Tools for CWR genetic diversity analysis

Helena Korpelainen

University of Helsinki, Finland

Biodiversity: three levels

- ECOSYSTEM DIVERSITY

  - The broadest level of biological diversity (e.g. deserts, rainforests, the deep sea, ponds and mountain)

  - Diversity is determined by the types of plants, animals and microorganisms, as well as by the physical characteristics (e.g. substrate, light, nutrients, etc.) and interactions (e.g. predator-prey relationships)

- TAXONOMIC DIVERSITY

  -Species, subspecies, cultivars,...

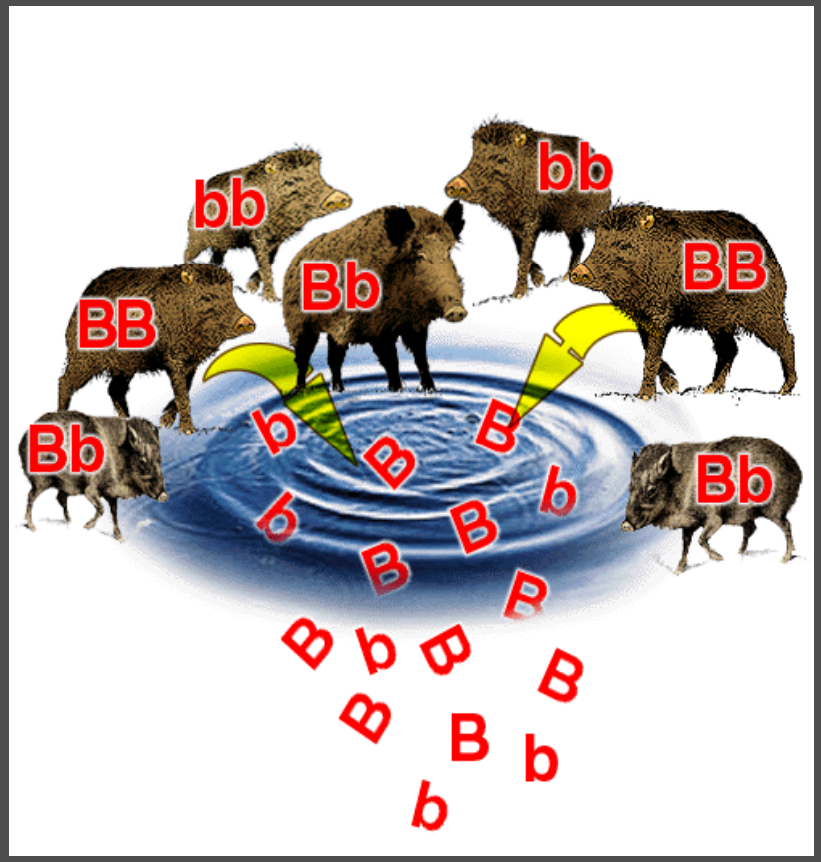
  -Species diversity combines the number of different species (species richness) with the relative abundance of a species within a given area

- **GENETIC DIVERSITY**

  -Genetic variation within and among populations of a species

Genetic variation

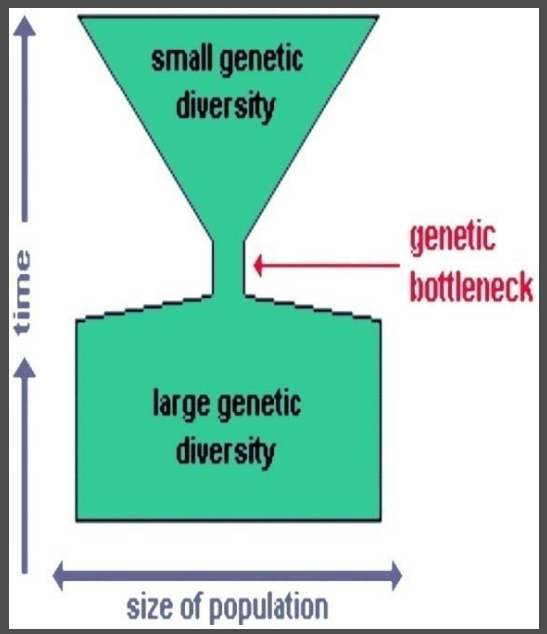
- Genetic variation is an absolute prerequisite for adaptation and evolution
- Abundant variation creates genetic flexibility and is thus usually advantageous



Gene pool (www.brooklyn.cuny.edu/.../C21_GenePool_2.GIF)

What explains the level of genetic diversity?

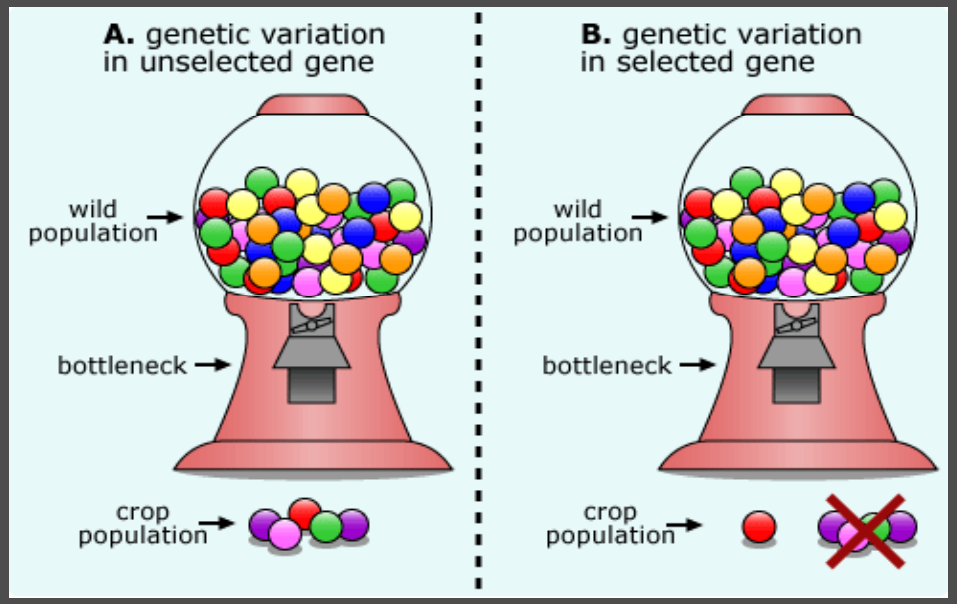
- Historical and current population sizes, population bottlenecks
- Reproductive system (crossbreeding and random mating promote)
- Natural selection (heterosis, frequency- or density-dependent selection, spatial or temporal environmental heterogeneity promote)
- Mutation rates
- Immigration and emigration among populations



Threats to genetic diversity

- Extinction of species, populations, subspecies
- Extinction of alleles
  (due to genetic drift or directional selection)
- Inbreeding reducing heterozygosity
  (alleles maintained but allocated to homozygotes → possibly inbreeding depression due to the homozygosity of deleterious recessive alleles)

Genetic diversity in wild and crop plants



Crop plants tend to have less genetic variation than their counterparts in wild populations (selection and drift)

Wild and domesticated sunflowers
(http://evolution.berkeley.edu/evolibrary/news/070201_corn)

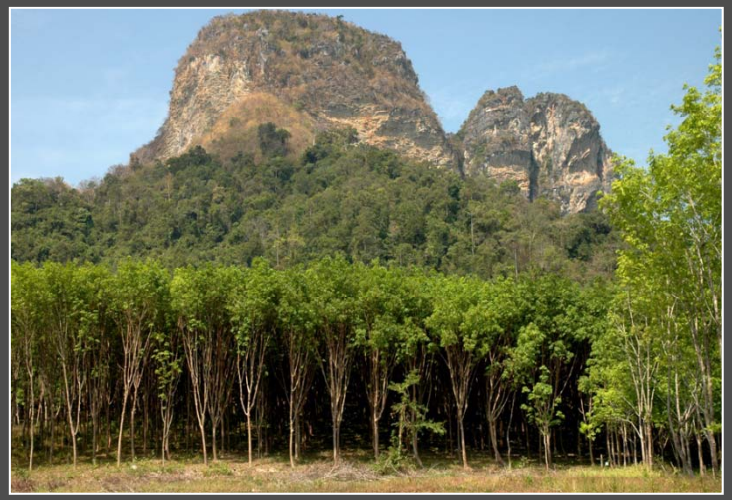
Causes of genetic erosion in CWR: Natural factors

- Genetic diversity can be lost through natural disasters, such as large-scale floods, wild fires, and severe and prolonged drought
- These events are beyond the control of humans

Flooded banana grove
(http://www.junglephotos.com/amazon/ampeople/workplay/flooding.shtml)

Causes of genetic erosion in CWR: Action of farmers

- Farmers may engage in activities that promote genetic erosion, e.g., clearing of virgin land, and the choice of planting material (narrow genetic-based cultivars)
- Especially in developed economics, farmers primarily grow improved seeds instead of landraces
- Monoculture tends to narrow genetic diversity
- Extending grazing lands into wild habitats by livestock farmers destroys wild species and wild germplasm resources



Former rain forest, now a rubber tree plantation in Thailand
(Photos: Jouko Rikkinen)

Causes of genetic erosion in CWR: General public action

- Increasing demand of land with increasing populations and demand for alternative use of land (e.g., for recreation, industry, roads) to meet the general needs of modern society

- Such actions tend to place wild germplasm in jeopardy, often include clearing of virgin land where wild species occur

Former rain forest in Thailand
(Photo: Jouko Rikkinen)

Analysis of genetic diversity

- Morphological and agronomic traits often used for basic characterization

  - such information of high interest to the users of PGR

  - such characterization can de done using simple techniques but it requires much human labour

  - allows the interpretation of relationships between the genotype and environmental conditions

- Molecular marker techniques and DNA sequencing

  - allow direct investigations of variation at the DNA level, thereby excluding all environmental influences

  - can be employed at very early growth stages

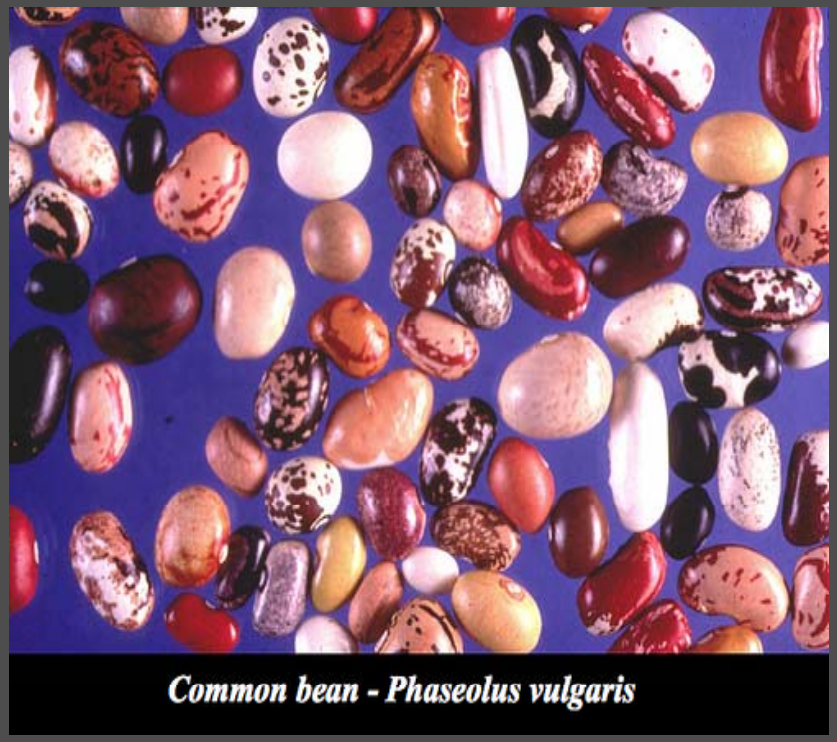
  - have marginalized other methods in diversity analyses

Analysis of genetic diversity

- Genetic variation is necessary for adaptation to changing environmental conditions
- Analysis of genetic diversity and relatedness between populations and individuals is important when aiming to

  - understand population and evolutionary biology

  - characterize and utilize available genetic resources



http://live.psu.edu/image/4951053856

Molecular markers

- What are they?

  - A detectable sequence of DNA or protein whose inheritance can be monitored

- To be useful, molecular markers must possess certain characteristics:

  - Polymorphic
  - Reproducible
  - Co-dominant inheritance (both forms detectable in heterozygotes) very useful in some applications
  - Fast and inexpensive to detect

Adaptive markers

- Many studies on diversity and plant genetic resources based on neutral molecular markers
- But also: gene-specific markers developed based on known genes of interest (may be adaptive)
- Possible to discover large amounts of expressed (i.e., protein coding) sequence information, such as expressed sequence tags (ESTs)

  - a unique chance to screen and detect molecular variation of genes at a genome-wide level

  -possible to discover polymorphisms that affect the performance of organisms (e.g., stress tolerance)

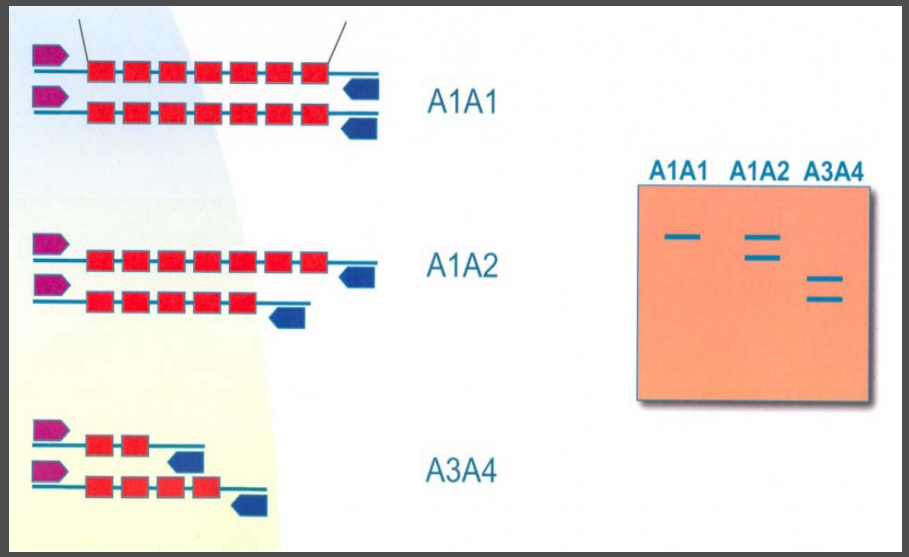
Microsatellites (simple sequence repeats, SSRs)

- Tandemly repeated DNA motifs composed of di-, tri-, tetra- and sometimes greater repeated nucleotide sequences, e.g., $[AT]_n$
- Different alleles vary in the number of units of the repeat motifs
- Developed markers species-specific (often usable within genus)
- Highly variable codominant markers

Example: $[gata]_{15}$

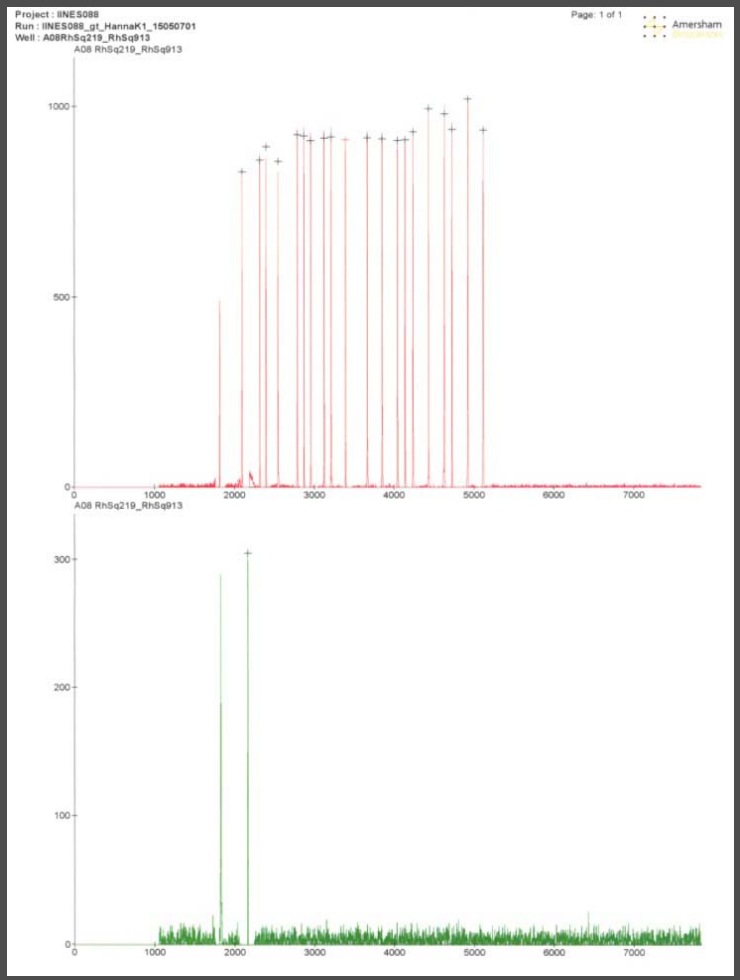
```
1   aatttttgta tttttttag agacggggtt tcaccatgtt ggtcaggctg actatggagt
61  tattttaagg ttaatatata taaagggtat gatagaacac ttgtcatagt ttagaacgaa
121 ctaacgatag atagatagat agatagatag atagatagat agatagatag atagatagat
181 tgatagtttt tttttatctc actaaatagt ctatagtaaa catttaatta ccaatatttg
241 gtgcaattct gtcaatgagg ataaatgtgg aatcgttata attcttaaga atatatattc
301 cctctgagtt tttgatacct cagattttaa ggcc
```

Microsatellites



Homozygote A1A1 (7 repeats)
Heterozygote A1A2 (7 and 5 repeats)
Heterozygote A3A4 (2 and 4 repeats)

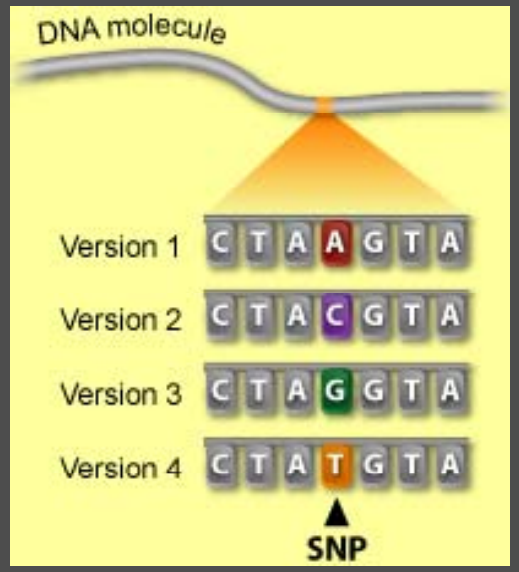
(http://www.landfood.ubc.ca/courses/agro/424)



Genotyping using capillary electrophoresis: a heterozygote individual (two peaks, green) and a ladder (red)

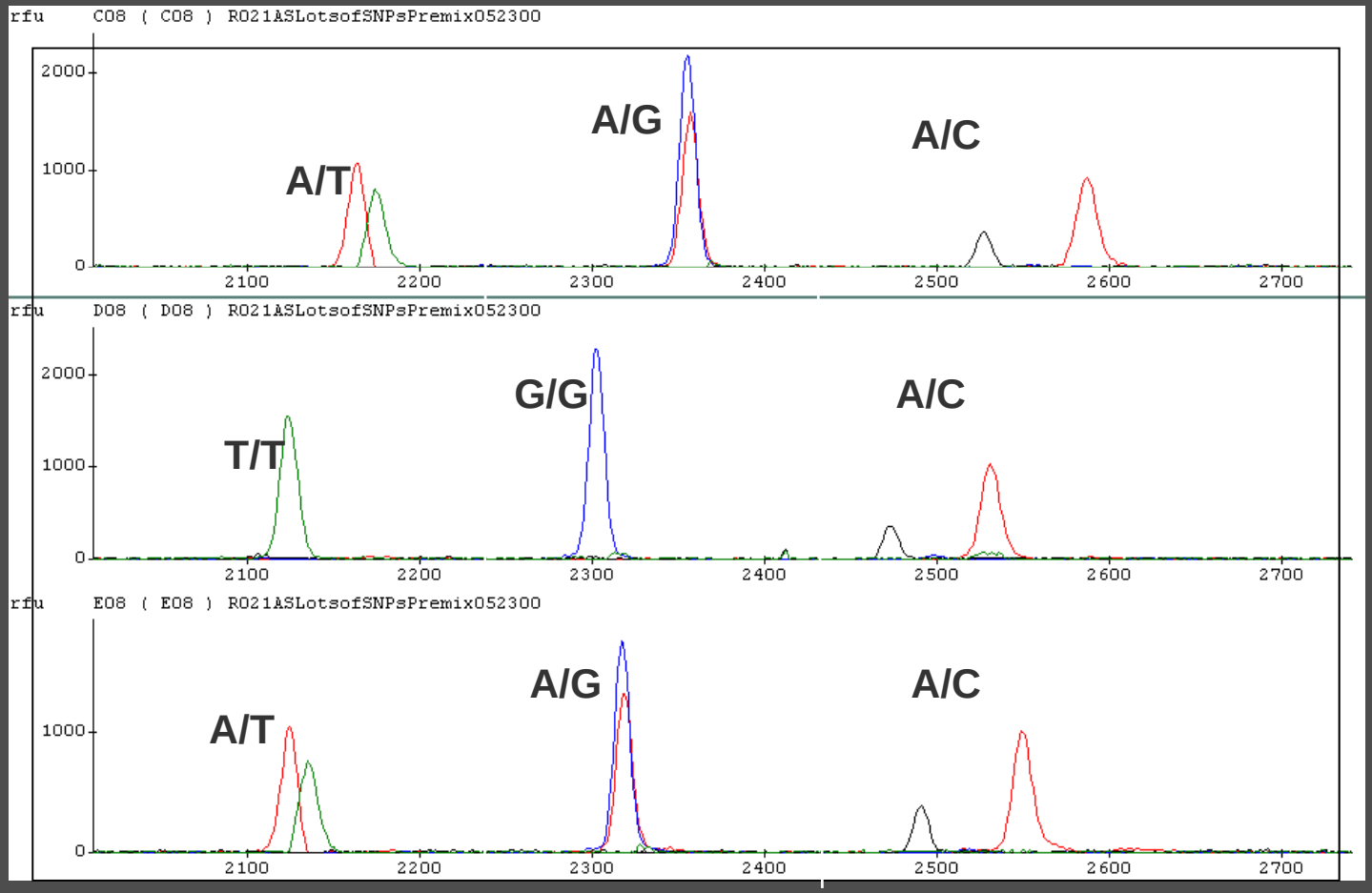
Single nucleotide polymorphisms (SNPs)

- Single-base changes in DNA, discovered by sequencing
- Variation level varies, e.g., 1-2 bp per 1000 bp in humans, every 40 bp in maize
- For some applications (e.g., germplasm identification) SNPs are not as informative as microsatellites with multiple alleles
- Several detection methods; typical approach: design primers and conduct sequencing (also commercial kits available for high-throughput SNP analysis)



http://learn.genetics.utah.edu

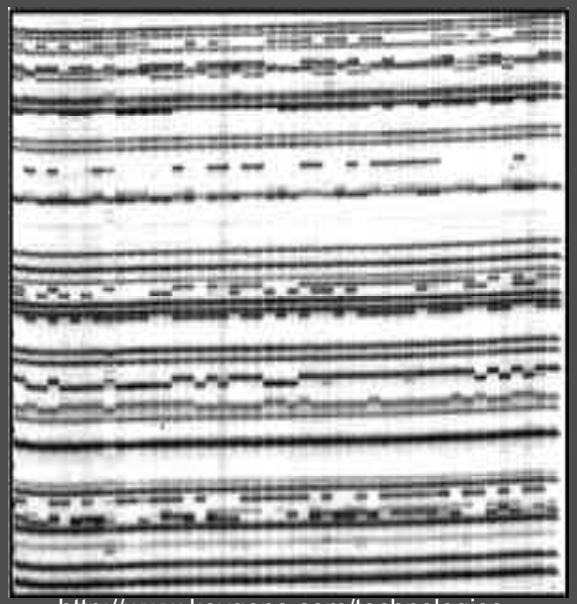
Single nucleotide polymorphisms (SNPs)



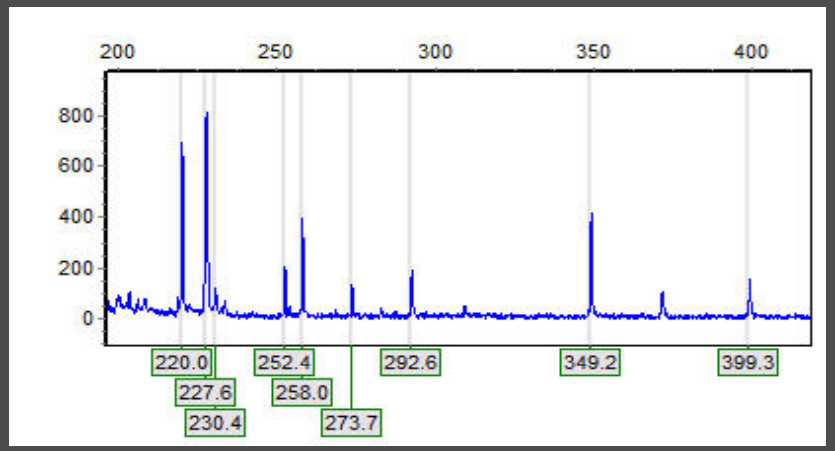
Using a commercial analysis kit, three reactions injected into each of three capillaries (Amersham Biosciences)

Amplified fragment length polymorphisms (AFLP)

- Usually two amplification steps
- Amplification products detected by electrophoresis (a gel or capillary system)
- Advantages: capability to amplify between 50 and 100 fragments at one time; no prior sequence information needed
- Disadvantages: potential allelism problems, dominant markers (usually heterozygotes cannot be distinguished)



http://www.keygene.com/technologies



Data analysis of molecular marker data

- First, genotype data from population samples (e.g., 30 individuals from each of N populations of species X), then

  - Descriptive statistics

  - Population comparisons and genetic structure analysis

  - Other analyses: demographic inference, detecting recent immigrants, neutrality tests, detecting loci under selection,…

Himalayan balsam (*Impatiens glandulifera*), highly invasive ornamental plant (http://www.luontoportti.com)

Descriptive statistics

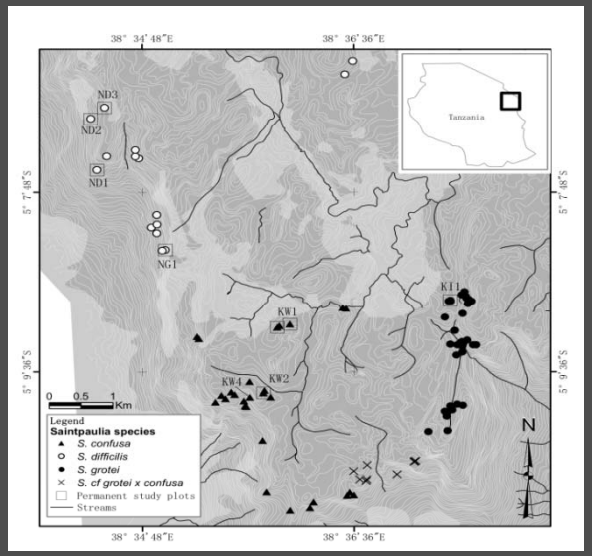
- Basic indices of genetic diversity within populations, such as
  - degree of heterozygosity
  - number of alleles
  - number of polymorphic loci
- Also more advanced statistics
  - allelic richness (number of alleles standardized to the smallest sample size)
  - private alleles (specific to some populations; possibly important variation)
  - Hardy-Weinberg equilibrium (HWE) testing at all loci (deviations due to genetic drift, selection etc.)
  - linkage disequilibrium testing between pairs of loci (common in threatened species as their population sizes are small)

Population comparisons and genetic structure analysis

- Genetic distances between populations
- Phylogenetic relationships (expressed as trees)
- Analyses of genetic variance (variation within and among populations, between regions, …)
- Gene flow estimates
- Comparisons between genetic and geographical distances, spatial genetic structure analyses (population and individual levels)

Hierarchical AMOVA of *Saintpaulia ionantha* ssp. *grotei* morphotypes "confusa", "difficilis", "hybrid" and "grotei" in Tanzania (from Kolehmainen & Korpelainen 2010).

Source of variation	% total variance
Among morphotypes	2.84
Among populations within morphotypes	3.47
Within populations	93.69



Other analyses

- Demographic inference (past population expansions/bottlenecks, effective population size estimates)
- Detecting recent immigrants (different multilocus genotypes than expected for natives)
- Neutrality tests
- Detecting loci under selection (adaptive loci)

Rare *Malus sylvestris*; how much hybridization with *M. domestica*? (www.helsinki.fi/biosci/pinkka)

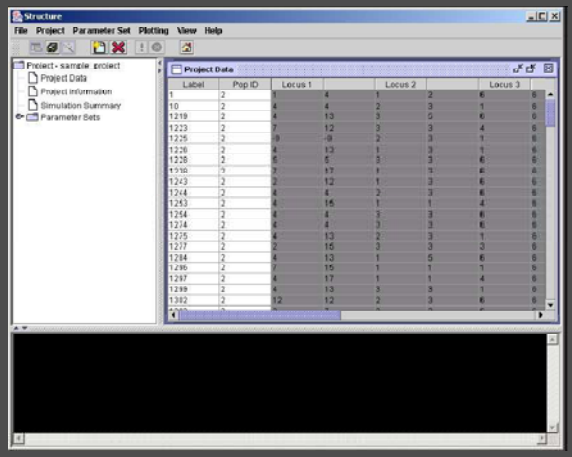
- LOTS OF DIFFERENT SOFTWARE AVAILABLE

Data analysis: software examples

- **STRUCTURE** (http://pritch.bsd.uchicago.edu/software)

  -A model-based structuring method for inferring population structure using genotype data consisting of unlinked markers

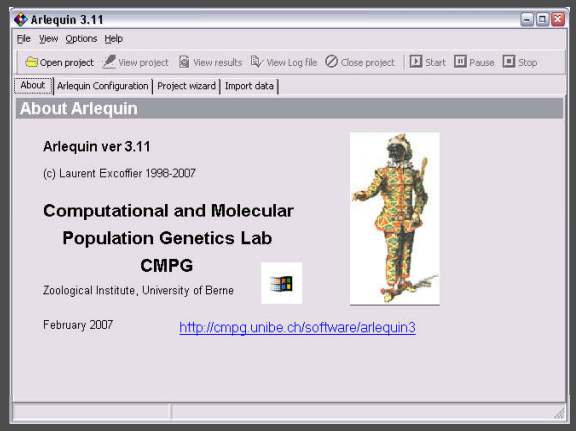
  -Applications include demonstrating the presence of population structure, identifying distinct genetic populations, assigning individuals to populations, and identifying migrants and admixed individuals



- **ARLEQUIN** (http://cmpg.unibe.ch/software/arlequin3/)

  -Provides a large set of basic methods and statistical tests to extract information on genetic features of populations

  -Data can be genotypes, haplotypes (i.e. combination of alleles at one or more loci) or allele frequencies



The case of the Lima bean (*Phaseolus lunatus*)

(Motta-Aldana et al. 2010. Crop Science 50.1773-1787)

- Question: genetic relationships between wild and domesticated *P. lunatus*
- Material: 59 wild and 50 domesticated Lima bean accessions originating from a region ranging from Mexico to Argentina
- Molecular method: sequencing of two intergenic spacers of cpDNA (atbB-rbcL and trnL-trnF) and the nuclear internal transcribed spacer ITS
- Results: at least two independent domestication events, a severe reduction in diversity because of domestication → conservation need for CWR germplasm

Steps of CWR genetic diversity analysis

- Sampling (x populations, y individuals/population)
- Molecular analyses (which specific method?)
- Data analyses (descriptive statistics, what else?)

→ conclusions, actions

Thank you!