# Analysis of genetic diversity with molecular markers

Isabel Roldán-Ruiz

August 2012

# Overview

**GENETIC DIVERSITY IS THE FOUNDATION OF BIODIVERSITY**

Without genetic diversity and variation - adaptation and evolution cannot occur in natural populations
Without genetic diversity and variation - selection is not possible in breeding populations

It follows that:
**GENETIC DIVERSITY IS THE FOUNDATION OF BREEDING**

- Genetic variation and population genetics
- Concept of population
- Hardy-Weinberg principle
- Questions addressed by population geneticists and breeders
- Forces that act on genetic diversity in natural and selected populations
- Quantifying genetic variation
  - Within populations: polymorphism and heterozygosity
  - Among populations: genetic differentiation, F-statistics
- Calculating genetic distances
  - Between genotypes
  - Between populations
- Displaying genetic relationships of a group of individuals or populations
- Examples

# Genetic variation

- Genetic variation can be described at three levels:
  1. Genetic variation within individuals (heterozygosity)
  2. Genetic differences among individuals (within-population diversity)
  3. Genetic differences among populations (genetic differentiation and fixation)

- DNA-markers are tools that allow quantification of diversity at these three levels

- **Population genetics** is the discipline that handles these aspects. **It consist in the study of genetic variation in populations and how that variation changes over time and space**. In other words, how much variation exists in natural populations, and how can we explain variation in terms of origin, maintenance, and evolutionary processes?

# Population

**Several definitions available**

- Ecology: a group of individuals of the same species that occur in the same habitat area at the same time (sometimes called a provenance, usually ‘isolated’ from similar groups of the same species)

- Genetics: an **interbreeding** group of individuals

**Population size**

- Census size N: the number of individuals
- Effective population size $N_e$: the number of individuals that stand an equal chance to mate and pass their genes to the next generation (smaller than the census size N)

$$N_e < N$$

due to skewed sex ratios, some non-breeders, some degree of inbreeding, variation in progeny survival; depends on the genetic parameter and the generation considered

$N_e = N$ if all individuals in population have equal probability of being parents of any individual of the next generation (requires panmixia, no overlapping generations, no migration, etc.)

# Hardy Weinberg principle

✓ Hardy-Weinberg principle is a model that relates **allele** frequencies to **genotype** frequencies

✓ central concept in traditional genetic diversity and differentiation models; independently formulated in 1908 by the mathematician Godfrey H. Hardy and physician Wilhelm Weinberg

**Based on five basic assumptions**

✓ population is infinitely large - no effects of genetic drift, no chance effects

✓ mating is random - no internal ‘structure’

✓ no (natural) selection - at least for the traits under study

✓ no mutation – no new alleles

✓ no migration – no ‘import’ of alleles from other populations

If these assumptions are met, the population will be in genetic equilibrium (H-W equilibrium).

**Makes two predictions (if assumptions met)**

✓ allele frequencies do not change over generations

✓ after one generation of random mating (i.e., zygotes form by random combinations of gametes, in proportion to the abundance of the alleles in the population), the genotypic frequencies will be:

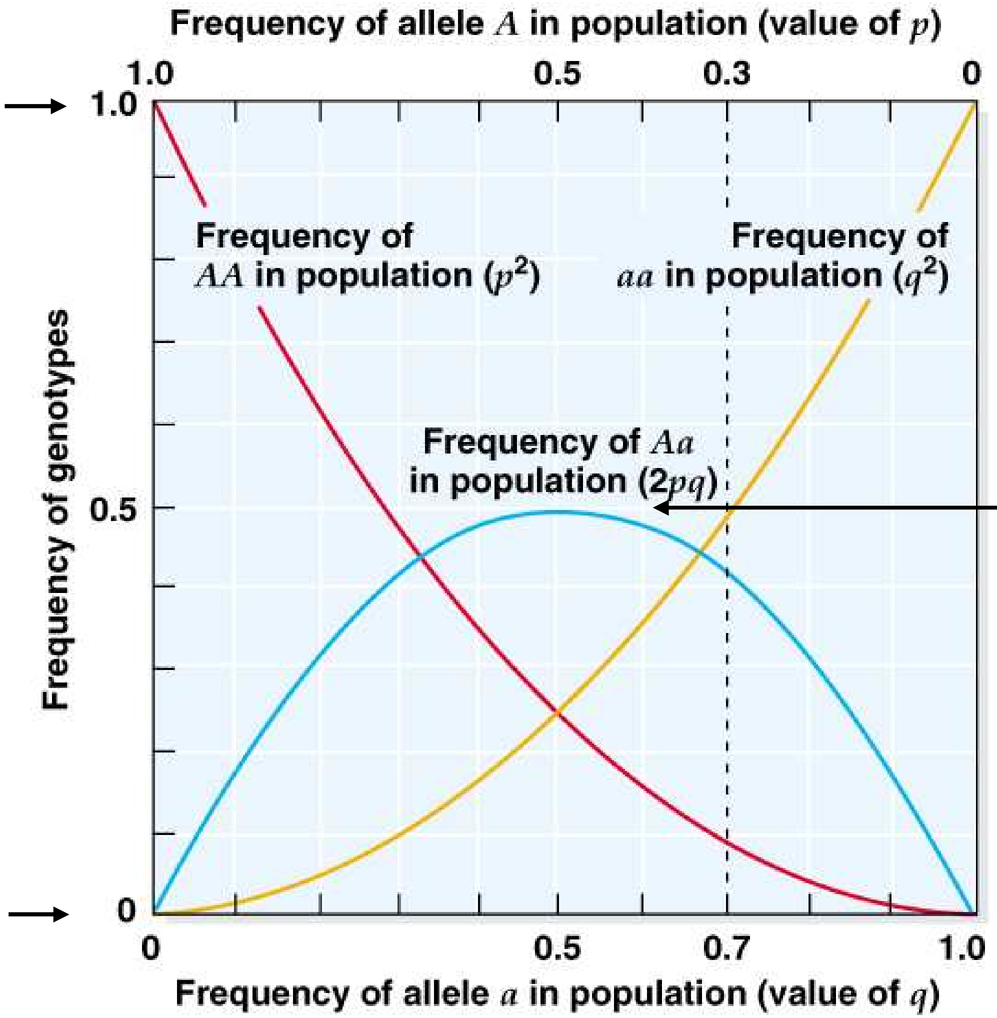
| | |
|---|---|
| $p^2$ | (frequency of genotype AA) |
| $2pq$ | (frequency of genotype Aa) |
| $q^2$ | (frequency of genotype aa) |

$p$ = allelic frequency of A
$q$ = allelic frequency of a

$$\boxed{p^2 + 2pq + q^2 = 1}$$

# Hardy Weinberg principle

Frequencies of genotypes AA, Aa, and aa relative to the frequencies of alleles A and a in populations at Hardy-Weinberg equilibrium



**p= 1**
***A* allele fixed**

**p = 0**
***A* allele lost**

**Max. heterozygosity**
**p = q = 0.5**

**q = 0**
***a* allele lost**

**q = 1**
***a* allele fixed**

# Hardy Weinberg principle

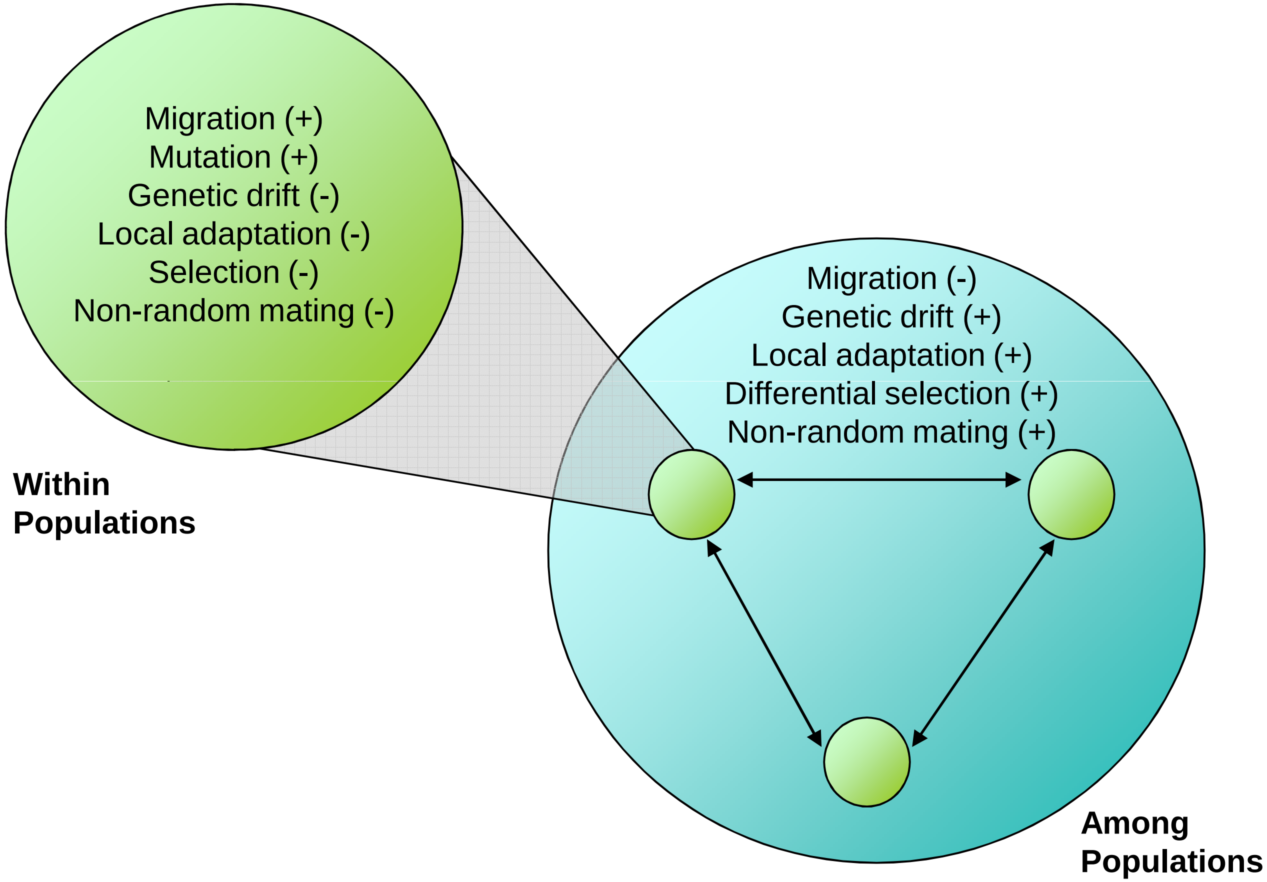
- ✓ H-W describes the properties of an ‘ideal population’, but **real populations are rarely in H-W equilibrium**:

  - Mutations may create new alleles
  - Selection may favor particular alleles or genotypes
  - Mating may be not- random => genotype frequencies will deviate from expectation
  - Population is finite => random changes in allele frequencies will happen; this is called genetic drift
  - Immigrants (i.e. by seed or by pollen) may import alleles with different frequencies, or new alleles

- ✓ How to check for H-W equilibrium?
  - test observed and expected genotype proportions with a goodness of fit test, such as a chi-square test
  - if deviation is significant, begin to determine which of the five assumptions of the Hardy-Weinberg law are violated

# Forces that act on genetic diversity

Forces that destroy H-W equilibrium are the forces that act on genetic diversity



# Questions addressed by population geneticists

- ✓ How much variation is contained in (natural) population(s)?
- ✓ What processes control and influence the observed variation?
- ✓ If two populations are differentiated (= genetically different), what forces are responsible for divergence among populations?
- ✓ How do demographic factors (such as breeding system, fecundity, changes in population size, and age structure) influence the gene pool in the population?
- ✓ Which are the genetic relationships among different accessions in genebanks or in breeding populations?
- ✓ Definition of ‘core collections’ in genebanks
- ✓ What genes were influenced by crop domestication?
- ✓ …….

# Quantifying genetic variation within populations

- ✓ **Polymorphism** (PLP): % of polymorphic loci; proportion of markers that are polymorphic
  - Usually a locus is considered polymorphic if the frequency of the most common allele is less than 95%
  - If 20 out of 50 marker loci sampled in a population have an allelic frequency of > 95% for a single allele, PLP=30/50 = 60%
- ✓ **Allelic richness** (Ar): number of alleles at a locus – standardized measures have been developed considering the number of individuals sampled in the population
- ✓ **Heterozygosity**: percentage of loci at which the average individual is heterozygous

average observed heterozygosity $H_O$ = mean frequency of heterozygotes observed at a particular locus averaged over all loci surveyed

average expected heterozygosity $H_e$ ; calculated by subtracting from 1 the expected frequency of homozygotes at a locus; averaged over all loci

calculation of the **expected heterozygosity**:

- locus j with two alleles (a and A) $\mathbf{h_j = 1 - p_a^2 - p_A^2}$
- locus j with i alleles (p denotes the allelic frequency) $\mathbf{h_j = 1 - \sum p_i^2}$
- averaged over several loci (L = number of loci) $\mathbf{H_e = \sum h_j/L}$

# $H_j$ in a two-allele system

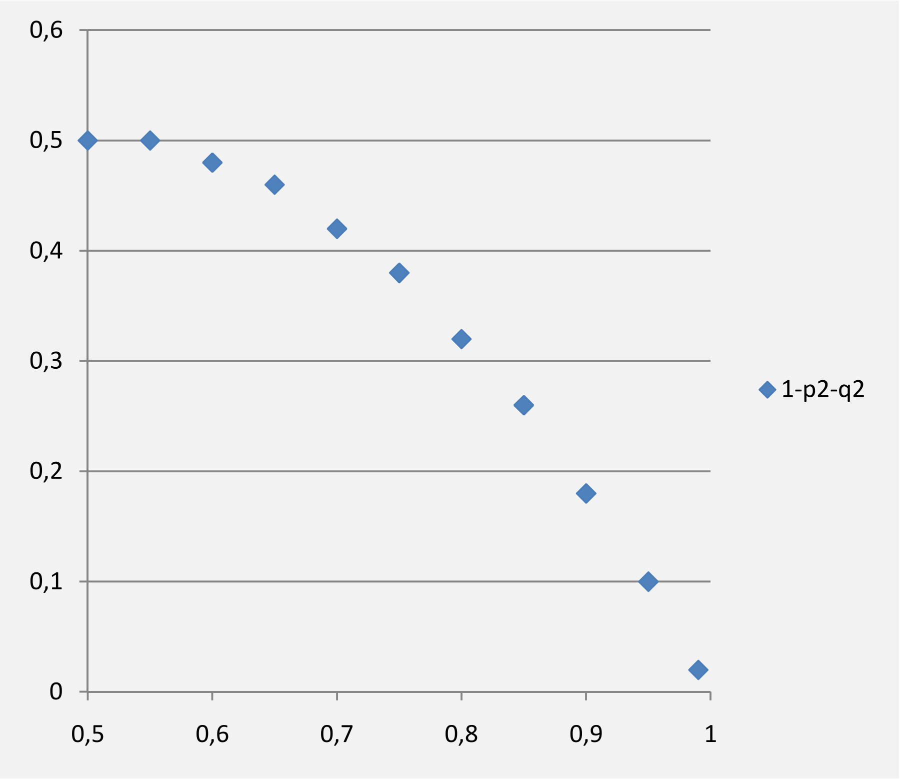
Calculate the expected heterozygosity for different values of p, p being the more common of the 2 alleles

| p | q | | | h |
|---|---|---|---|---|
| 0.5 | | | | |
| 0.6 | | | | |
| 0.7 | | | | |
| 0.8 | | | | |
| 0.9 | | | | |

# $H_j$ in a two-allele system

Calculate the expected heterozygosity for different values of p, p being the more common of two alleles

| p | q | $p^2$ | $q^2$ | $1-p^2-q^2$ |
|---|---|---|---|---|
| 0.5 | 0.5 | 0.25 | 0.25 | 0.50 |
| 0.6 | 0.4 | 0.36 | 0.16 | 0.48 |
| 0.7 | 0.3 | 0.49 | 0.09 | 0.42 |
| 0.8 | 0.2 | 0.64 | 0.04 | 0.32 |
| 0.9 | 0.1 | 0.81 | 0.01 | 0.18 |



=> between p=0.5 and p=0.75 slow change of heterozygosity, beyond more rapid decrease

# $H_j$ with more alleles

Calculate the expected heterozygosity for different numbers of alleles/locus, with equal frequencies for each allele!

| $i$ | $p_i$ | | | $h_j$ |
|---|---|---|---|---|
| 2 | 0.5 | | | |
| 4 | | | | |
| 5 | | | | |
| 10 | | | | |
| 100 | | | | |

# $H_j$ with more alleles

Calculate the expected heterozygosity for different numbers of alleles, with equal frequencies for each allele!

| i | $p_i$ | $p_i^2$ | $\Sigma p_i^2$ | $1-\Sigma p_i^2$ |
|---|---|---|---|---|
| 2 | 0.5 | 0.25 | 0.5 | 0.5 |
| 4 | 0.25 | 0.062 | 0.25 | 0.75 |
| 5 | 0.2 | 0.04 | 0.2 | 0.8 |
| 10 | 0.1 | 0.01 | 0.1 | 0.9 |
| 100 | 0.01 | 0.001 | 0.01 | 0.99 |

| $p_1$ | $p_2$ | $p_1^2$ | $p_2^2$ | $1-\Sigma p_i^2$ |
|---|---|---|---|---|
| 0.6 | 0.4 | 0.36 | 0.16 | 0.48 |
| 0.7 | 0.3 | 0.49 | 0.09 | 0.42 |
| 0.8 | 0.2 | 0.64 | 0.04 | 0.32 |
| 0.9 | 0.1 | 0.81 | 0.01 | 0.18 |

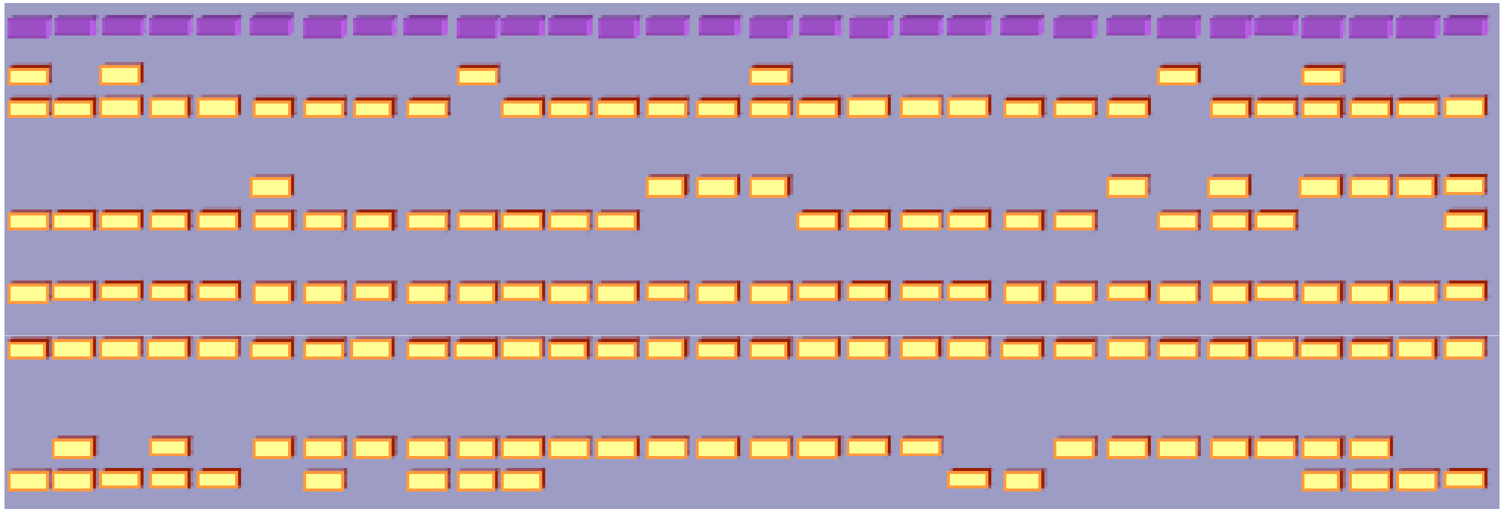
In general terms:

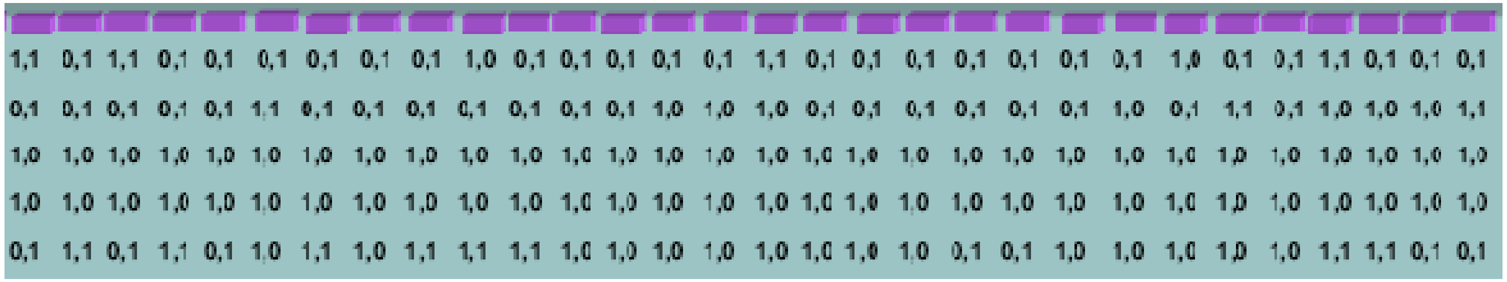
$h_{max}=1$

- ✓ More alleles at a locus mean a higher level of expected heterozygosity
- ✓ The expected heterozygosity is higher when the frequencies of the different alleles at a locus are equal (~ evenness)

# $H_O$: co-dominant data

e.g. SSR





| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Locus A** | 1,1 | 0,1 | 1,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,0 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,0 | 0,1 | 0,1 | 1,1 | 0,1 | 0,1 | 0,1 |
| **Locus B** | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,0 | 1,0 | 1,0 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 1,0 | 0,1 | 1,1 | 0,1 | 1,0 | 1,0 | 1,0 | 1,1 |
| **Locus C** | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 |
| **Locus D** | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 |
| **Locus E** | 0,1 | 1,1 | 0,1 | 1,1 | 0,1 | 1,0 | 1,1 | 1,0 | 1,1 | 1,1 | 1,1 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 0,1 | 0,1 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,1 | 1,1 | 0,1 | 0,1 |

Average observed heterozygosity $H_o = [(4/30)+(3/30)+(0/30)+(0/30)+(8/30)]/5=0.1$

# Genetic diversity: co-dominant data

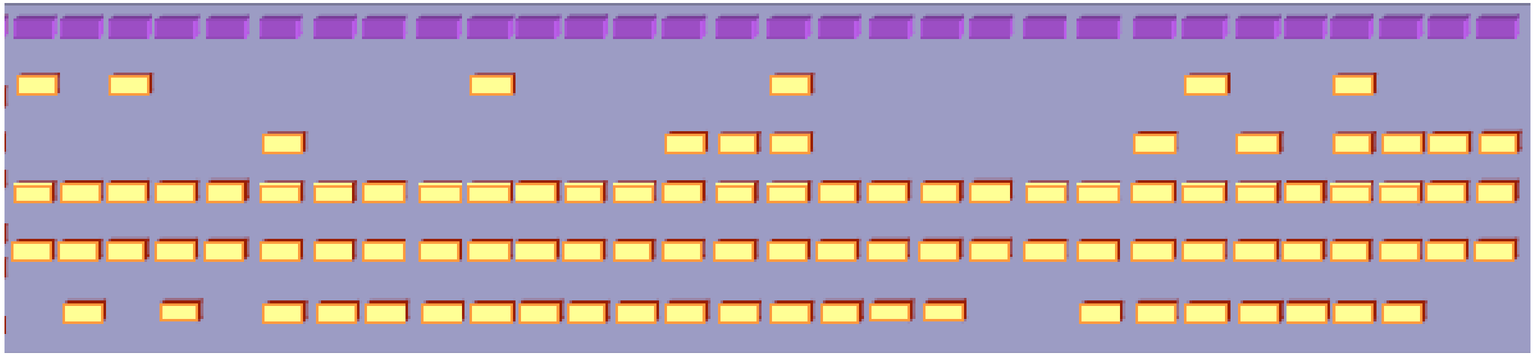
Average observed heterozygosity $H_o$ = 0.1

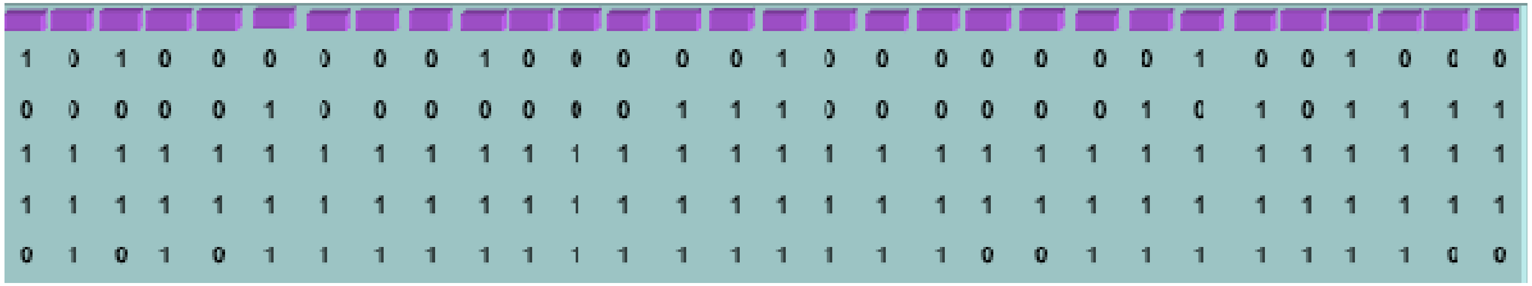
Average expected heterozygosity $H_e$

| Locus | Data analysis | | | | | allele frequency | | $H_j$ (1-$p^2$-$q^2$) | $H_e$ |
|---|---|---|---|---|---|---|---|---|---|
| A | genotypes | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | total | p | q | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 2 | 4 | 24 | 30 | | | | |
| | gen. freq. (obs.) | 0.07 | 0.13 | 0.8 | 1 | 8/60= 0.13 | 52/60= 0.87 | 0.23 | |
| B | genotypes | $B_1B_1$ | $B_1B_2$ | $B_2B_2$ | total | | | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 7 | 3 | 20 | 30 | | | | |
| | gen. freq. (obs.) | 0.23 | 0.1 | 0.67 | 1 | 17/60= 0.28 | 43/60= 0.72 | 0.41 | |
| E | genotypes | $E_1E_1$ | $E_1E_2$ | $E_2E_2$ | total | | | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 15 | 8 | 7 | 30 | | | | |
| | gen. freq. (obs.) | 0.5 | 0.27 | 0.23 | 1 | 38/60= 0.63 | 22/60= 0.37 | 0.46 | 0.22 |

# $H_O$ : dominant data

dominant data, e.g. AFLP





| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Locus A | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Locus B | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| Locus C | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Locus D | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Locus E | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

1 = fragment present in two copies (homozygote dominant) or in one copy (heterozygote)
0 = fragment absent (homozygote recessive)

with dominant data: observed heterozygosity cannot be estimated

# Genetic diversity: dominant data

Average observed heterozygosity $H_o$ ????

Average expected heterozygosity $H_e$

| Locus | Data analysis | | | | | allele frequency | | $1-p^2-q^2$ | $H_e$ |
|---|---|---|---|---|---|---|---|---|---|
| A | genotypes | AA | Aa | aa | total | p | q | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 6 | | 24 | 30 | | | | |
| | gen. freq. (obs.) | 0.2 | | 0.8 | 1 | 0.11 | 0.89 | 0.19 | |
| B | genotypes | BB | Bb | bb | total | | | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 10 | | 20 | 30 | | | | |
| | gen. freq. (obs.) | 0.33 | | 0.67 | 1 | 0.18 | 0.82 | 0.30 | |
| E | genotypes | EE | Ee | ee | total | | | | |
| | gen. freq. (exp.) | $p^2$ | 2pq | $q^2$ | 1 | | | | |
| | individuals (no.) | 23 | | 7 | 30 | | | | |
| | gen. freq. (obs.) | 0.77 | | 0.23 | 1 | 0.52 | 0.48 | 0.50 | 0.198 |

Expected heterozygosity can be calculated because we assume H-W

# Quantifying genetic variation among populations

Heterozygosity is ‘hypothetical’: refers to the probability that individuals would be heterozygous

- ✓ The concept of heterozygosity can be extended from a single population to multiple populations
- ✓ The probability that two genes at a given locus, drawn at random from two or more populations, are different (heterozygous) => heterozygosity

# Genetic differentiation

- Consider 2 populations (A and B) of the same size
- We can estimate the heterozygosity in A, in B and in the combined population (AB)
  - typically H will be higher in AB than in A or B separately

**If $p_i$ is the frequency of a given allele in the total sample of plants (AB)**, the allele frequency $p_i$ will be higher (+d) or lower (-d) in each subpopulation, with d = difference between populations

e.g., A: $p_i+d$ and B: $p_i-d$

1. Homozygosity in the total AB population = probability to draw the same allele from A and B:

$(p_i+d)(p_i-d)= p_i^2-d^2$

The average heterozygosity **<u>between</u>** the subpopulations is then

(remember $h_j = 1 – \sum p_i^2$)

$H_D=1-\Sigma p_i^2+\Sigma d^2$

2. Homozygosity within the subpopulations is

A: $(p_i+d)^2 =p_i^2+d^2+2p_id$ //// B: $(p_i-d)^2 =p_i^2+d^2-2p_id \Rightarrow$ average $p_i^2+d^2$

The average heterozygosity **<u>within</u>** the subpopulations is then

$H_S=1-\Sigma p_i^2-\Sigma d^2$

$\Rightarrow$ Heterozygosity is $2\Sigma d^2$ greater between the two populations than within them

# Genetic differentiation

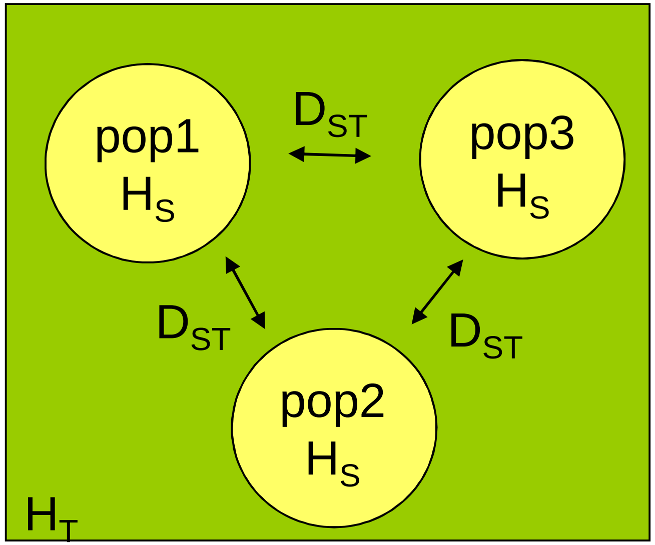
**We define then:**

Heterozygosity in the total population as: $H_T=1-\Sigma p_i^2$

Heterozygosity within the subpopulations: $H_S=1-\Sigma p_i^2-\Sigma d^2$

It follows: $H_T=H_S+\Sigma d^2$

As a result, the total genetic variation can be partitioned into within / between / among subpopulations (with d or $D_{ST}$ = the difference in diversity between populations)



# Sewall Wright‘s $F_{ST}$

Fixation index **$F_{ST}$** measures the reduction in heterozygosity (*H*) expected with non-random mating at any one level of population hierarchy relative to another more inclusive hierarchical level

$$\mathbf{F_{ST} = (H_{Total} - H_{subpop})/H_{Total}}$$

# Genetic differentiation: F statistics (Sewall Wright)

$$F_{ST}= 1 – (H_S/H_T)$$
$$F_{IT} = 1 – (H_I/H_T)$$
$$F_{IS} = 1 – (H_I/H_S)$$

with

$H_T$ = expected heterozygosity in the total population as estimated from pooled allele frequencies

$H_I$ = average observed heterozygosity in a group of populations

$H_S$ = average expected heterozygosity estimated for each subpopulation

$F_{IT}$ / $F_{IS}$ = the deficiency or excess of heterozygotes in a group of populations / each subpopulation

$F_{ST}$ = degree of gene differentiation among populations

$F_{ST}$ ranges between 0 and 1

| | |
|---|---|
| = 0 | ⇒ no genetic differentiation |
| 0 – 0.05 | ⇒ little differentiation |
| 0.05 – 0.15 | ⇒ moderate genetic differentiation |
| 0.15 – 0.25 | ⇒ large genetic differentiation |
| > 0.25 | ⇒ very large genetic differentiation |
| = 1.0 | ⇒ populations fixed for alternate/different alleles |

# Genetic differentiation: F statistics

2 populations, 1 locus with 2 alleles

F fixation index: Hexp-Hobs/Hexp

| | | Genotype frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | $p_i$ | $q_i$ | $2\ p_i\ q_i$ | F |
| | Pop 1 | 0.4 | 0.3 | 0.3 | 0.55 | 0.45 | 0.4950 | 0.3939 |
| | Pop 2 | 0.6 | 0.2 | 0.2 | 0.70 | 0.30 | 0.4200 | 0.5238 |
| expected | $\mathbf{H_T}$ | 2(0.625)(0.375) = 0.4688 | | | $\mathbf{p_O}$ | (0.55 + 0.70)/2 = 0.625 | | |
| observed | $\mathbf{H_I}$ | (0.3 + 0.2)/2 = 0.25 | | | $\mathbf{q_O}$ | (0.45 + 0.30)/2 = 0.375 | | |
| expected | $\mathbf{H_S}$ | (0.495 + 0.420)/2 = 0.4575 | | | | | | |

$$\mathbf{F_{IT} = 1 – (0.25/0.4688) = 0.4667}$$

$$\mathbf{F_{IS} = 1 – (0.25/0.4575) = 0.4536}$$

$$\mathbf{F_{ST} = 1 – (0.4575/0.4688) = 0.0241}$$

- ✓ low differentiation in allele frequencies among populations
- ✓ all the heterozygote deficit due to nonrandom mating within the populations

# Calculating genetic distances

Genetic distance can be any quantitative measure of genetic difference, be it at the sequence level or the allele frequency level that is calculated between individuals, populations or species

Refers to the genetic elements (alleles, genes, genotypes) that the two samples do not share

$$\mathbf{D = 1 - s}$$

distance D = 1 when the two samples have no genetic elements in common

similarity index s= 0 when the two samples have no genetic elements in common

Possible applications:

- ✓ establish relatedness of individuals in breeding pool? (inter-genotype similarities)
- ✓ study distance among populations? (inter-population differences)

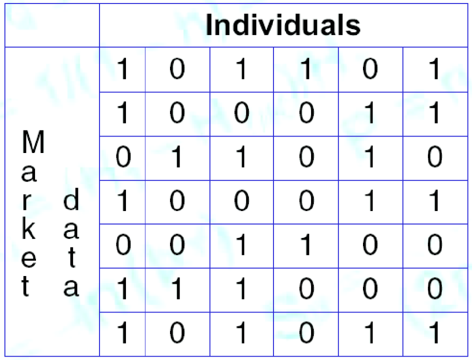
Steps:

- ✓ Calculation of genetic similarity/distance matrix
- ✓ Analysis of GS/GD matrix using clustering algorithm(s)
- ✓ Graphical presentation and interpretation

# Patterns of genetic variation: general approach

describe the diversity

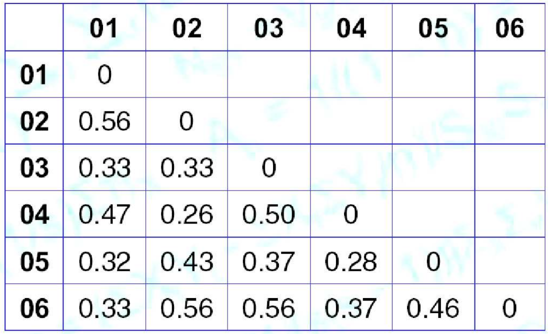
- ✓ within a population or between populations
- ✓ may extend to larger units, such as areas and regions



| | Individuals | | | | | |
|---|---|---|---|---|---|---|
| Market data | 1 | 0 | 1 | 1 | 0 | 1 |
| | 1 | 0 | 0 | 0 | 1 | 1 |
| | 0 | 1 | 1 | 0 | 1 | 0 |
| | 1 | 0 | 0 | 0 | 1 | 1 |
| | 0 | 0 | 1 | 1 | 0 | 0 |
| | 1 | 1 | 1 | 0 | 0 | 0 |
| | 1 | 0 | 1 | 0 | 1 | 1 |

calculate relationships between the entities

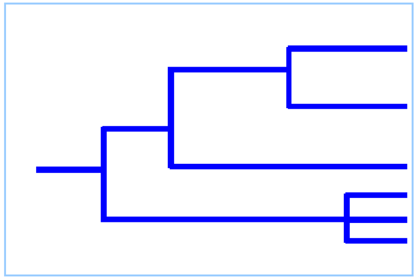
- ✓ calculate the distances (geometric or genetic) among all pairs of subjects in the study



| | **01** | **02** | **03** | **04** | **05** | **06** |
|---|---|---|---|---|---|---|
| **01** | 0 | | | | | |
| **02** | 0.56 | 0 | | | | |
| **03** | 0.33 | 0.33 | 0 | | | |
| **04** | 0.47 | 0.26 | 0.50 | 0 | | |
| **05** | 0.32 | 0.43 | 0.37 | 0.28 | 0 | |
| **06** | 0.33 | 0.56 | 0.56 | 0.37 | 0.46 | 0 |

express the relationships

- ✓ any classification and/or ordination method
- ✓ possible to compare the results of molecular study with other data (e.g. geographical)



# Genetic distance: between genotypes

## Similarity indices for dominant data

Simple Matching coefficient,
or simple concordance coefficient: **(a + d)/(a + b + c + d)**

Jaccard coefficient
(absent data are treated as missing): **a/(a + b + c)**

Nei-Li coefficient, or Dice: **2a/(2a + b + c)**

| | | Indiv. i | |
|---|---|---|---|
| | | 1 | 0 |
| Indiv.j | 1 | a | c |
| | 0 | b | d |

| individual i | | individual j | | count | condition |
|---|---|---|---|---|---|
| present | 1 | present | 1 | **a** | positive match |
| present | 1 | absent | 0 | **b** | mismatch |
| absent | 0 | present | 1 | **c** | mismatch |
| absent | 0 | absent | 0 | **d** | negative match |

# Genetic distance: between genotypes

## Similarity indices for co-dominant data

*e.g.*, Roger's distance

$$RD_{ij} = 1/2\left[\sum (X_{ai} - X_{aj})^2\right]^{1/2}$$

where:

$X_{ai}$ = frequency of allele a for individual i

= 0 if allele not present

= 0.5 if allele present in one copy

= 1 if allele present in two copies

for comparison: Euclidean distance

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

# Genetic distance: between populations

Nei‘s genetic distance $D_{xy}$ between populations i and j:

$$\mathbf{D_{xy} = -\ln (I_{xy})} \quad \text{with} \quad \mathbf{I_{xy}} = \frac{\mathbf{J_{xy}}}{\sqrt{(\mathbf{J_x J_y})}}$$

with $I_{xy}$ = genetic identity
$J_X$ = average homozygosity in population X
$J_Y$ = average homozygosity in population Y
$J_{XY}$ = average interpopulation homozygosity

# Genetic distance: calculating Nei‘s genetic distance

example: 3 populations (i), 13 loci (j) and # no. alleles/locus (k)
10 monomorphic and 3 polymorphic loci

| | | **pop1** | **pop2** | **pop3** |
|---|---|---|---|---|
| A | $A_1$ | 0.8 | 0.74 | 0.65 |
| | $A_2$ | 0.2 | 0.26 | 0.35 |
| Locus heterozygosity | $h_{ijk}$ | 0.32 | 0.3848 | 0.455 |
| B | $B_1$ | 0.86 | 0.81 | 1 |
| | $B_2$ | 0.01 | 0.1 | 0 |
| | $B_3$ | 0.13 | 0.09 | 0 |
| Locus heterozygosity | $h_{ijk}$ | 0.2434 | 0.3258 | 0 |
| D | $D_1$ | 0 | 1 | 0.3 |
| | $D_2$ | 1 | 0 | 0.7 |
| Locus heterozygosity | $h_{ijk}$ | 0 | 0 | 0.42 |
| **Average heterozygosity** | $\mathbf{H_i}$ | 0.0433 | 0.0547 | 0.0673 |
| **Average homozygosity** | $\mathbf{J_i}$ | 0.9567 | 0.9453 | 0.9327 |
| **Average interpop homozygosity** | $\mathbf{J_{ii'}}$ | $J_{1,2}$=0.8733 | $J_{1,3}$=0.9346 | $J_{2,3}$=0.8986 |
| **Genetic identity** | $\mathbf{I_{ii'}}$ | $I_{1,2}$=0.9183 | $I_{1,3}$=0.9894 | $I_{2,3}$=0.9570 |
| **Genetic distance** | $\mathbf{D_{ii'}}$ | $D_{1,2}$=0.0852 | $D_{1,3}$=0.0107 | $D_{2,3}$=0.0440 |

# Displaying relationship: cluster analysis

- Groups individuals or objects (i.e. populations) based on their similarity relationships, so that
- Objects with similar descriptions are mathematically gathered into the same cluster

1. hierarchical methods

   group similar entities (individuals or populations) together into classes, and arrange the classes into a hierarchy

   1. nearest neighbour = single linkage
   2. furtherst neighbour = complete linkage
   3. UPGMA = average linkage

2. non-hierarchical methods

   groups similar entities (individuals or populations) together into classes without hierarchical structure

   1. PCA
   2. PCO

3. model-based methods
   1. maximum likelihood
   2. Bayesian methods

# Neighbours

**simple linkage – 'nearest neighbour'**

✓minimizes the inter-group distance by taking the distance to the neighbour with the highest similarity

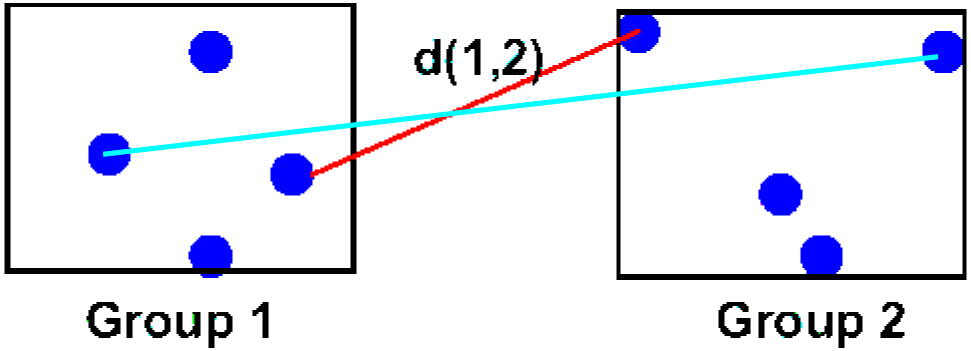
✓works with regular and compact groups, but is highly influenced by distant individuals

✓inconvenient when there are different groups that are not well distributed in (mathematical) space

**complete linkage – 'farthest neighbour'**

✓minimizes the inter-group distance by taking the distance to the individual with minimal similarity

✓works well with regular and compact groups but, again, it is influenced by distant individuals



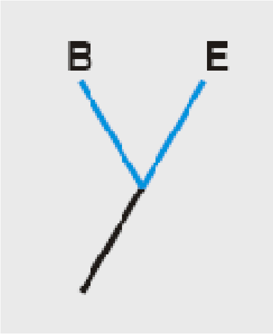
# UPGMA

UPGMA = unweigthed pair-group average using arithmetic means (average linkage)

✓minimizes the inter-group distance by taking the average pairwise distance among all individuals of the sample

✓frequently used method

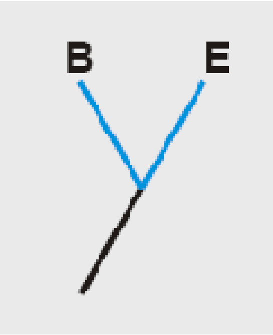
1. matrix of distances among individuals or genotypes
2. find the smallest distance; these two entities (B and E) form a first cluster

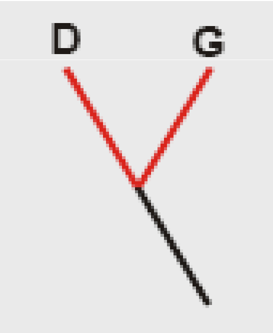


| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | - | 63 | 94 | 111 | 67 | 23 | 107 |
| B | 63 | - | 79 | 96 | 16 | 58 | 92 |
| C | 94 | 79 | - | 47 | 83 | 89 | 43 |
| D | 111 | 96 | 47 | - | 100 | 106 | 20 |
| E | 67 | **16** | 83 | 100 | - | 62 | 96 |
| F | 23 | 58 | 89 | 106 | 62 | - | 102 |
| G | 107 | 92 | 43 | 20 | 96 | 102 | - |

# UPGMA

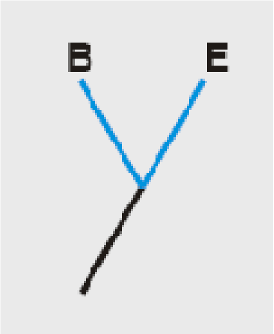
3. calculate the similarity of the newly created cluster to the rest of the entities as the the mean of the similarities of B and E
4. find the smallest distance in this matrix and merge the new entity into the cluster (DG)

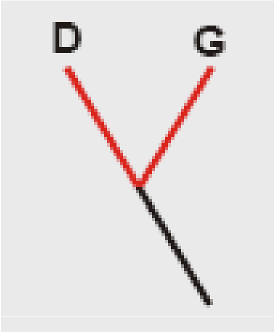




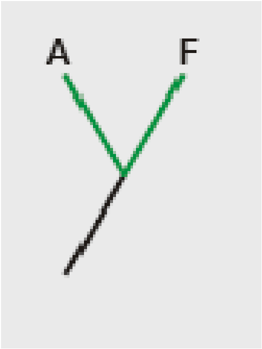
| | A | C | D | F | G | BE |
|---|---|---|---|---|---|---|
| A | - | 94 | 111 | 23 | 107 | 65 |
| C | 94 | - | 47 | 89 | 43 | 81 |
| D | 111 | 47 | - | 106 | 20 | 98 |
| F | 23 | 89 | 106 | - | 102 | 60 |
| G | 107 | 43 | **20** | 102 | - | 94 |
| BE | 65 | 81 | 98 | 60 | 94 | |

# Hierarchical clustering: UPGMA

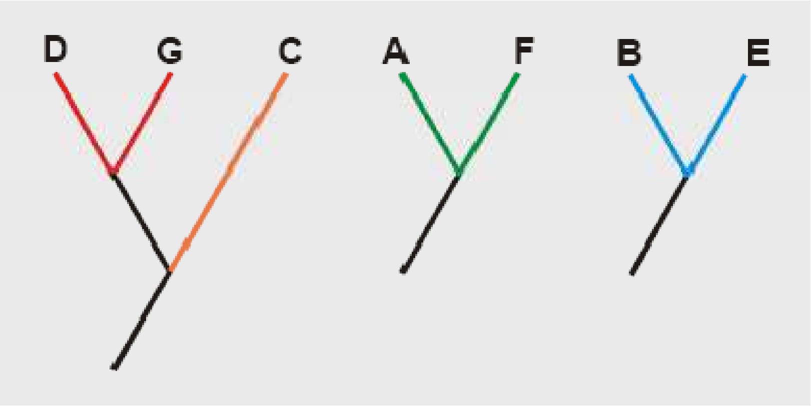




|  | A | C | F | BE | DG |
|---|---|---|---|---|---|
| A | - | 94 | 23 | 65 | 109 |
| C | 94 | - | 89 | 81 | 45 |
| F | **23** | 89 | - | 60 | 104 |
| BE | 65 | 81 | 60 |  | 96 |
| DG | 109 | 45 | 104 | 96 |  |

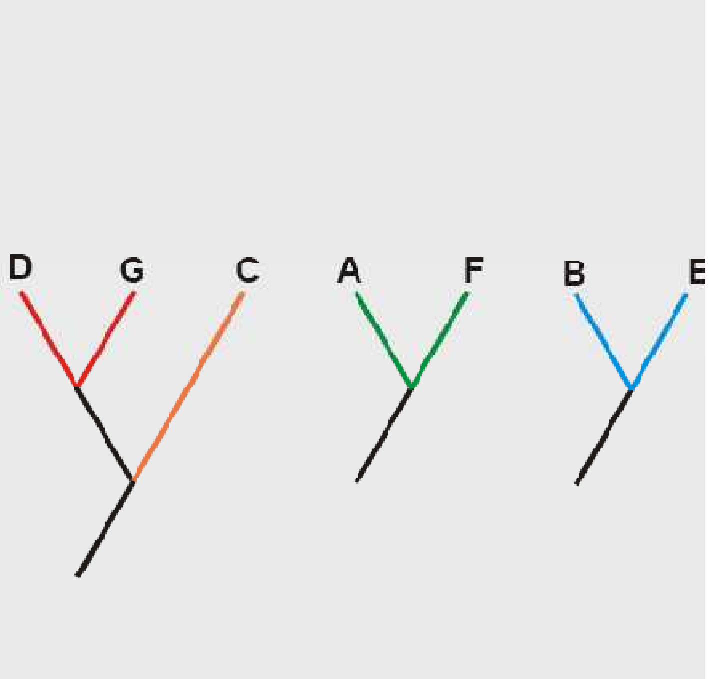


# Hierarchical clustering: UPGMA

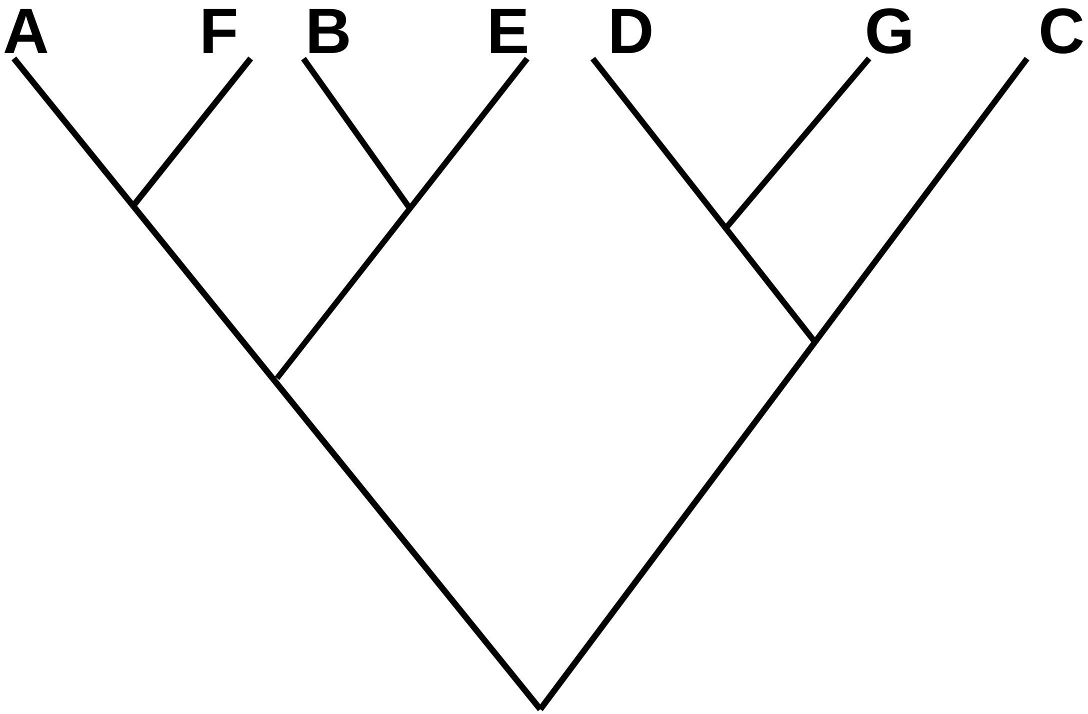


| | C | BE | DG | AF |
|---|---|---|---|---|
| C | | 81 | 34 | 92 |
| BE | 81 | | 96 | 63 |
| DG | **34** | 96 | | 107 |
| AF | 92 | 63 | 107 | |

# Hierarchical clustering: UPGMA



| | BE | DGC | AF |
|---|---|---|---|
| BE | | 91 | 63 |
| DGC | 91 | | 102 |
| AF | **63** | 102 | |



# PCA

**PCA: Principal components analysis**

To represent a multidimensional dataset (including n individuals and m characteristics) into a reduced number of dimensions (e.g. 2 or 3-dimensional plot)

PCA can be used for dimensionality reduction in a data set by retaining those characteristics of the data set that contribute most to its variance

# PCA

How?

- By linear transformation of the original m variables into a new set of uncorrelated (orthogonal) variables: principal components
- The principal components are used as a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

# PCA

1/ Calculate correlation or covariance matrix between the characters in the datamatrix

2/ Eigenanalysis of this matrix. The eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the data set

3/ Convert the original data onto this new coordinate system using the eigenvectors

# PCA

Example

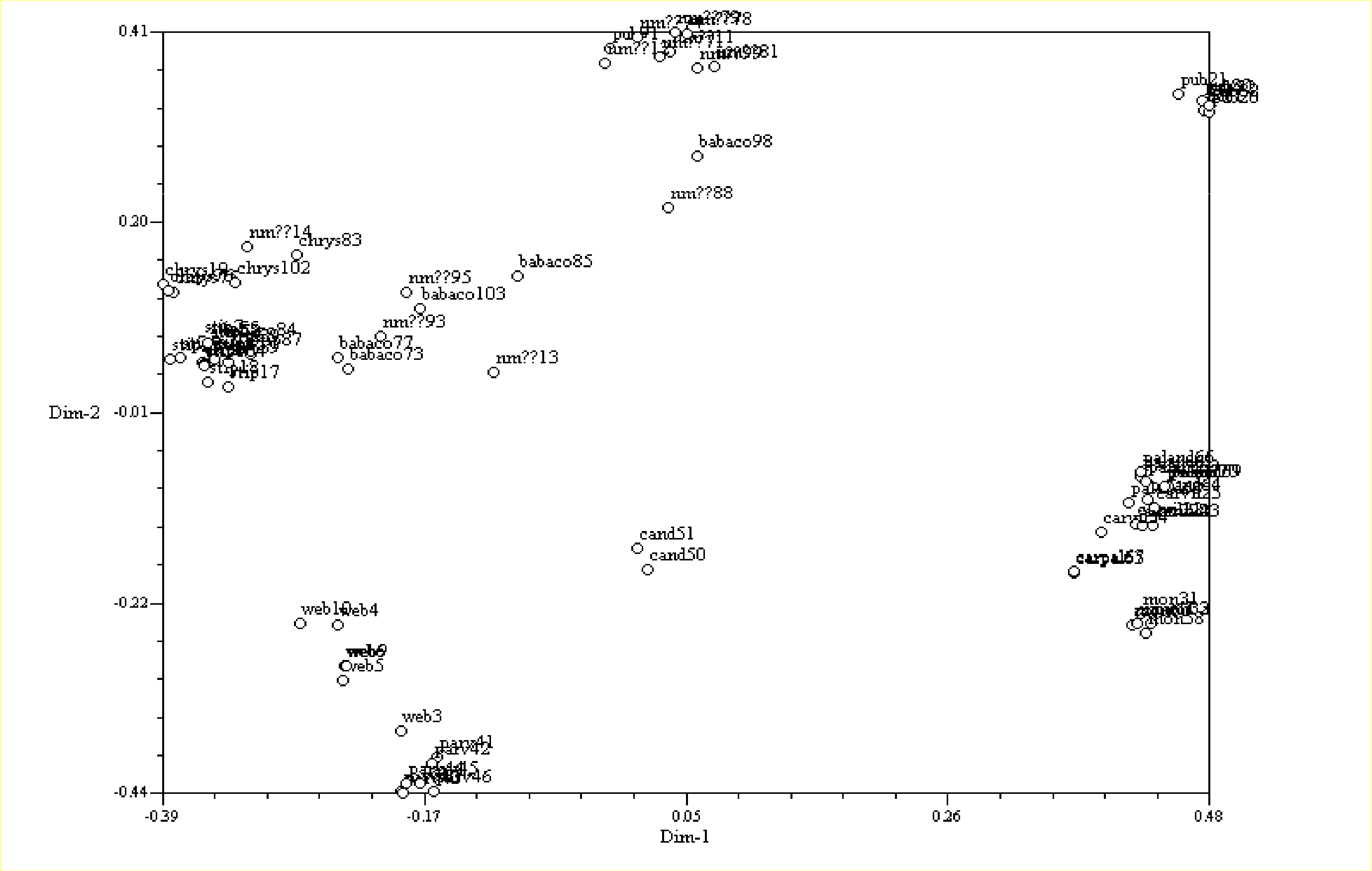
| i | Eigenvalue | Percent of variance | Cumulative |
|---|---|---|---|
| 1 | 181.93252427 | 72.4831 | 72.4831 |
| 2 | 5.49418088 | 2.1889 | 74.6720 |
| 3 | 4.60785573 | 1.8358 | 76.5078 ⇒ |
| 4 | 4.27376157 | 1.7027 | 78.2105 |
| 5 | 2.85174878 | 1.1362 | 79.3466 |
| 6 | 2.54575460 | 1.0142 | 80.3609 |
| 7 | 2.22088287 | 0.8848 | 81.2457 |

77 % of the total variance can be visualized in 3 dimensions

# PCO

**PCO:** **Principal co-ordinate analysis**

Variant of PCA, starts from a dissimilarity/distance matrix to calculate the eigenvalues



# AFLP in sweetpotato and wild relatives

- ⇒ Huang et al (2002)

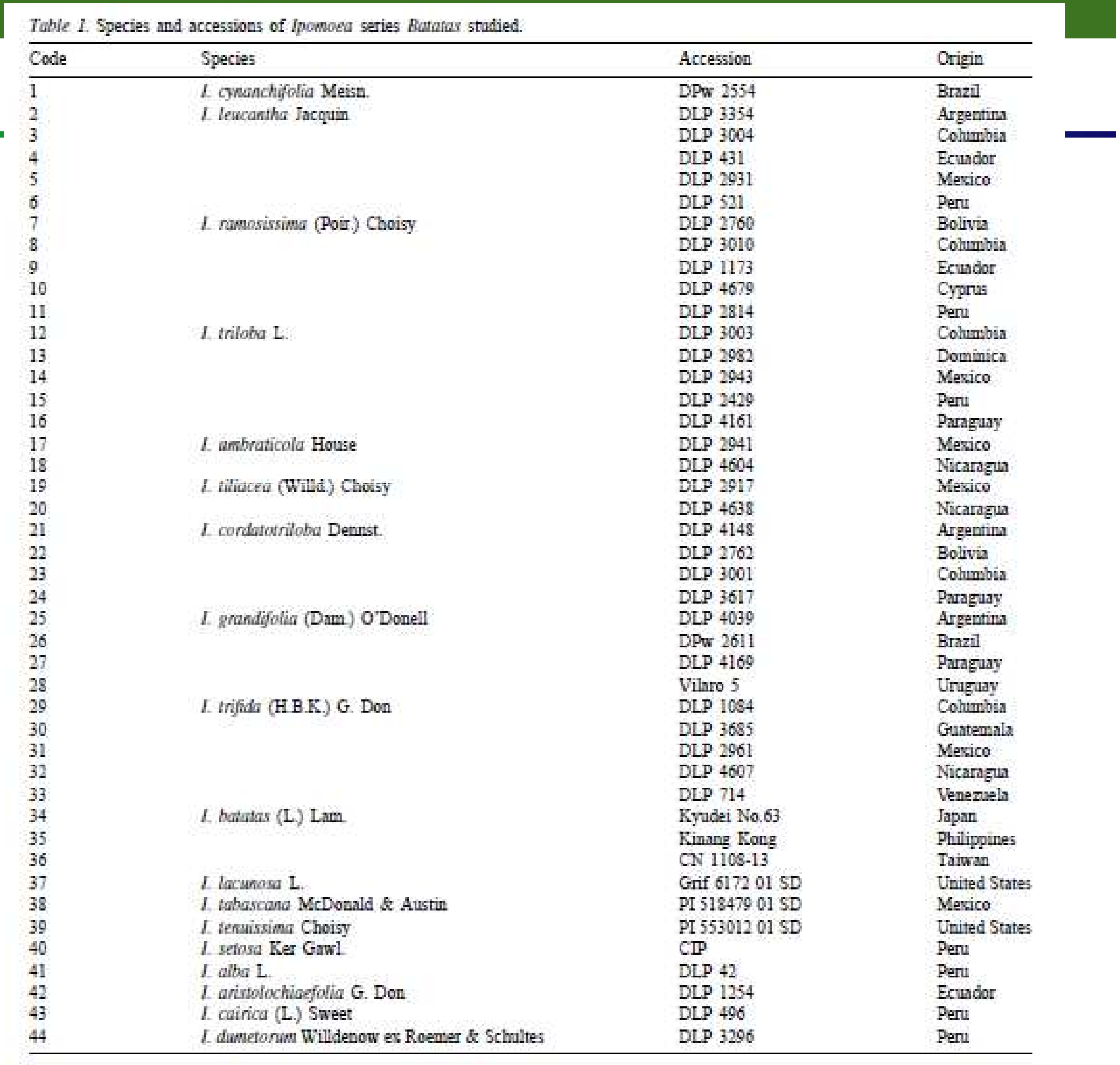


*Table 1.* Species and accessions of *Ipomoea* series *Batatas* studied.

| Code | Species | Accession | Origin |
|---|---|---|---|
| 1 | *I. cynanchifolia* Meisn. | DPw 2554 | Brazil |
| 2 | *I. leucantha* Jacquin | DLP 3354 | Argentina |
| 3 | | DLP 3004 | Columbia |
| 4 | | DLP 431 | Ecuador |
| 5 | | DLP 2931 | Mexico |
| 6 | | DLP 521 | Peru |
| 7 | *I. ramosissima* (Poir.) Choisy | DLP 2760 | Bolivia |
| 8 | | DLP 3010 | Columbia |
| 9 | | DLP 1173 | Ecuador |
| 10 | | DLP 4679 | Cyprus |
| 11 | | DLP 2814 | Peru |
| 12 | *I. triloba* L. | DLP 3003 | Columbia |
| 13 | | DLP 2982 | Dominica |
| 14 | | DLP 2943 | Mexico |
| 15 | | DLP 2429 | Peru |
| 16 | | DLP 4161 | Paraguay |
| 17 | *I. umbraticola* House | DLP 2941 | Mexico |
| 18 | | DLP 4604 | Nicaragua |
| 19 | *I. tiliacea* (Willd.) Choisy | DLP 2917 | Mexico |
| 20 | | DLP 4638 | Nicaragua |
| 21 | *I. cordatotriloba* Dennst. | DLP 4148 | Argentina |
| 22 | | DLP 2762 | Bolivia |
| 23 | | DLP 3001 | Columbia |
| 24 | | DLP 3617 | Paraguay |
| 25 | *I. grandifolia* (Dam.) O'Donell | DLP 4039 | Argentina |
| 26 | | DPw 2611 | Brazil |
| 27 | | DLP 4169 | Paraguay |
| 28 | | Vilaro 5 | Uruguay |
| 29 | *I. trifida* (H.B.K.) G. Don | DLP 1084 | Columbia |
| 30 | | DLP 3685 | Guatemala |
| 31 | | DLP 2961 | Mexico |
| 32 | | DLP 4607 | Nicaragua |
| 33 | | DLP 714 | Venezuela |
| 34 | *I. batatas* (L.) Lam. | Kyudei No.63 | Japan |
| 35 | | Kinang Kong | Philippines |
| 36 | | CN 1108-13 | Taiwan |
| 37 | *I. lacunosa* L. | Grif 6172 01 SD | United States |
| 38 | *I. tabascana* McDonald & Austin | PI 518479 01 SD | Mexico |
| 39 | *I. tenuissima* Choisy | PI 553012 01 SD | United States |
| 40 | *I. setosa* Ker Gawl. | CIP | Peru |
| 41 | *I. alba* L. | DLP 42 | Peru |
| 42 | *I. aristolochiaefolia* G. Don | DLP 1254 | Ecuador |
| 43 | *I. cairica* (L.) Sweet | DLP 496 | Peru |
| 44 | *I. dumetorum* Willdenow ex Roemer & Schultes | DLP 3296 | Peru |

# AFLP in sweetpotato and wild relatives

⇒ Huang et al (2002)

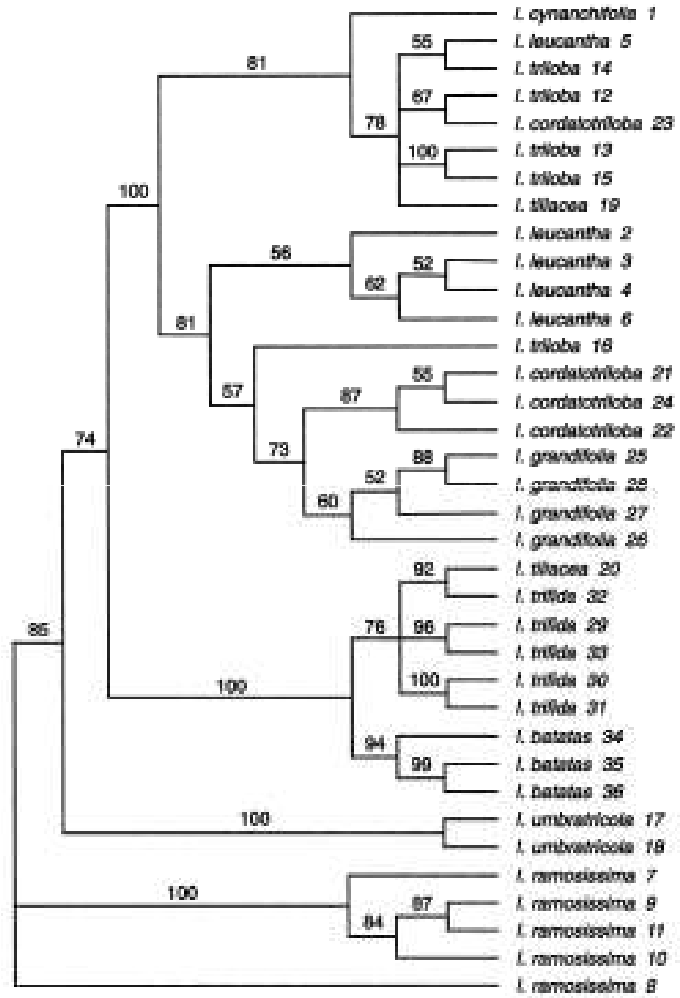


*Figure 3.* A single most parsimonious tree based on combined AFLP data set generated with all six primer combinations. Numbers above branches are bootstrap values. Numbers following each species name represent the accession code as given in Table 1.

# Management of germplasm collections

**Wheat (allohexaploid)**

⇒ Low genetic diversity at most genetic loci (using DNA-markers)

⇒ However, prone to mutation and easy to cross with other species (many disease resistances obtained by inter-specific crosses)

⇒ High phenotypic diversity revealed and exploited by farmers and breeders worldwide

⇒ Balfourier *et al* (2007)

  ⇒ INRA Clermont-Ferrand collection of more than 10,000 accessions of hexaploid wheat

  ⇒ Morphologically well-characterized

  ⇒ Is it necessary to keep all these accessions or can we preserve the same amount of genetic diversity with a smaller number of plants?

  ⇒ DNA-markers (SSRs) can assist to create a **Core collection**

**A core collection is a subset of a larger germplasm collection that contains the maximum possible genetic diversity of the species with a minimum of repetitiveness**

- Several possibilities
- M strategy: genetic markers are used to sample the collection while maximizing **allele richness at each marker locus**

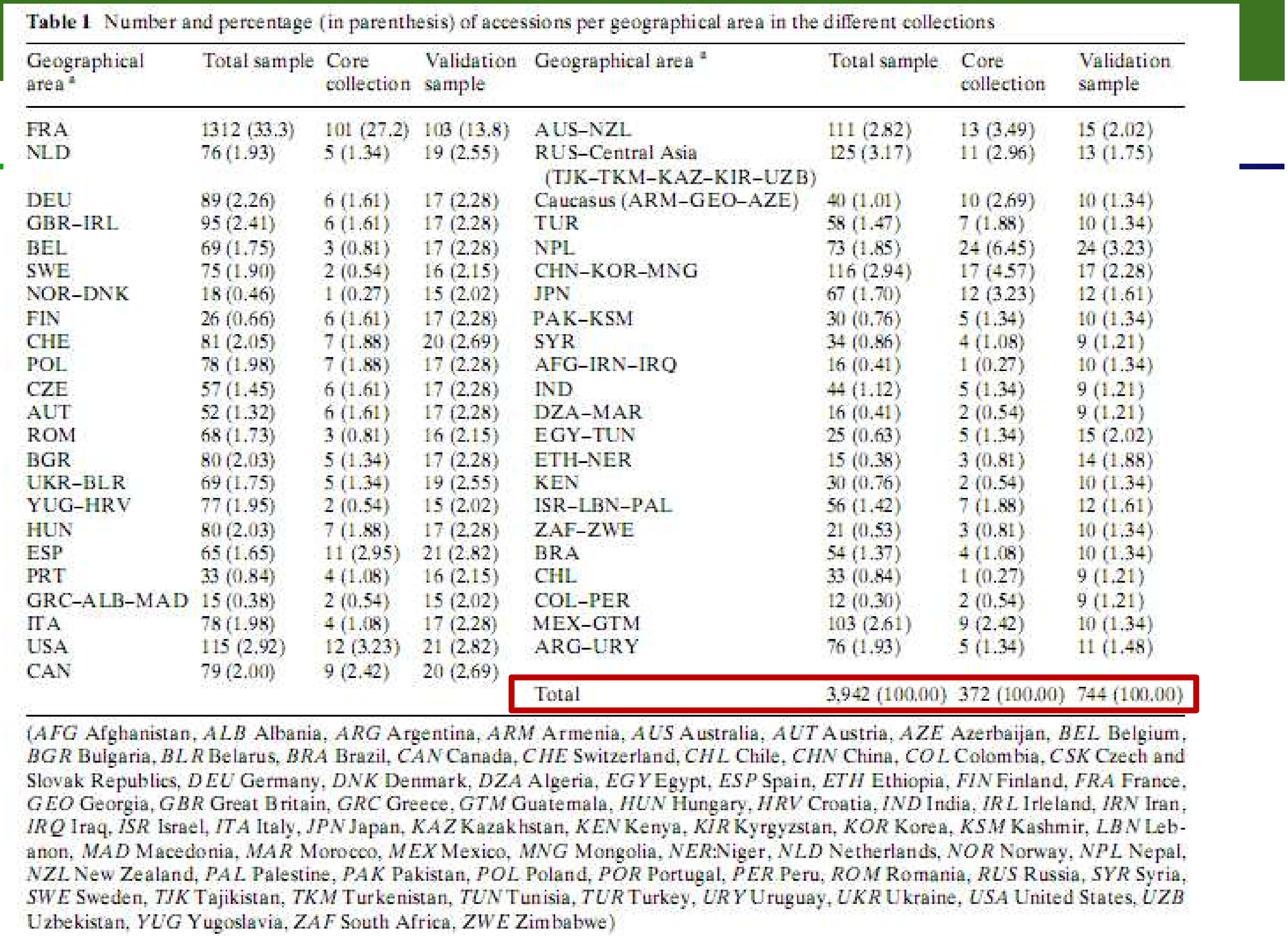


**Table 1** Number and percentage (in parenthesis) of accessions per geographical area in the different collections

| Geographical area [a] | Total sample | Core collection | Validation sample | Geographical area [a] | Total sample | Core collection | Validation sample |
|---|---|---|---|---|---|---|---|
| FRA | 1312 (33.3) | 101 (27.2) | 103 (13.8) | AUS–NZL | 111 (2.82) | 13 (3.49) | 15 (2.02) |
| NLD | 76 (1.93) | 5 (1.34) | 19 (2.55) | RUS–Central Asia (TJK–TKM–KAZ–KIR–UZB) | 125 (3.17) | 11 (2.96) | 13 (1.75) |
| DEU | 89 (2.26) | 6 (1.61) | 17 (2.28) | Caucasus (ARM–GEO–AZE) | 40 (1.01) | 10 (2.69) | 10 (1.34) |
| GBR–IRL | 95 (2.41) | 6 (1.61) | 17 (2.28) | TUR | 58 (1.47) | 7 (1.88) | 10 (1.34) |
| BEL | 69 (1.75) | 3 (0.81) | 17 (2.28) | NPL | 73 (1.85) | 24 (6.45) | 24 (3.23) |
| SWE | 75 (1.90) | 2 (0.54) | 16 (2.15) | CHN–KOR–MNG | 116 (2.94) | 17 (4.57) | 17 (2.28) |
| NOR–DNK | 18 (0.46) | 1 (0.27) | 15 (2.02) | JPN | 67 (1.70) | 12 (3.23) | 12 (1.61) |
| FIN | 26 (0.66) | 6 (1.61) | 17 (2.28) | PAK–KSM | 30 (0.76) | 5 (1.34) | 10 (1.34) |
| CHE | 81 (2.05) | 7 (1.88) | 20 (2.69) | SYR | 34 (0.86) | 4 (1.08) | 9 (1.21) |
| POL | 78 (1.98) | 7 (1.88) | 17 (2.28) | AFG–IRN–IRQ | 16 (0.41) | 1 (0.27) | 10 (1.34) |
| CZE | 57 (1.45) | 6 (1.61) | 17 (2.28) | IND | 44 (1.12) | 5 (1.34) | 9 (1.21) |
| AUT | 52 (1.32) | 6 (1.61) | 17 (2.28) | DZA–MAR | 16 (0.41) | 2 (0.54) | 9 (1.21) |
| ROM | 68 (1.73) | 3 (0.81) | 16 (2.15) | EGY–TUN | 25 (0.63) | 5 (1.34) | 15 (2.02) |
| BGR | 80 (2.03) | 5 (1.34) | 17 (2.28) | ETH–NER | 15 (0.38) | 3 (0.81) | 14 (1.88) |
| UKR–BLR | 69 (1.75) | 5 (1.34) | 19 (2.55) | KEN | 30 (0.76) | 2 (0.54) | 10 (1.34) |
| YUG–HRV | 77 (1.95) | 2 (0.54) | 15 (2.02) | ISR–LBN–PAL | 56 (1.42) | 7 (1.88) | 12 (1.61) |
| HUN | 80 (2.03) | 7 (1.88) | 17 (2.28) | ZAF–ZWE | 21 (0.53) | 3 (0.81) | 10 (1.34) |
| ESP | 65 (1.65) | 11 (2.95) | 21 (2.82) | BRA | 54 (1.37) | 4 (1.08) | 10 (1.34) |
| PRT | 33 (0.84) | 4 (1.08) | 16 (2.15) | CHL | 33 (0.84) | 1 (0.27) | 9 (1.21) |
| GRC–ALB–MAD | 15 (0.38) | 2 (0.54) | 15 (2.02) | COL–PER | 12 (0.30) | 2 (0.54) | 9 (1.21) |
| ITA | 78 (1.98) | 4 (1.08) | 17 (2.28) | MEX–GTM | 103 (2.61) | 9 (2.42) | 10 (1.34) |
| USA | 115 (2.92) | 12 (3.23) | 21 (2.82) | ARG–URY | 76 (1.93) | 5 (1.34) | 11 (1.48) |
| CAN | 79 (2.00) | 9 (2.42) | 20 (2.69) | | | | |
| | | | | Total | 3,942 (100.00) | 372 (100.00) | 744 (100.00) |

(*AFG* Afghanistan, *ALB* Albania, *ARG* Argentina, *ARM* Armenia, *AUS* Australia, *AUT* Austria, *AZE* Azerbaijan, *BEL* Belgium, *BGR* Bulgaria, *BLR* Belarus, *BRA* Brazil, *CAN* Canada, *CHE* Switzerland, *CHL* Chile, *CHN* China, *COL* Colombia, *CSK* Czech and Slovak Republics, *DEU* Germany, *DNK* Denmark, *DZA* Algeria, *EGY* Egypt, *ESP* Spain, *ETH* Ethiopia, *FIN* Finland, *FRA* France, *GEO* Georgia, *GBR* Great Britain, *GRC* Greece, *GTM* Guatemala, *HUN* Hungary, *HRV* Croatia, *IND* India, *IRL* Ireland, *IRN* Iran, *IRQ* Iraq, *ISR* Israel, *ITA* Italy, *JPN* Japan, *KAZ* Kazakhstan, *KEN* Kenya, *KIR* Kyrgyzstan, *KOR* Korea, *KSM* Kashmir, *LBN* Lebanon, *MAD* Macedonia, *MAR* Morocco, *MEX* Mexico, *MNG* Mongolia, *NER*:Niger, *NLD* Netherlands, *NOR* Norway, *NPL* Nepal, *NZL* New Zealand, *PAL* Palestine, *PAK* Pakistan, *POL* Poland, *POR* Portugal, *PER* Peru, *ROM* Romania, *RUS* Russia, *SYR* Syria, *SWE* Sweden, *TJK* Tajikistan, *TKM* Turkenistan, *TUN* Tunisia, *TUR* Turkey, *URY* Uruguay, *UKR* Ukraine, *USA* United States, *UZB* Uzbekistan, *YUG* Yugoslavia, *ZAF* South Africa, *ZWE* Zimbabwe)

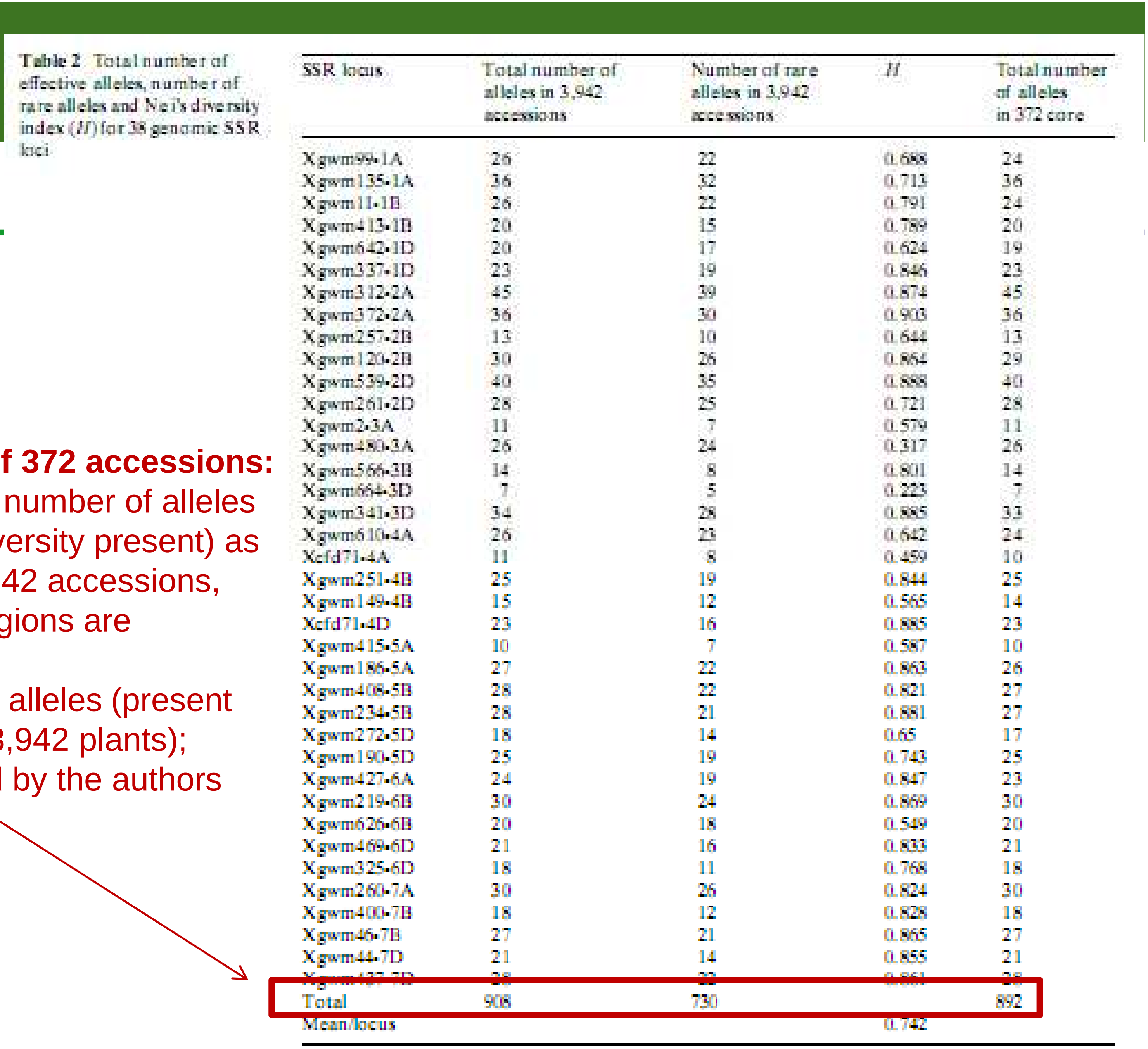
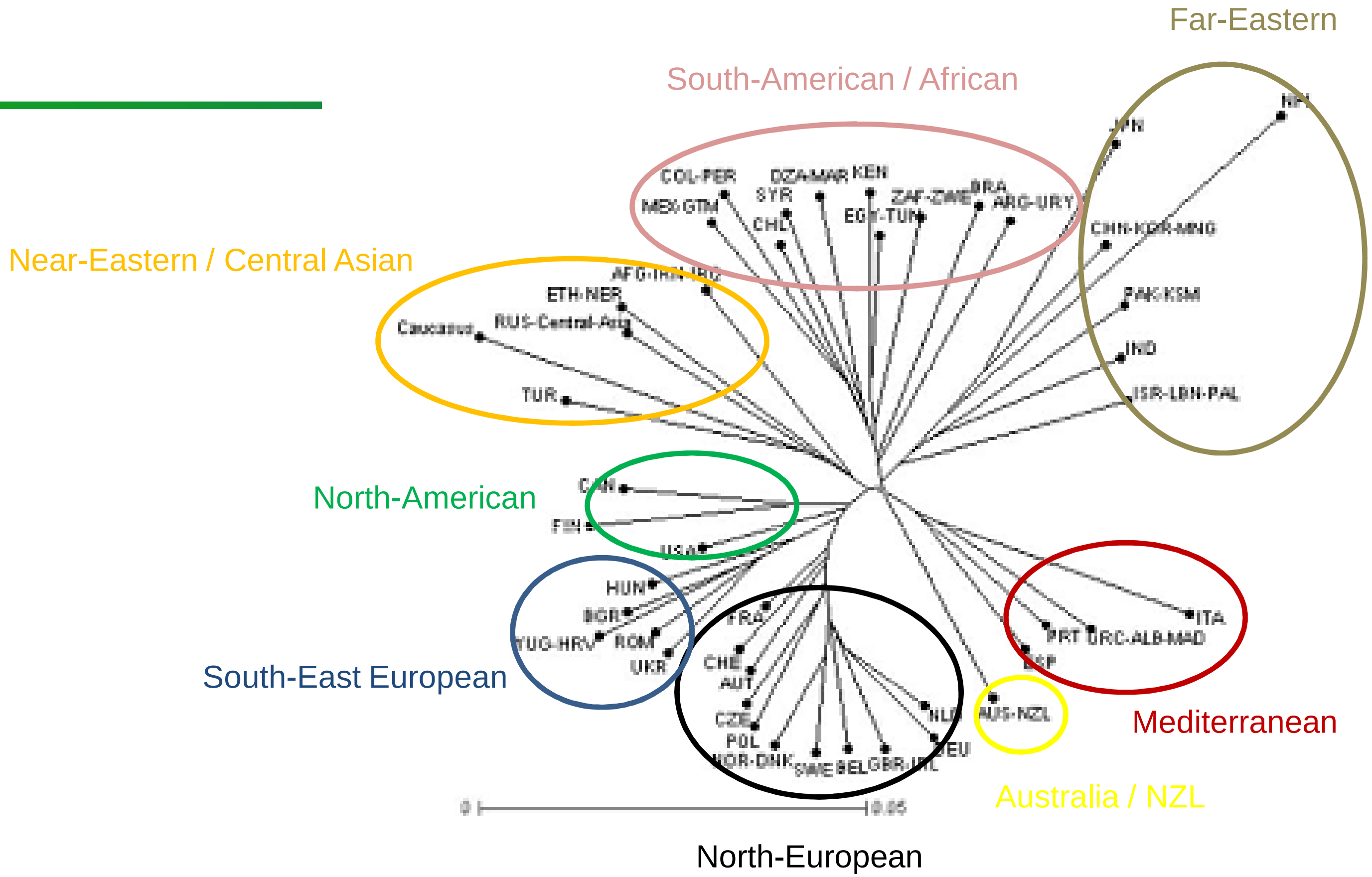


**Table 2** Total number of effective alleles, number of rare alleles and Nei's diversity index (*H*) for 38 genomic SSR loci

| SSR locus | Total number of alleles in 3,942 accessions | Number of rare alleles in 3,942 accessions | *H* | Total number of alleles in 372 core |
|---|---|---|---|---|
| Xgwm99-1A | 26 | 22 | 0.688 | 24 |
| Xgwm135-1A | 36 | 32 | 0.713 | 36 |
| Xgwm11-1B | 26 | 22 | 0.791 | 24 |
| Xgwm413-1B | 20 | 15 | 0.789 | 20 |
| Xgwm642-1D | 20 | 17 | 0.624 | 19 |
| Xgwm337-1D | 23 | 19 | 0.846 | 23 |
| Xgwm312-2A | 45 | 39 | 0.874 | 45 |
| Xgwm372-2A | 36 | 30 | 0.903 | 36 |
| Xgwm257-2B | 13 | 10 | 0.644 | 13 |
| Xgwm120-2B | 30 | 26 | 0.864 | 29 |
| Xgwm539-2D | 40 | 35 | 0.898 | 40 |
| Xgwm261-2D | 28 | 25 | 0.721 | 28 |
| Xgwm2-3A | 11 | 7 | 0.579 | 11 |
| Xgwm480-3A | 26 | 24 | 0.317 | 26 |
| Xgwm566-3B | 14 | 8 | 0.801 | 14 |
| Xgwm664-3D | 7 | 5 | 0.223 | 7 |
| Xgwm341-3D | 34 | 28 | 0.885 | 33 |
| Xgwm610-4A | 26 | 23 | 0.642 | 24 |
| Xcfd71-4A | 11 | 8 | 0.459 | 10 |
| Xgwm251-4B | 25 | 19 | 0.844 | 25 |
| Xgwm149-4B | 15 | 12 | 0.565 | 14 |
| Xcfd71-4D | 23 | 16 | 0.885 | 23 |
| Xgwm415-5A | 10 | 7 | 0.587 | 10 |
| Xgwm186-5A | 27 | 22 | 0.863 | 26 |
| Xgwm408-5B | 28 | 22 | 0.821 | 27 |
| Xgwm234-5B | 28 | 21 | 0.881 | 27 |
| Xgwm272-5D | 18 | 14 | 0.65 | 17 |
| Xgwm190-5D | 25 | 19 | 0.743 | 25 |
| Xgwm427-6A | 24 | 19 | 0.847 | 23 |
| Xgwm219-6B | 30 | 24 | 0.869 | 30 |
| Xgwm626-6B | 20 | 18 | 0.549 | 20 |
| Xgwm469-6D | 21 | 16 | 0.833 | 21 |
| Xgwm325-6D | 18 | 11 | 0.768 | 18 |
| Xgwm260-7A | 30 | 26 | 0.824 | 30 |
| Xgwm400-7B | 18 | 12 | 0.828 | 18 |
| Xgwm46-7B | 27 | 21 | 0.865 | 27 |
| Xgwm44-7D | 21 | 14 | 0.855 | 21 |
| Xgwm437-7D | 28 | 22 | 0.861 | 28 |
| Total | 908 | 730 | | 892 |
| Mean/locus | | | 0.742 | |

**The core collection of 372 accessions:**

1. contains the same number of alleles (estimate of the diversity present) as the collection of 3942 accessions,
2. all geographical regions are represented
3. contains all unique alleles (present only in one of the 3,942 plants); restriction imposed by the authors



**SSR- based genetic relationships among geographical origins for the 372 accessions included in the core collection**