m,g_a) \) | \( T(g_o|g_m) \) | \( L(H_1,H_2) \) |
|---------------------------------|---------------------------------|----------------------------|-------------------|-----------------|----------------|
| BB                              | BB                              | BB                        | 1                 | b               | 1/b            |
| BB                              | BX                              | BB                        | 1/2               | b               | 1/2b           |
| BB                              | BB                              | BX                        | 1/2               | b/2             | 1/b            |
| BB                              | BX                              | BX                        | 1/4               | b/2             | 1/b            |
| BC                              | BB                              | CC                        | 1                 | b               | 1/b            |
| BC                              | BX                              | CC                        | 1/2               | b               | 1/2b           |
| BC                              | BB                              | CY                        | 1/2               | b/2             | 1/b            |
| BC                              | BX                              | CY                        | 1/4               | b/2             | 1/2b           |
| BC                              | BB                              | BC                        | 1/2               | (b + c)/2       | 1/(b + c)      |
| BC                              | BY                              | BC                        | 1/4               | (b + c)/2       | 1/2(b + c)     |
| BC                              | BC                              | BC                        | 1/2               | (b + c)/2       | 1/(b + c)      |

Advantages and Disadvantages of Likelihood

✦ **Advantages:**

- **Flexibility:** can be extended in many ways
  - Compensate for errors in genotyping
  - Incorporate factors influencing mating success: fecundity, distance, and direction
- Compensates for lack of exclusion power
  - Fractional paternity

✦ **Disadvantages**

- Often results in ambiguous paternities
- Difficult to determine proper cutoff for LOD score
Study Site Locations

<table>
<thead>
<tr>
<th>Site</th>
<th>Elev (m)</th>
<th>Precip (mm)</th>
<th>Deg Days &lt;18</th>
<th>Deg Days &gt;10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clatskanie</td>
<td>5</td>
<td>1446</td>
<td>2873</td>
<td>944</td>
</tr>
<tr>
<td>Corvallis</td>
<td>100</td>
<td>1085</td>
<td>2735</td>
<td>1122</td>
</tr>
<tr>
<td>Vinson</td>
<td>550</td>
<td>350</td>
<td>3097</td>
<td>1309</td>
</tr>
</tbody>
</table>

Marchel/ Davidson Hill

Clatskanie

Boardman

Vinson
Gene Flow in Wild Populations

Davidson Hill, Corvallis
Is high level of gene flow due to relatively small sampling area? Lack of near neighbors?
What is the shape of the curve for the immediate neighborhood? Where does tail begin?
Eastern Oregon population, showing population location and number of female and male trees per population.
Eastern Oregon Populations

- 203 adult ramets, 54 genets in 10 km radius circle
- Disjunct, degraded stands
- 849 progeny analyzed for 28 females

- Unique paternity assigned for 355 seedlings
- 29 seedlings had multiple possible fathers
- 58% of pollen gametes came from outside the sampled population (>3-10 km, mean 5.5 km)
Pollen Dispersal, Eastern Oregon

- Tail of distribution is unestimated (inestimable?)
- Negative exponential fits well to near-neighbor pollination success

\[ y = 0.69e^{-0.008d} \]
No apparent relationship

Minimum isolation distance is large enough to capture all local gene flow; pollen cloud is extensive
Pollen Dispersal, Comparison of Sites

Eastern Oregon
- 849 Seedlings
- 28 Mothers
- 54 Males
- 10 km Sampling Radius
- 58% Gene Flow

Davidson Hill
- 423 Seedlings
- 5 Mothers
- 57 Males
- 1 km Sampling Radius
- 74.7% Gene Flow

Is high level of gene flow due to limited pollen production within populations, and/or phenological incompatibility?
Marchel Tract

- 287 ramets, 221 genets in a 700 x 700 m area
- Includes mature trees in riparian strip along Willamette River, as well as regeneration in gravel pit and a 2.5 ha hybrid poplar plantation
- 306 seedlings from 9 mother trees
- Density and continuity are greater than in other populations, expect less gene flow
Marchel Tract, Location of Clones and Mother Trees
103 seedlings had unique paternity
27 seedlings had multiple paternity
43% Gene flow (mean 225 m)
Pollen Dispersal, Comparison of Sites

Eastern Oregon

849 Seedlings
28 Mothers
54 Males
10 km Sampling Radius
58% Gene Flow

Marchel Tract

239 Seedlings
5 Mothers
221 Males
0.25 km Sampling Radius
42% Gene Flow

Davidson Hill

423 Seedlings
5 Mothers
57 Males
1 km Sampling Radius
74.7% Gene Flow

Network size is consistent across sites, but is sensitive to spatial distribution

East: $y = 0.69e^{-0.008d}$

Marchel: $y = 0.65e^{-0.006d}$
Next Time

- Inferring population history and origin using molecular markers