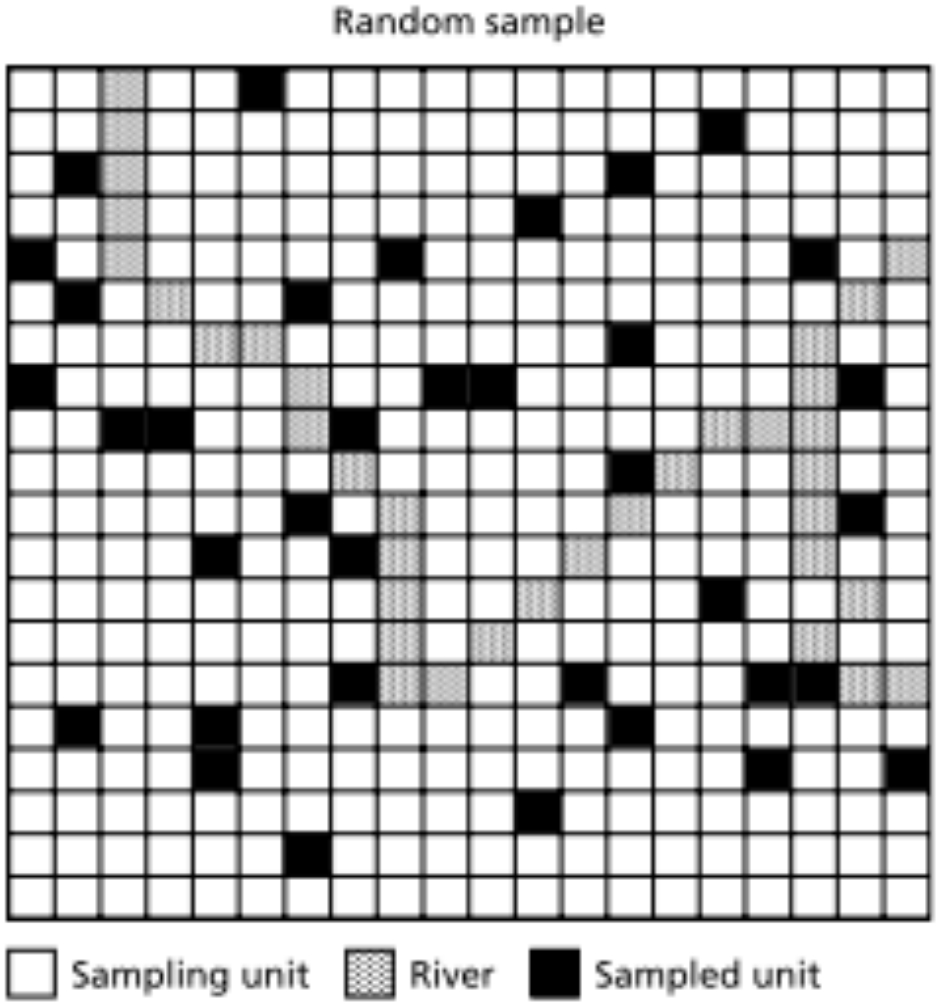
SAMPLING DESIGN

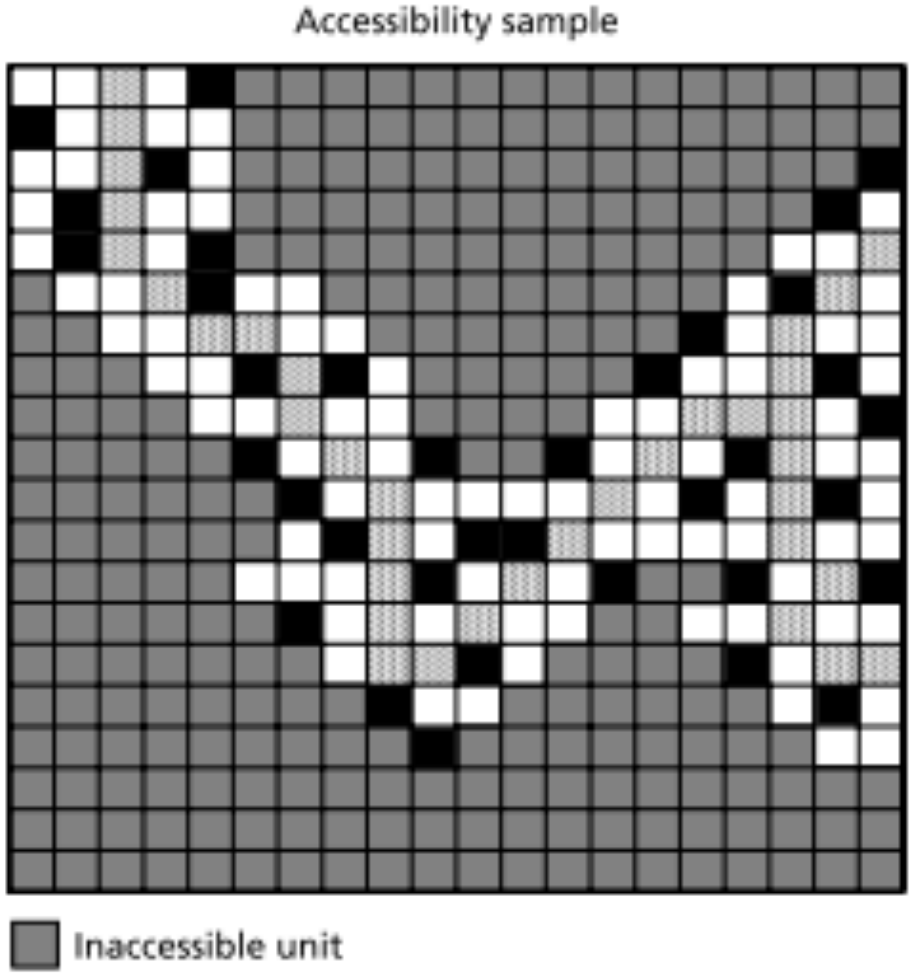
Simple random sampling

Statistical population consisting of N sampling units, from which n units are selected in such way that **every unit has an equal probability**



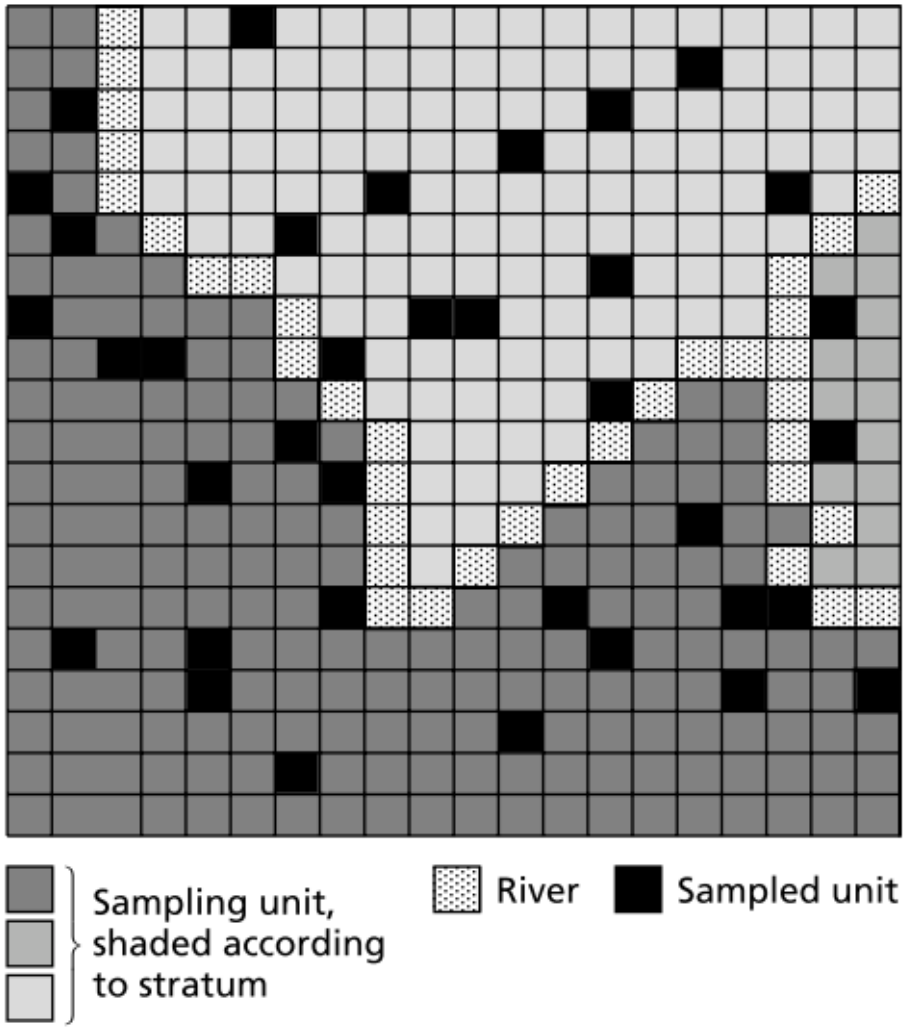
Simple random sampling

Frequently **some areas are inaccessible** (e.g. cliffs) and thus, we need to design our random sampling excluding them.



Stratified sampling

The statistical population of N units is divided into L non-overlapping strata and each strata is sampled separately (e.g. different habitats in the population)



Stratified random sampling
Each strata is sampled using simple random sampling

Systematic sampling

The samples are made at fixed points on a line, grid, or physical feature (e.g., road, river)

Systematic sampling

Random sampling

- ✓ **Sample evenly the study region**
- ✓ **Minimize sampling closely related individuals**
- ✓ **Useful if a gradient or cline exist in the study region**

MOLECULAR MARKERS

Codominant: we can distinguish heterozygotes and homozygotes

SNPs
Microsatellites
…

Dominant: we cannot distinguish heterozygotes and homozygotes

AFLPs
RAPDs
…

	Allozymes	RFLP	RAPD	AFLP
Abudance in genome	Low	High	Very high	High
Level of polymorphism	Low	Medium	Medium	Medium
Dominance	Usually codominant	Codominant	Dominant	Dominant
Development costs	Low	Medium	Low	Medium
Reproducibility	Medium/high	High	Low	Medium

	SSR	SNP	Sequencing
Abudance in genome	Medium	High	High
Level of polymorphism	High	Medium	Medium
Dominance	Codominant	Codominant	Codominant
Development costs	High	High	Medium
Reproducibility	High	High	High

COMPARISON SNPs vs. STRs

Microsatellites	SNPs
Many alleles per marker	Two alleles per marker
Less common	Much more common
PCR product size 100 to 400 bps	PCR product < 100 bps
Multiplex > 10 markers per reaction	Potentially multiplex 1000s of SNP per chip

Let's see how to introduce different data sets in **gstudio**

Rodney J. Dyer

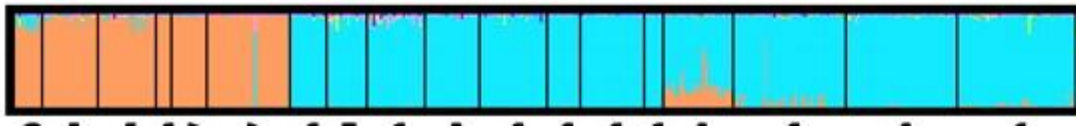
June 2, 2014

Genepop on the Web

GenAlEx
Genetic Analysis in Excel 6.5
©2006 to 2012

BayeScan

Arlequin



STRUCTURE

A lot of softwares for population genetics
….what is the best?

EXPLORATORY ANALYSIS

gstudio

Workflow

1. Partition of data: analyze different subgroups of data in our dataset
2. Plot study sites in a map
3. Estimate allelic frequencies
4. Plot allelic frequencies

Let's see an example with populations of *Miconia affinis*

GENETIC STRUCTURE

Analysis of Molecular Variance (AMOVA)

Method of estimating population differentiation directly from molecular data and testing hypotheses about such differentiation.

Any kind of raw molecular data as **Boolean vector p_i**

	220	228	230	300	340
Ind 1	1	1	1	0	0
Ind 2	0	0	1	1	0
Ind 3	1	1	0	0	1
Ind 4	0	0	0	1	1

Analysis of Molecular Variance (AMOVA)

statistics at different levels of hierarchical subdivision

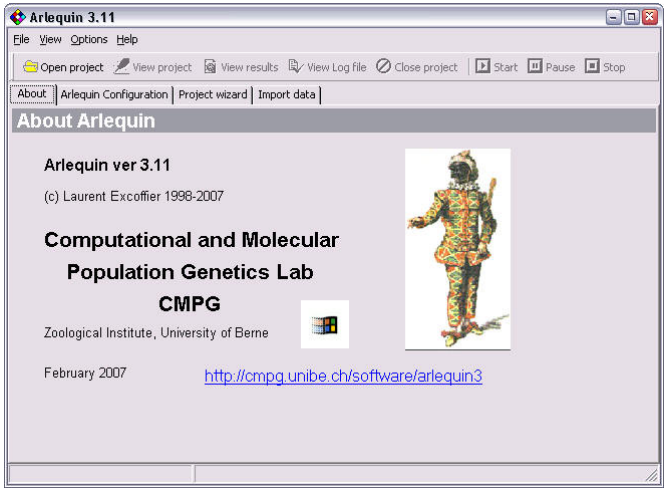
Molecular variance

- Among demes within group
- Among groups within population
- Among demes within population

is tested using a **permutational approach**

No normal distribution because data are 0 and 1

Analysis of Molecular Variance (AMOVA)



GenAlEx
Genetic Analysis in Excel
6.5
©2006 to 2012

GenAlEx
Genetic Analysis in Excel 6.5
©2006 to 2012

1. AMOVA
2. PCA
3. MANTEL TEST

You can also do it in R (gstudio, ecodist, vegan)

Let's see an example with marginal and core populations of the understory shrub *Daphne laureola*!

STRUCTURE (Bayesian clustering method)

Identify distinct genetic populations

Identification of migrants and admixed individuals

There are **K populations** each of which is **characterized by a set of allele frequencies at each locus**

Individuals are **assigned (probabilistically) to populations**, or jointly to two or more populations if their genotypes indicate admixture

STRUCTURE (Bayesian clustering method)

Assumptions of the model

Hardy-Weinberg equilibrium

Linkage equilibrium

The model does **not assume a particular mutation process**

Most of molecular markers

Ancestry Models

No admixture:

Individuals are discretely from one population or another

Admixture model:

Each individual draws some fraction of their genome from each of the K populations

Linkage model:

Similar to admixture model, but linked loci are more likely to come from the same population

Models with informative priors:

Allow to Structure to use information about sampling locations either to assist clustering with weak data, to detect migrants, or to pre-define some populations

Allelic Frequencies Models

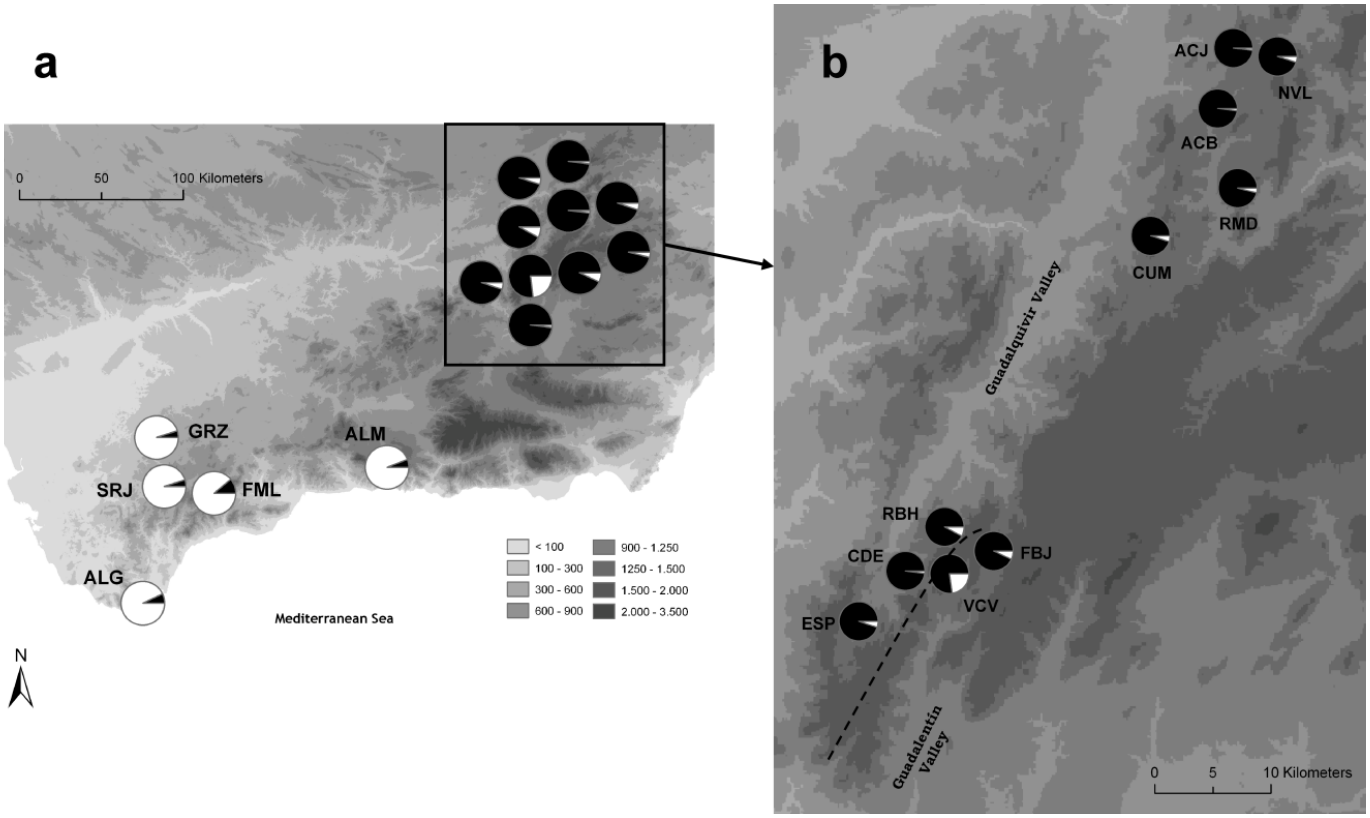
Estimating λ:

Allele frequencies in each population are independent draws from a distribution that is specified by a parameter called λ.

Correlated allele frequencies:

Assumes that frequencies in the different populations are likely to be similar (migration and shared ancestry)

STRUCTURE



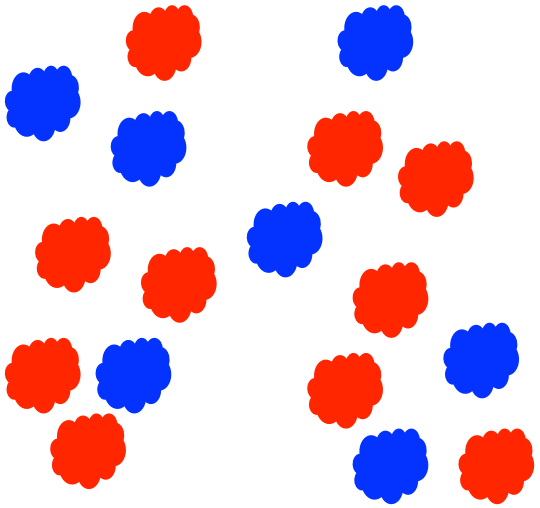
Two main genetic clusters corresponding to western and eastern populations

FINE SCALE-SGS: SPAGeDi

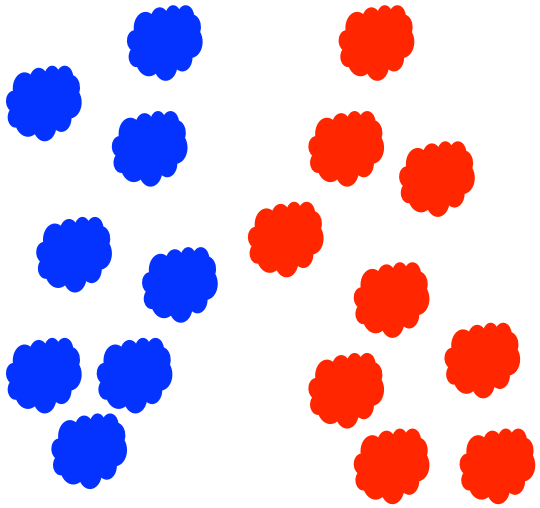
http://ebe.ulb.ac.be/ebe/SPAGeDi.html

Objective : spatial genetic structure of mapped individuals and/or mapped populations using genotype data of any ploidy level.

NO SGS



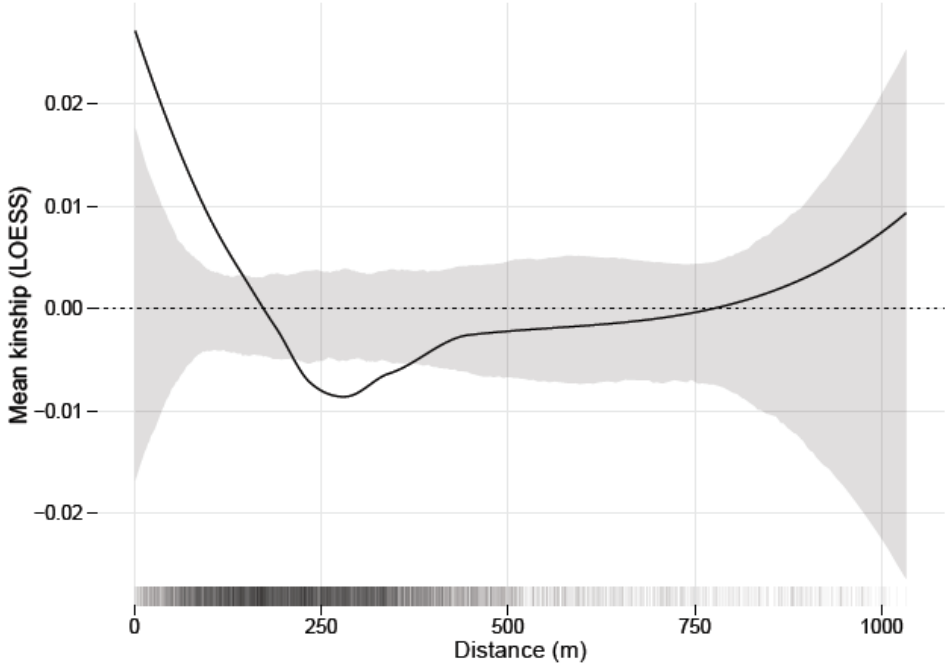
STRONG SGS



FINE SCALE-SGS: SPAGeDi

http://ebe.ulb.ac.be/ebe/SPAGeDi.html

Objective : spatial genetic structure of mapped individuals and/or mapped populations using genotype data of any ploidy level.



You can also estimate pairwise Kinship using **gstudio** (*function: genetic_relatedness*)

Kinship coefficient (Coancestry coefficients)

Kinship coefficient (F) is often defined as the **probability of identity by descent** of the gene copies compared (Ritland 1996)

Estimators based on genetic markers actually estimate **“relative kinship“** that can be defined as **ratios of differences of probabilities of identity in state** (Rousset 2002; Vekemans and Hardy 2004)

Identity by state (IBS)

A DNA segment is identical by state (IBS) in two or more individuals if They have identical nucleotide sequences in this segment.

Identity by descent (IBD)

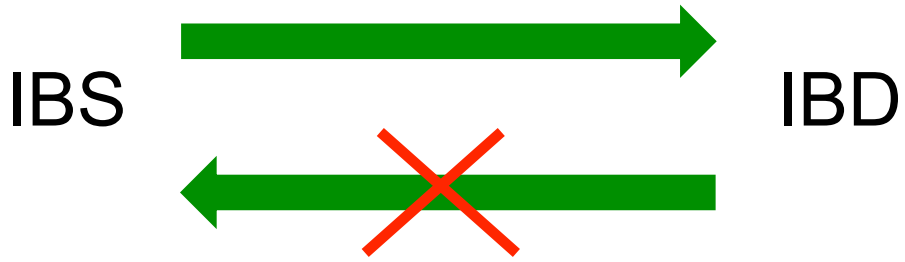
An IBS segment is identical by descent (IBD) in two or more individuals if they have inherited it from a common ancestor without recombination.

Identity by state (IBS)

A DNA segment is identical by state (IBS) in two or more individuals if They have identical nucleotide sequences in this segment.

Identity by descent (IBD)

An IBS segment is identical by descent (IBD) in two or more individuals if they have inherited it from a common ancestor without recombination.



Kinship coefficient (Coancestry coefficients)

$$F_{ij} = (Q_{ij} - Q_m) / (1 - Q_m)$$

Q_{ij} : probability of identity by state for random gene copies from i and j individuals

Q_m : average probability of identity by state for gene copies coming from random individuals in the sample.

Kinship coefficient (Coancestry coefficients)

$$\mathbf{Fij} = (Qij - Qm) / (1 - Qm)$$

Fij > 0 Two individuals **are more related than random** individuals

Fij = 0 Two individuals are equally related than random individuals

Fij < 0 Two individuals **are less related than random** individuals

Individual-level SGS

- Kinship coefficients (3)
- Relationship coefficients (6)
- Fraternity coefficients (2)
- Rousset distance (1)
- Kinship analog based on allele size (1)
- Kinship analog based on allele distances (1)

14 different coefficients for the individual-level spatial genetic structure analyses!!

Let’s see an example!

We want to analyze if an island population of the understory tropical tree *Miconia affinis* exhibit fine-scale spatial genetic structure

PARENTAGE ANALYSIS

Exclusion

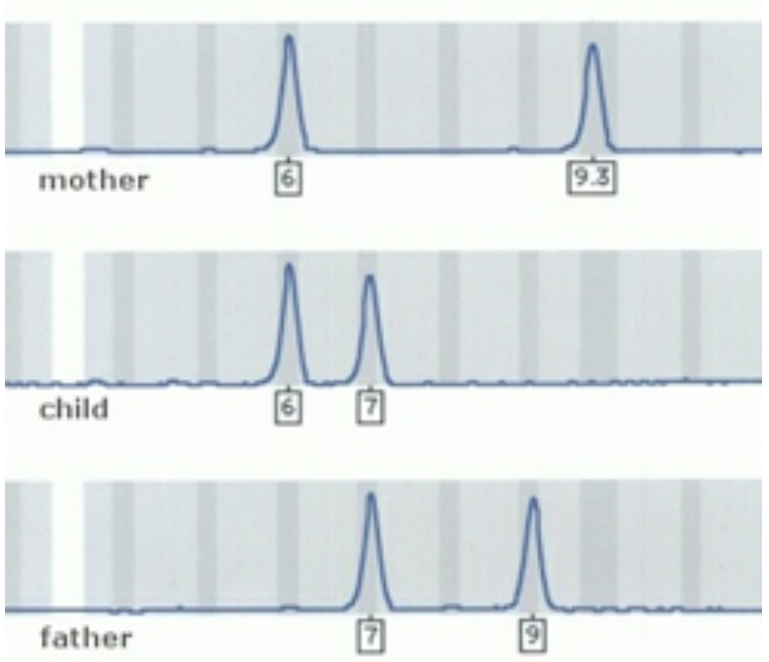
Uses incompatibilities between parents and offspring to reject particular parent-offspring hypotheses.

Error sources

✓Genotyping errors

✓Null alleles

✓Mutations



PARENTAGE ANALYSIS

Fractional

Split the offspring among all compatible males

Disadvantage: It does not make sense from the biological viewpoint

Advantage: Better statistical properties for the evaluation of some hypotheses (Jones and Ardren 2003)

- ✓ Less biased estimates of the proportion of offspring in a population parented by each of the adults
- ✓ Comparing the reproductive success of different categories of males
- ✓ For incorporating prior information about the biology of the species into the analyses

Let’s see an example with some simulated data in **Gstudio**!

PARENTAGE ANALYSIS

Categorical allocation

Uses likelihood-based approaches to select the most likely parent from a pool of non-excluded parents.

For each locus

- The frequency of the offspring allele or alleles that could come from a candidate father
- Candidate parent is heterozygous or homozygous

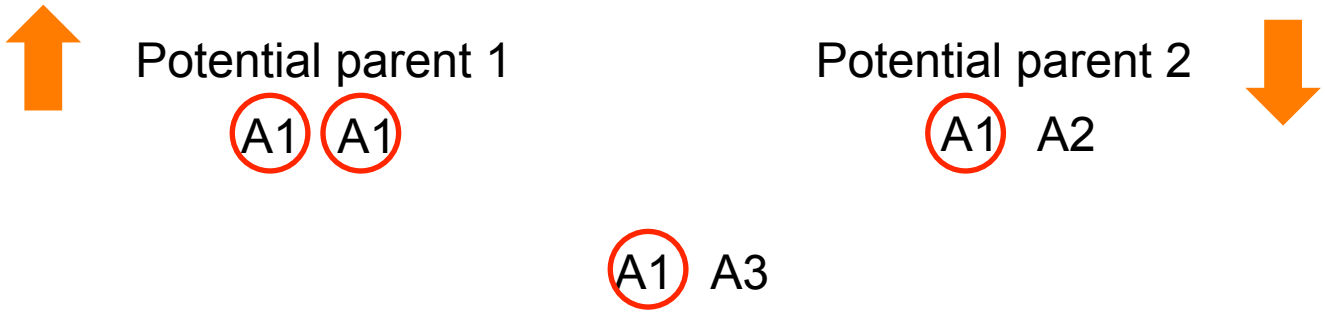
PARENTAGE ANALYSIS

Categorical allocation

Uses likelihood-based approaches to select the most likely parent from a pool of non-excluded parents.

For each locus

- The frequency of the offspring allele or alleles that could come from a candidate father
- Candidate parent is heterozygous or homozygous



PARENTAGE ANALYSIS

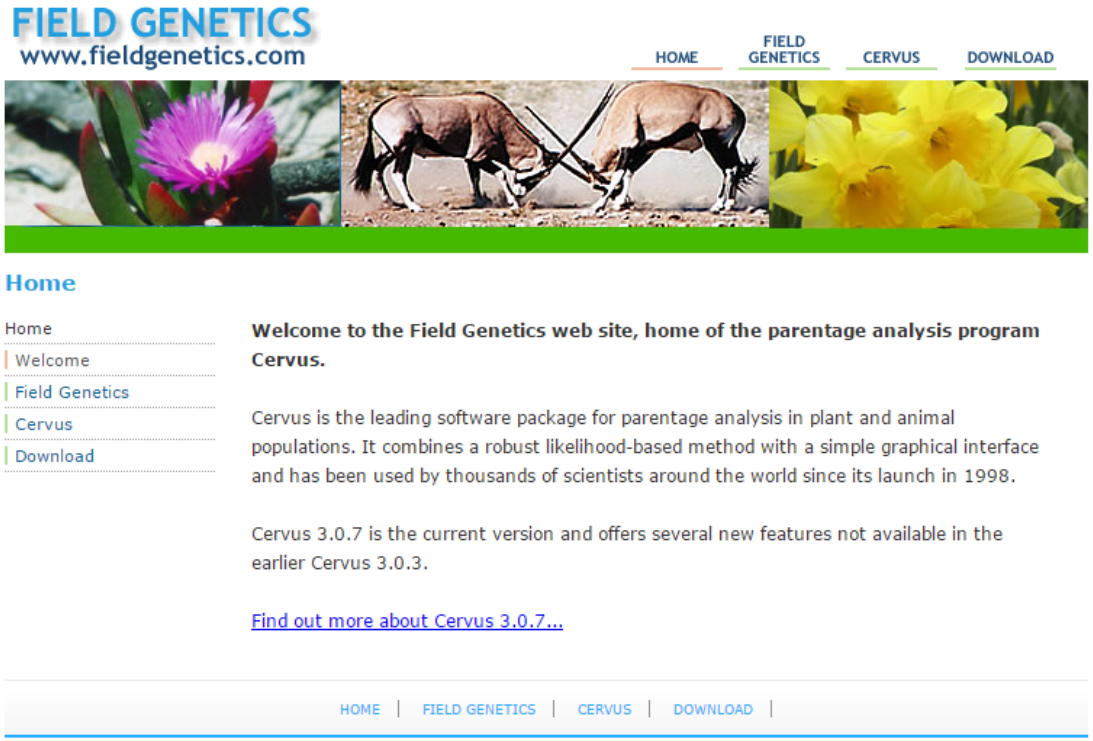
Categorical allocation

LOD score: the likelihood of an individual (or pair of individuals) being the parent (or parents) of a given offspring divided by the likelihood of these individuals being unrelated.

Offspring are assigned to parent (or parental pair) with the highest LOD score!!

Categorical allocation: Cervus software

Uses a likelihood-base approach to assign parentage combined with simulation of parentage analysis to determine the confidence of parentage assignments



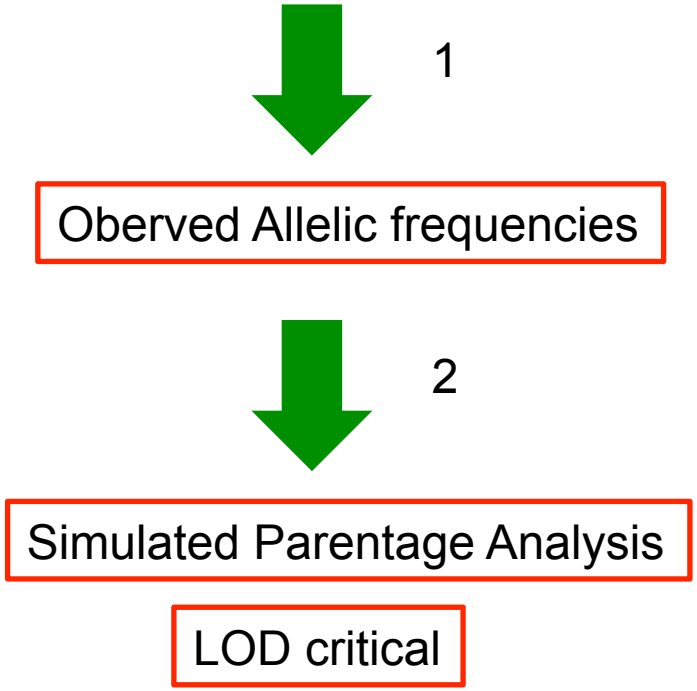
http://www.fieldgenetics.com/pages/home.jsp

Categorical allocation: Cervus software

1

Oberved Allelic frequencies

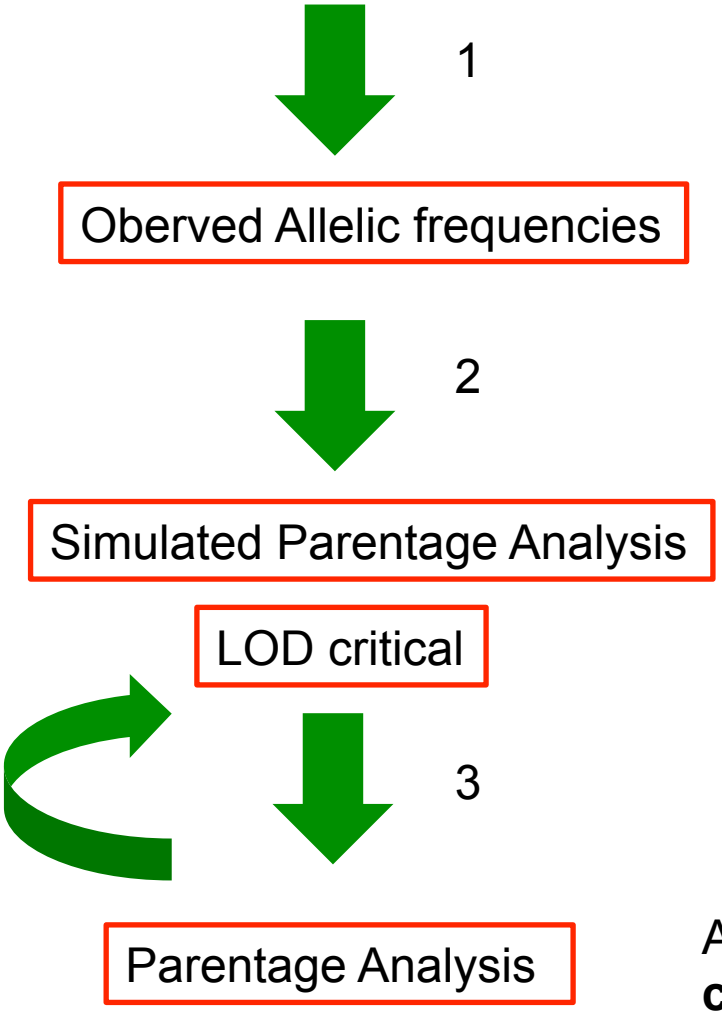
Categorical allocation: Cervus software



- ✓ Observed allele frequencies
- ✓ No. candidate parents
- ✓ Proportion of candidate parents sampled
- ✓ Completeness of genetic typing
- ✓ Estimated genotyping error rate

Critical LOD for a **relaxed confidence of 80 % and strict confidence of 95 %**

Categorical allocation: Cervus software



Any candidate parent with **LOD exceeding this critical value** is assigned parentage with 95 or 80 % confidence

Thank you very much!!!

Nichole Bennett

Nate Pope