Microsatellite data analysis

Tomáš Fér & Filip Kolář
Multilocus data

- **dominant** – heterozygotes and homozygotes cannot be distinguished
- **binary** – biallelic data (fragments)
  - presence (dominant allele/heterozygote)
  - absence (recessive allele)
  - i.e., 0-1 scoring
- **anonymous** – unknown genomic origin
- **multilocus** – simultaneous analysis of hundreds of loci, i.e. analysis covers „whole genome“

- **codominant** – heterozygotes and homozygotes can be distinguished
- **alellic** – known allelic frequencies in loci, populations...

- **anonymous** – unknown genomic origin
- **multilocus** – usually analysis of few loci (5-20)

- RAPD, ISSR, AFLP...

- microsatellites (SSRs), isozymes
### AFLP

**Advantage**
- high variability – many loci
- many independent loci (*multilocus method*)
- covering „whole“ genome
- statistical apparatus for data analysis

**Drawbacks**
- anonymous marker
- asymmetry in probability of loss and gain of fragments – yes/no?
- dominant – impossible to distinguish homozygotes and heterozygotes
- evaluation subjectivity
- unknown rate of mutation accumulation (impossible to use molecular clock)
- problematic (impossible) addition of further samples

### microsatellites

**Advantage**
- usually high variation – many alleles
- codominant – distinguish among homozygotes and heterozygotes, allelic frequencies
- models of allele evolution – „known“ relationships among alleles
- more objective evaluation
- statistical apparatus for data analysis
- possible to add further samples

**Drawbacks**
- species-specific markers
- simultaneous analysis of limited number of loci
- more limited representation of the „whole“ genome
SNPs
= single nucleotide polymorphisms

**Advantage**
- combined advantage of AFLP and SSR
- codominant – mainly biallelic
- sequence-based (NGS, e.g., RADseq)
- non-anonymous
- multilocus
- substitution changes => evolutionary models
- up to tens of thousands of loci
- study of neutral and adaptive variability (selection)

**Drawbacks**
- not yet stabilized statistical apparat
- not-stabilized laboratory techniques
- RADseq: null alleles and coverage bias
Evaluation of codominant data

diploids vs. tetraploids
Interpretation – assumptions

• alleles can be recognized
  • known ploidy
  • stutter bands
  • +A/-A PCR artefacts
  • artefact peaks peaks recognized
• allele drop-out
  • null alleles
    – mutation in priming site
    – poor-quality DNA prevents amplification of some alleles
  • longer alleles are amplified with lower probability
• non-scoring alleles out of the usual range
## Data matrix – SSRs

<table>
<thead>
<tr>
<th>Locus</th>
<th>Repeat Length</th>
<th>Length of Flanking Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLGA1</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>NLGA2</td>
<td>74</td>
<td>100</td>
</tr>
<tr>
<td>NLGA3</td>
<td>102</td>
<td>100</td>
</tr>
<tr>
<td>NLGA4</td>
<td>46</td>
<td>100</td>
</tr>
<tr>
<td>NLGA5</td>
<td>71</td>
<td>100</td>
</tr>
</tbody>
</table>

### Population Code
- **A**: outbred (d) or inbred (h) individual
- **B**: outbred (d) or inbred (h) individual
- **D**: outbred (d) or inbred (h) individual

### Population Group
- **Group 1**: 160, 160
- **Group 2**: 152, 160
- **Group 3**: 152, 162

### Missing Data
- **nd**: not determined
- **-1**: missing data

### One- or Two-Column Format
- **One-Column Format**: A, B, D
- **Two-Column Format**: NLGA1, NLGA2, NLGA3, NLGA4, NLGA5
Analysis options

• relationships among individuals – basic orientation in the structure
  • distance trees (NJ, UPGMA), networks
  • multidimensional analysis (PCoA)
  • Bayesian clustering

• population-genetic parameters
  • diverzity (% polymorphic alleles, diversity indices)
  • divergence (% of unique alleles, DW-index)
  • F-statistics, R-statistics

• testing and detecting spatial structure
  • AMOVA
  • Bayesian estimates, Mantel tests, spatial autocorrelation

• testing specific hypotheses
  • similarity and evolutionary relationships of identified groups
  • hybridization
  • origin of polyploids
  • ...
SSRs example data

<table>
<thead>
<tr>
<th></th>
<th>Typha latifolia</th>
<th>Typha × glauca</th>
<th>Typha angustifolia</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. latifolia</td>
<td>176 176 278 278</td>
<td>176 176 278 278</td>
<td>176 176 278 278</td>
</tr>
<tr>
<td>T. angustifolia</td>
<td>210 210 286 286</td>
<td>190 190 286 286</td>
<td>196 196 286 286</td>
</tr>
<tr>
<td>T. × glauca</td>
<td>180 210 278 286</td>
<td>190 190 278 286</td>
<td>269 287 179 193</td>
</tr>
<tr>
<td>advanced hybrid</td>
<td>176 210 278 286</td>
<td>190 190 278 286</td>
<td>193 193 101 101</td>
</tr>
</tbody>
</table>

Cattail (Typha) hybridization in the USA – T. × glauca (F1) is an invasive species.

Hybridization dynamics? Are there crosses among F1 and are F2 produced? Backcrossing?

Snow et al. 2010
Practical part 1

1. Make PCA/PCoA diagram and unrooted tree based on selected distance matrix from MSA
   
   • run MSA
   
   • modify matrix from MSA to CSV format (text separated by semicolons – using Excel)
   
   • run script in R to make PCoA biplot and unrooted tree
   
   • alternative: copy distance matrix to PAST and make PCoA/NJ tree using “User similarity“

2. Make AMOVA in R
   
   • run script in R
Bayesian clustering

• searching for an optimal partitioning of individuals to $K$ clusters, i.e., with maximum negative logarithm of the marginal likelihood

• result is an optimal number of clusters (i.e., „real populations“) and assignment of all individuals to that clusters

• within populations (clusters) deviation from H-W and linkage equilibrium are minimized (the individuals are assigned to clusters in the way to reach this goal)
  • mixture – each samples to just one population
  • admixture – probabilistic assignment of a sample to more populations

• software
  • BAPS 3.2 – Bayesian Analysis of Population Structure (Corander et al.)
    (stochastic optimization)
  • STRUCTURE (Pritchard et al.) (MCMC)
Proportion of membership of each pre-defined population in each of the 6 clusters

<table>
<thead>
<tr>
<th>Given Pop</th>
<th>Inferred Clusters</th>
<th>Number of Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1:</td>
<td>0.086</td>
<td>0.012</td>
</tr>
<tr>
<td>2:</td>
<td>0.011</td>
<td>0.037</td>
</tr>
<tr>
<td>3:</td>
<td>0.094</td>
<td>0.010</td>
</tr>
<tr>
<td>4:</td>
<td>0.789</td>
<td>0.005</td>
</tr>
<tr>
<td>5:</td>
<td>0.251</td>
<td>0.631</td>
</tr>
</tbody>
</table>

Allele-freq. divergence among pops (Net nucleotide distance), computed using point estimates of P.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-42.6055</td>
<td>-90.4091</td>
<td>-62.0519</td>
<td>-25.8868</td>
<td>-119.1667</td>
</tr>
<tr>
<td>2</td>
<td>-42.6055</td>
<td>-</td>
<td>-58.3355</td>
<td>-83.5292</td>
<td>-56.3593</td>
<td>-100.9015</td>
</tr>
<tr>
<td>3</td>
<td>-90.4091</td>
<td>-58.3355</td>
<td>-</td>
<td>-114.9450</td>
<td>-64.8990</td>
<td>-54.0048</td>
</tr>
<tr>
<td>4</td>
<td>-62.0519</td>
<td>-83.5292</td>
<td>-114.9450</td>
<td>-</td>
<td>-54.0265</td>
<td>-139.2714</td>
</tr>
<tr>
<td>5</td>
<td>-25.8868</td>
<td>-56.3593</td>
<td>-64.8990</td>
<td>-54.0265</td>
<td>-</td>
<td>-77.1662</td>
</tr>
<tr>
<td>6</td>
<td>-119.1667</td>
<td>-100.9015</td>
<td>-54.0048</td>
<td>-139.2714</td>
<td>-77.1662</td>
<td>-</td>
</tr>
</tbody>
</table>

Average distances (expected heterozygosity) between individuals in same cluster:

- Cluster 1: 1091.5688
- Cluster 2: 1109.7404
- Cluster 3: 1127.4684
- Cluster 4: 1034.3344
- Cluster 5: 1047.3957
- Cluster 6: 1094.7986

Inferred ancestry of individuals:

<table>
<thead>
<tr>
<th>Label (%Miss)</th>
<th>Pop</th>
<th>Inferred clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 110 (0)</td>
<td>1</td>
<td>0.002 0.003 0.005 0.985 0.002 0.002</td>
</tr>
<tr>
<td>2 111 (0)</td>
<td>1</td>
<td>0.332 0.037 0.140 0.421 0.062 0.008</td>
</tr>
<tr>
<td>3 112 (0)</td>
<td>1</td>
<td>0.452 0.036 0.015 0.462 0.031 0.003</td>
</tr>
<tr>
<td>4 113 (0)</td>
<td>1</td>
<td>0.033 0.010 0.007 0.942 0.006 0.002</td>
</tr>
<tr>
<td>5 114 (0)</td>
<td>1</td>
<td>0.009 0.003 0.002 0.962 0.022 0.002</td>
</tr>
<tr>
<td>6 115 (0)</td>
<td>1</td>
<td>0.016 0.012 0.047 0.906 0.014 0.006</td>
</tr>
<tr>
<td>7 116 (0)</td>
<td>1</td>
<td>0.009 0.005 0.003 0.972 0.009 0.001</td>
</tr>
<tr>
<td>8 118 (0)</td>
<td>1</td>
<td>0.004 0.003 0.003 0.983 0.004 0.002</td>
</tr>
<tr>
<td>9 119 (0)</td>
<td>1</td>
<td>0.002 0.004 0.003 0.986 0.002 0.002</td>
</tr>
<tr>
<td>10 120 (0)</td>
<td>1</td>
<td>0.005 0.003 0.003 0.986 0.002 0.002</td>
</tr>
</tbody>
</table>
STRUCTURE results evaluation

- Structure-sum (R script)
  - summary of all runs with a respect to K
  - similarity coefficient among runs for the same K
  - estimation of best K using deltaK approach (optimal number of clusters)
STRUCTURE results evaluation

Distruct (Rosenberg 2004)

- graphical representation of sample assignment to individual clusters

K2

K3
STRUCTURE results evaluation

graphical interface
Practical part 2

Find optimal partitioning to K clusters with STRUCTURE

- export data from MSA to the STRUCTURE format (or use PGDSpider)
- modify the matrix that it now includes column with numbers indicating the particular species (look at Typha_US_Structure_populcodes.txt)
- run STRUCTURE with parameters 10 000 burnin/20 000 run for K=1-6 with five runs for each K
- summarize STRUCTURE results with R scripts Structure-sum
  - which K has the highest LnP(D) ?
  - which Ks have high similarity coefficient?
  - which K has the highest deltaK?
- draw colour barplot using Distruct for converging K