Genetic diversity analysis with molecular marker data: Learning module

Measures of genetic diversity

Contents

- Basic genetic diversity analysis
- Types of variables
- Quantifying genetic diversity:
  - Measuring intrapopulation genetic diversity
  - Measuring interpopulation genetic diversity
- Quantifying genetic relationships:
  - Diversity and differentiation at the nucleotide level
  - Genetic distance
- Displaying relationships:
  - Classification or clustering
  - Ordination
- Appendices

Basic genetic diversity analysis

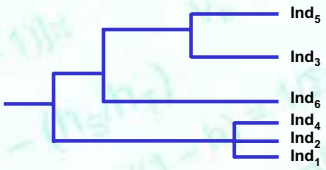
1. Description of variation within and between populations, regions, etc.

	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

2. Assessment of relationships among individuals, populations, regions, etc.

	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

3. Expression of relationships between results obtained from different sets of characters



Most of the genetic diversity analysis that we might want to do will involve the following steps:

1. *Describing the diversity.* This may be done within a population or between populations. It may also extend to larger units such as areas and regions.

2. *Calculating the relationships between the units analysed in step one.* This entails calculating the distances (geometric or genetic) among all pairs of subjects in the study.

3. *Expressing these relationships with any classification and/or ordination method at hand.* Some of these methods will permit comparing the results of our molecular study with other types of data (e.g. geographical). In the slide, Ind_1, Ind_2, … may represent, instead of individuals, populations or regions.

Types of variable

- Qualitative. These refer to characters or qualities, and are either binary or categorical:
  - Binary, taking only two values: present (1) or absent (0)
  - Categorical, taking a value among many possibilities, and are either ordinal or nominal:
    - Ordinal: categories that have an order
    - Nominal: categories that are unrelated
- Quantitative. These are numerical and are either continuous or discrete:
  - Continuous, taking a value within a given range
  - Discrete, taking whole or decimal numbers

Examples of qualitative variables:
- Binary: e.g. leaf hairiness: present (1), absent (0)
- Categorical:
  - Ordinal: e.g. stalk hairs: rare (1), common (2), abundant (3) or petiole length: short (1), medium (2), long (3)
  - Nominal: e.g. petal colour: yellow (1), red (2), white (3), purple (4)

Examples of quantitative variables:
- Continuous: e.g. root weight (g); leaf length (cm)
- Discrete: e.g. number of stamens: 2, 3, 4, ...
  number of fruits: 1, 2, 3, …

Categorical variables can be converted to binary variables, but with limitations because, as we will see later, some similarity coefficients give a weight to the category of a character and this may bias against other characters being evaluated. That is, the more categories a variable has, the more weight it has when combined with other binary or categorical variables with few categories.

An example of converting a categorial variable into a binary one:
Petiole length: short (1), medium (2), long (3)
- Short: present (1), absent (0)
- Medium: present (1), absent (0)
- Long: present (1), absent (0)

Quantitative variables can also be converted to binary variables, for example:
From 0 to 3 fruits: present (1), absent (0)
From 4 to 7 fruits: present (1), absent (0), ...

Quantifying genetic diversity: measuring intrapopulation genetic diversity

- Based on the number of variants
  - Polymorphism or rate of polymorphism (Pj)
  - Proportion of polymorphic loci
  - Richness of allelic variants (A)
  - Average number of alleles per locus
- Based on the frequency of variants
  - Effective number of alleles (A_e)
  - Average expected heterozygosity (H_e; Nei's genetic diversity)

Polymorphism or rate of polymorphism (Pj)

A gene is defined as polymorphic if the frequency of one of its alleles is less than or equal to 0.95 or 0.99

$Pj = q \leq 0.95$ or **$Pj = q \leq 0.99$**

Where,

Pj = rate of polymorphism

q = allele frequency

- This measure provides criteria to demonstrate that a gene is showing variation.
- Its calculation is through direct observation of whether the definition is fulfilled.
- It can be used with codominant markers and, very restrictively, with dominant markers. This is because the estimate based on dominant markers would be biased below the real number.

A polymorphic gene is usually one for which the most common allele has a frequency of less than 0.95. Rare alleles are defined as those with frequencies of less than 0.005. The limit of allele frequency, which is set at 0.95 (or 0.99) is arbitrary, its objective being to help identify those genes in which allelic variation is common.

Reference

Cavalli–Sforza, L.L. and W.F. Bodmer. 1981. Genética de las Poblaciones Humanas. Ed. Omega, Barcelona.

Proportion of polymorphic loci

This the number of polymorphic loci divided by the total number of loci (polymorphic and monomorphic), that is:

$$P = n_{pj}/n_{total}$$

Where,

P = proportion of polymorphic loci

n_{pj} = number of polymorphic loci

n_{total} = total number of loci

- It expresses the percentage of variable loci in a population.
- Its calculation is based on directly counting polymorphic and total loci.
- It can be used with codominant markers and, very restrictively, with dominant markers (see previous slide for explanation).

Richness of allelic variants (A)

- Refers to the number of variants in a sample
- The measure of diversity is (A - 1) variants because, within a monomorphic population, the degree of diversity is zero (A - 1 = 0)

- For a given gene in a sample, this measure tells how many allelic variants can be found.
- It is sensitive to sample size.
- Although the distribution of alleles does not matter, the maximum number of alleles does.
- The measure can only be applied with codominant markers.

Average number of alleles per locus

It is the sum of all detected alleles in all loci, divided by the total number of loci

$$n = (1/K)\sum_{i=1}^{K} n_i$$

Where,

K = the number of loci

n_i = the number of alleles detected per locus

- This measure provides complementary information to that of polymorphism.
- It requires only counting the number of alleles per locus and then calculating the average.
- It is best applied with codominant markers, because dominant markers do not permit the detection of all alleles.

Effective number of alleles (A_e)

It is the number of alleles that can be present in a population

$$A_e = 1/(1 - h) = 1/\Sigma p_i^2$$

Where,

p_i = frequency of the i^{th} allele in a locus

$h = 1 - \Sigma p_i^2$ = heterozygosity in a locus

- The measure tells about the number of alleles that would be expected in a locus in each population.
- It is calculated by inverting the measure of homozygosity in a locus.
- It can be used with codominant marker data.
- Its calculation may be affected by sample size.

This measure of diversity may be informative for establishing collecting strategies. For example, we estimate it in a given sample. We then verify it in a different sample or the entire collection. If the figure obtained the second time is less than the first estimated number, this could mean that our collecting strategy needs revising.

Calculating the A_e: an example

Loci (A, B, C)	Population 1			Population 2		
Individual 1	A_1 A_1	B_1 B_1	C_1 C_1	A_1 A_1	B_1 B_3	C_1 C_1
Individual 2	A_1 A_2	B_1 B_2	C_2 C_2	A_1 A_1	B_2 B_3	C_1 C_1
Individual 3	A_1 A_1	B_1 B_1	C_1 C_3	A_2 A_2	B_1 B_4	C_1 C_1
Individual 4	A_1 A_3	B_1 B_3	C_2 C_3	A_2 A_2	B_1 B_1	C_1 C_1
Individual 5	A_3 A_3	B_3 B_3	C_3 C_3	A_1 A_2	B_4 B_4	C_1 C_1
Number of alleles	3	3	3	2	4	1
Frequency of allele 1	0.60	0.60	0.30	0.50	0.40	1.00
Frequency of allele 2	0.10	0.10	0.30	0.50	0.10	0.00
Frequency of allele 3	0.30	0.30	0.40	—	0.20	0.00
Frequency of allele 4	—	—	—	—	0.30	—
Heterozygosity (h)	0.54	0.54	0.66	0.50	0.70	0.00
Effective number of alleles	**2.17**	**2.17**	**2.94**	**2.00**	**3.33**	**1.00**

The table on the slide shows an example for calculating the effective number of alleles. The two populations each have 5 individuals. For each individual, 3 loci are analysed, each with a different number of alleles, depending also on the population (locus A has 3 alleles in population 1 and only 2 alleles in population 2, and so on). First, allele frequencies are calculated for each locus and each population; then, heterozygosity in each locus; and finally, the A_e, according to the formula shown in the previous slide.

Average expected heterozygosity (H_e) (Nei's genetic diversity [D])

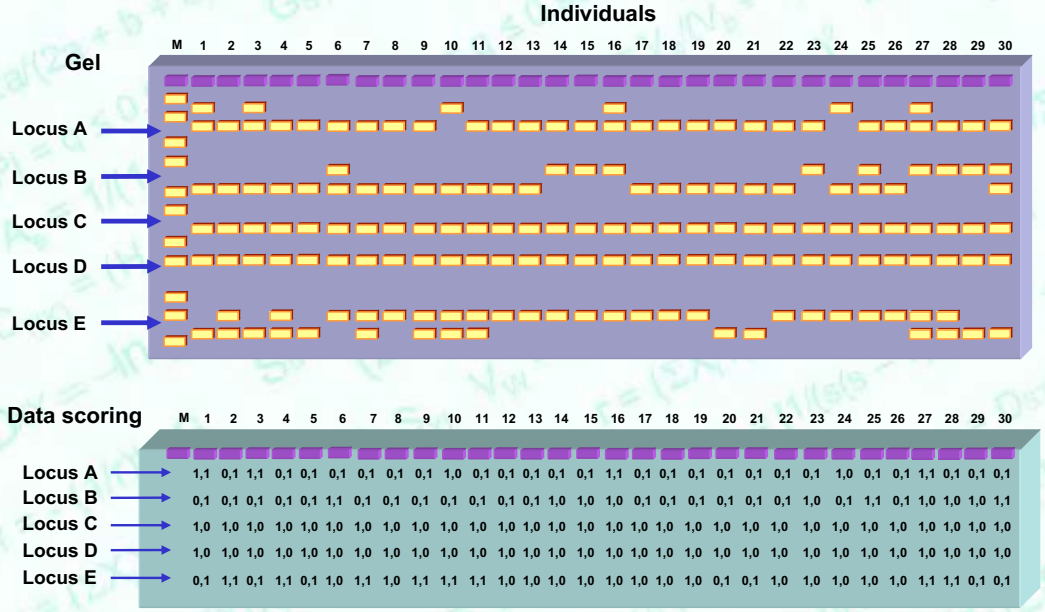
- It is the probability that, at a single locus, any two alleles, chosen at random from the population, are different to each other
- Three calculations are possible:
  - A locus with two alleles: $h_j = 1 - p^2 - q^2$
  - A locus j with i alleles: $h_j = 1 - \Sigma p_i^2$
  - Average for several loci: $H = \Sigma_j^L hj/L$
- The average H_e over all loci is an estimate of the extent of genetic variability in the population

Where,

h_j = heterozygosity per locus
p and q = allele frequencies
H = average heterozygosity for several loci
L = total number of loci

- The average expected heterozygosity is calculated by substracting from 1 the expected frequencies of homozygotes in a locus. The operation is repeated for all loci and the average then performed.
- It can be applied to all markers, both codominants and dominants.
- The estimated value may be affected by those alleles present at higher frequencies.
- It ranges from 0 to 1.
- It is maximized when there are many alleles at equal frequencies.
- A minimum of 30 loci in 20 individuals per population should be analysed to reduce the risk of statistical bias.

Calculating diversity with a codominant molecular marker



Data scoring

	M	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
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

(continued on next slide)

The top half of this slide shows a drawing of a gel with a size marker on the left (M) and 30 individuals analysed with a codominant marker, which detected five loci (A, B, C, D and E). Of these loci, only three are polymorphic (A, B and E).

The bottom half of the slide shows the results of scoring bands per individual and per locus. Note that, for ease of presentation, no more than two alleles per locus were depicted. Although the bands belonging to loci C and D were scored (1,0) for all individuals, scoring was not necessary, because the bands did not yield information on diversity.

See next slide for calculations.

Calculating diversity with a codominant molecular marker (continued)

Locus	Data analysis					Allele freq.		$h_j = (1 - p^2 - q^2)$	H_i
A	Genotypes	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	Total	p	q		
	Gen. freq. (exp.)	p^2	2pq	q^2	1				
	Individuals (no.)	2	4	24	30				
	Gen. freq. (obs.)	$P_{11} = 0.07$	$P_{12} = 0.13$	$P_{22} = 0.80$	1	0.13	0.87	0.23	
B	Genotypes	$B_1 B_1$	$B_1 B_2$	$B_2 B_2$	Total	p	q		
	Gen. freq. (exp.)	p^2	2pq	q^2	1				
	Individuals (no.)	7	3	20	30				
	Gen. freq. (obs.)	$P_{11} = 0.23$	$P_{12} = 0.10$	$P_{22} = 0.67$	1	0.28	0.72	0.41	
E	Genotypes	$E_1 E_1$	$E_1 E_2$	$E_2 E_2$	Total	p	q		
	Gen. freq. (exp.)	p^2	2pq	q^2	1				
	Individuals (no.)	15	8	7	30				
	Gen. freq. (obs.)	$P_{11} = 0.50$	$P_{12} = 0.27$	$P_{22} = 0.23$	1	0.63	0.37	0.46	0.22

1. First, we note that loci A, B and E are polymorphic because they fulfill the requirement of having allele frequencies below 0.99. Loci C and D are monomorphic. (exp. = expected value; obs. = observed value.)

2. The proportion of polymorphic loci is P = (3/5) = 0.6 or 60%. That is, the number of polymorphic loci is divided by the total number of loci analysed.

3. To calculate average heterozygosity (H_o), we:
   a. Count how many loci, out of the total, are heterozygous. For instance, $Individual_1$ has one heterozygous locus (A), $Individual_2$ also has one (E), $Individual_{27}$ has 2 heterozygous loci (A and E), ... In all, 16 individuals were monomorphic (i.e. they had only one band in each of the five loci), 13 individuals had 1 heterozygous locus and 1 individual had 2 heterozygous loci.
   b. Calculate the average observed heterozygosity as:

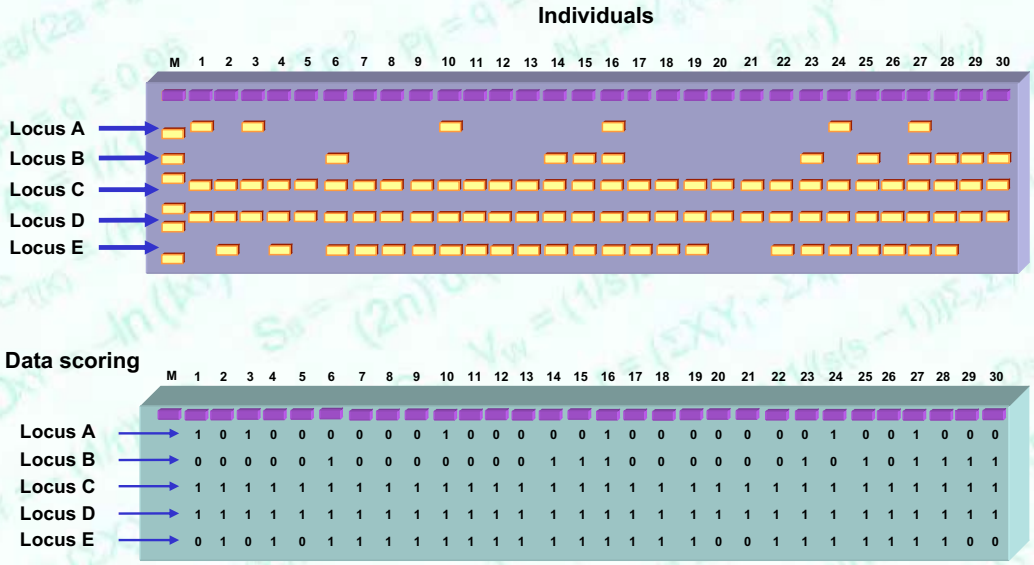
$$H_o = [16(0/5) + 13(1/5) + 1(2/5)]/(30) = 0.1$$

4. The intralocus gene diversity (h_j) is calculated for each locus according to the formula in the top row of the table, giving us locus A = 0.23, locus B = 0.41 and locus E = 0.46.

5. The average expected gene diversity (H_i) is calculated from the formula in slide number 12:

$$H_i = (0.23 + 0.41 + 0.46)/5 = 0.22$$

Calculating diversity with a dominant molecular marker



(continued on next slide)

The top half of this slide shows a drawing of a gel with a size marker on the left (M) and 30 individuals analysed with a dominant marker. Five loci are identified (A, B, C, D and E). Of the five loci detected, three are segregating (A, B and E), while the other two, C and D, are monomorphic.

The bottom half of the slide shows the results of scoring bands per individual and per locus. Because we are dealing with a dominant marker, bands are scored as 1 when present or 0 when absent. Scoring the bands for loci C and D can be either omitted or done as in the slide with '1' for every individual.

The next slide shows the calculations.

Calculating diversity with a dominant molecular marker (continued)

Locus	Data analysis					Allele freq.		$h_j = (1 - p^2 - q^2)$	H_i
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.)	$P_1 = 0.20$		$P_2 = 0.80$	1	0.11	0.89	0.19	
B	Genotypes	BB	Bb	bb	Total	p	q		
	Gen. freq. (exp.)	p^2	2pq	q^2	1				
	Individuals (no.)	10		20	30				
	Gen. freq. (obs.)	$P_1 = 0.33$		$P_2 = 0.67$	1	0.18	0.82	0.30	
E	Genotypes	EE	Ee	ee	Total	p	q		
	Gen. freq. (exp.)	p^2	2pq	q^2	1				
	Individuals (no.)	23		7	30				
	Gen. freq. (obs.)	$P_1 = 0.77$		$P_2 = 0.23$	1	0.52	0.48	0.50	0.198

1. First, we look at the polymorphism shown by all loci. Loci A, B and E fulfill the requirement of having allele frequencies below 0.99 and as such can be said to be polymorphic. Loci C and D are monomorphic. (exp. = expected value; obs. = observed value.)

2. The proportion of polymorphic loci (P) is P = (3/5) = 0.6 or 60%. The average heterozygosity (H_e) cannot be estimated because dominant markers do not allow discrimination between heterozygous and homozygous individuals.

3. Despite the above (2), the intralocus gene diversity (h_j) may be calculated for each locus using the formula that appears in the top row of the table, column 4, as follows: locus A = 0.19; locus B = 0.30; and locus E = 0.50.

4. The average gene diversity (H_i) is calculated from the formula in slide number 12:

$$H_i = (0.19 + 0.30 + 0.50)/5 = 0.198$$

Quantifying genetic diversity: measuring interpopulation genetic diversity

- Interpopulation differentiation for one locus (g_{ST})
- Interpopulation differentiation for several loci (G_{ST})
- Population's contribution to total genetic diversity
- F statistics (Wright)
- Analysis of molecular variance (AMOVA)

'Differentiation' refers to polymorphic differences between populations at different levels of structure (populations and individuals).

Interpopulation differentiation for one locus (g_{ST})

$$g_{ST} = 1 - (h_S/h_T)$$

h_S = population diversity

h_T = total diversity

Where,

$h_S = (\tilde{n}/(\tilde{n} - 1)[1 - (1/s)\sum\sum x_{ij}^2 - (h_o/2\tilde{n})]$

$h_T = 1 - \sum[(1/s)\sum x_{ij}]^2 + (h_S/\tilde{n}s) - (h_o/2\tilde{n}s)$

$\tilde{n}$ = harmonic average of population sizes

s = number of populations

h_o = average observed heterozygosity

x_{ij} = estimated frequency of the i^{th} allele in the j^{th} population

- The formula in the slide provides a measure of differentiation in terms of alleles per locus in two or more populations
- It ranges from 0 to 1. A negative value may be obtained if an error was made for sampling or an inappropriate marker system was used.
- Because of the complexity of its components, its calculation requires specialized computer software.
- It can be used with codominant markers and restrictedly with dominant markers. This is because it is a measure of heterozygosity. To have a fair estimate of the real value, several generations are needed.

Calculating the g_{ST}

Genotypes	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	p	q	$p^2 + q^2$
Pop. 1	20	30	50	0.35	0.65	0.545
Pop. 2	10	20	70	0.20	0.80	0.680
Pop. 3	60	10	30	0.65	0.35	0.545
$h_o = 1/3(0.3 + 0.2 + 0.1) = 0.20$				$s = 3$		$\sum(p^2 + q^2) = 1.77$
$1/\tilde{n} = 1/n_1 + 1/n_2 + 1/n_3 = 1/100 + 1/100 + 1/100 = 0.03$						$\tilde{n} = 33.33$
$h_s = (33.33/33.33 - 1)[1 - 1/3(1.77) - (0.20/2(33.33))] = 0.4196$						
$\sum[1/3\sum x_{ij}]^2 = (1/3(0.35))^2 + (1/3(0.65))^2 + (1/3(0.20))^2 + \ldots + (1/3(0.35))^2 = 0.1967$						
$h_T = 1 - 0.1967 + [0.4196/(33.33 \times 3)] - [0.20/(2 \times 33.33 \times 3)] = 0.8065$						
$g_{ST} = 1 - (h_s/h_T) = 1 - (0.4196/0.8065) = 0.4797$						

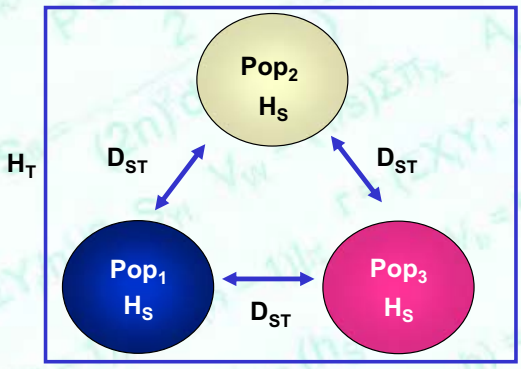
In this example, we have the number of individuals for each genotype for one locus (A) in three different populations. Using this number, we want to know the degree of differentiation in the three populations. In the table, the calculations are followed for all the necessary elements in the formula shown on the previous slide.

The result ($g_{ST} = 0.4797$) shows significant differentiation between populations in allele frequencies. We can therefore say that a high percentage of genetic diversity is distributed among populations.

Interpopulation differentiation for several loci (G_{ST})

G_{ST} is the coefficient of gene differentiation

$$G_{ST} = D_{ST}/H_T$$



Where,

H_T = total genic diversity = $H_S + D_{ST}$
H_S = intrapopulation genic diversity
D_{ST} = interpopulation diversity
$(H_T/H_T) = (H_S/H_T) + (D_{ST}/H_T) = 1$

- G_{ST} measures the proportion of gene diversity that is distributed among populations.
- A larger number of loci must be sampled.
- Equations are complex and should be calculated with specific computer software.

For example, assuming that:

$H_T = 0.263$
$H_S = 0.202$
$D_{ST} = 0.263 - 0.202 = 0.061$

Then, $G_{ST} = (D_{ST}/H_T) * 100 = (0.061/0.263) * 100 = 23.19\%$. This means that, in this species, a 23% differentiation among populations exist.

Population's contribution to total gene diversity

The contribution is calculated by removing a population so that its contribution to the total gene diversity may be evaluated

$$C_{T(K)} = (H_T - H_{T/K})/H_T$$

$$C_{S(K)} = (H_S - H_{S/K})/H_T$$

$$C_{ST(K)} = (D_{ST} - D_{ST/K})/H_T$$

Where,

$C_{T(K)}$ = contribution of K to total diversity
$C_{S(K)}$ = contribution of K to intrapopulation diversity
$C_{ST(K)}$ = contribution of K to interpopulation diversity
H_T = total gene diversity
H_S = intrapopulation genic diversity
D_{ST} = interpopulation diversity
$H_{T/K}$ = total gene diversity after removing population K
$H_{S/K}$ = intrapopulation gene diversity after removing population K
$D_{ST/K}$ = interpopulation gene diversity after removing population K

- The measure allows quantifying the variation of total gene diversity when a population is introduced to or removed from a site (e.g. when introducing a new variety into a farmer's field in an *in situ* conservation programme).
- It also serves to measure the impact of losing a population from a given site in terms of gene diversity.
- It can be used only with codominant markers.

F statistics (Wright)

The equation for the genetic structure of populations is:

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

$$F_{IT} = 1 - (H_I/H_T)$$

$$F_{IS} = 1 - (H_I/H_S)$$

$$F_{ST} = 1 - (H_S/H_T)$$

Where,

H_T = total gene diversity or expected heterozygosity in the total population as estimated from the pooled allele frequencies

H_I = intrapopulation gene diversity or average observed heterozygosity in a group of populations

H_S = average expected heterozygosity estimated from each subpopulation

The F statistics allow analysis of structures of subdivided populations. It may also be used to measure the genetic distance among subpopulations, a concept that is based on the idea that those subpopulations that are not intermating will have different allele frequencies to those of the total population.

Genetic distance also provides a way of measuring the probability of encounter between equal alleles (endogamy). The statistical indexes involved measure:

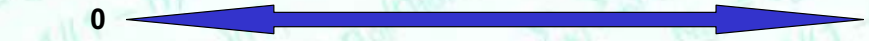
F_{IS} = the deficiency or excess of average heterozygotes in each population

F_{ST} = the degree of gene differentiation among populations in terms of allele frequencies

F_{IT} = the deficiency or excess of average heterozygotes in a group of populations

Interpreting F_{ST} values

The range of the F_{ST} is:



(no genetic divergence)

(fixation for alternate alleles in different subpopulations)

When F_{ST} is:	**then the genetic differentiation is:**
0 to 0.05	small
0.05 to 0.15	moderate
0.15 to 0.25	large
>0.25	very large

Calculating F statistics

Pop.	Genotype frequency						
	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	p_i	q_i	$2p_iq_i$	F
1	0.40	0.30	0.30	0.55	0.45	0.4950	0.3939
2	0.60	0.20	0.20	0.70	0.30	0.4200	0.5238
H_T	2(0.625)(0.375) = 0.4688			**p_o**	(0.55 + 0.70)/2 = 0.625		
H_I	(0.3 + 0.2)/2 = 0.25			**q_o**	(0.45 + 0.30)/2 = 0.375		
H_S	(0.495 + 0.420)/2 = 0.4575						

F_{IT} = 1 – (0.25/0.4688) = 0.4667

F_{IS} = 1 – (0.25/0.4575) = 0.4536

F_{ST} = 1 – (0.4575/0.4688) = 0.0241

(continued on next slide)

This slide shows an example of two populations and the analysis of one locus (A). The allele frequencies are calculated (p and q), as are their averages. The variables H_T, H_I and H_S are also estimated and used to calculate the F statistics. The analysis shows low differentiation in allele frequencies among the two populations (F_{ST}). We could conclude that almost all the heterozygote deficit was due to nonrandom mating within the populations (F_{IS} = 0.4536).

F = fixation index (first column on right of table) is the probability that two alleles carried by one individual will be the same. It should be calculated only with codominant markers. If done with dominant markers, a biased estimator may result.

Calculating F statistics (continued)

Pop.	Genotype frequency						
	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	p_i	q_i	$2p_iq_i$	F
1	0.25	0.50	0.25	0.50	0.50	0.500	0.0000
2	0.80	0.10	0.10	0.85	0.15	0.255	0.6078
H_T	2(0.675)(0.325) = 0.4388			p_o	(0.50 + 0.85)/2 = 0.675		
H_I	(0.5 + 0.1)/2 = 0.30			q_o	(0.50 + 0.15)/2 = 0.325		
H_S	(0.500 + 0.255)/2 = 0.3775						

F_{IT} = 1 – (0.30/0.4388) = 0.3163

F_{IS} = 1 – (0.30/0.3775) = 0.2053

F_{ST} = 1 – (0.3775/0.4388) = 0.1397

This is another example for which the procedures used in the previous slide were followed. Differentiation in allele frequencies between the two populations seems greater (F_{ST} = 0.1397), with only a moderate effect of nonrandom mating within the populations (F_{IS} = 0.2053).

Analysis of molecular variance (AMOVA)

- AMOVA is a method for studying molecular variation within a species
- It is based on a hierarchical or nested model
- It differs from an analysis of variance (ANOVA) in that:
  - It may contain different evolutionary assumptions without modifying the basic structure of the analysis
  - The driving hypothesis uses permutational methods that do not require the assumption of a normal distribution

The different hierarchical levels of gene diversity studied through AMOVA may include:

1. Continents, which may contain lesser hierarchical levels
2. Geographical regions within a continent
3. Areas within a region in a continent
4. Populations within an area of a region in a continent
5. Individuals within a population in an area of a region in a continent

The mathematical description of the model for situations 3 and 4 can be found in Appendices 2 and 3, respectively.

The next two slides illustrate how to analyse situation 4.

An example of AMOVA

Ind.	Pop. 1		Pop. 2		Pop. 3	
	A_1	A_2	A_1	A_2	A_1	A_2
1	0	0	0	1	1	1
2	1	1	0	1	1	1
3	0	0	1	1	0	1
4	1	0	1	0	1	1
5	0	0	0	1	0	1
6	0	0	0	1	0	0
7	1	1	1	1	1	1
8	0	0	1	1	0	0
9	1	0	1	1	1	0
10	1	1	1	0	0	1
11	1	0	0	1	1	1
12	0	0	1	1	1	0
13	1	1	1	1	0	1
14	1	1	1	0	1	0
15	1	1	0	1	1	0

$X..._k$	15	21	18	54
$X..._k^2$	225	441	324	990
$\sum\sum X_{i...k}^2$	27	33	28	88
$\sum\sum\sum X_{ijk}^2$	15	21	18	54
$X...^2$				2916

SS_a	0.6	MS_a	0.3
SS_b	11	MS_b	0.26190476
SS_w	10	MS_w	0.22222222

$A_1 = 1$	Present
$A_1 = 0$	Absent

(continued on next slide)

In this table, we show data obtained with 15 individuals from each of three populations in an analysis with a codominant marker. By means of an analysis of variance, these data will allow us to calculate the F statistics.

The first step is to convert the bands detected in the gels to binary variables with a value of either 0 or 1. Then, the sums of presences (1) are calculated so we may proceed with the sum of squares. Calculations are first done for one population and continued for the others until we have ($X..._k$). We have i = 15 individuals (effect b), j = 2 alleles (effect w), k = 3 populations (effect a).

Where,

$X..._k$ is the result of summing up all the band presences (1) in the individuals per population

$X..._k^2$ is the result of squaring the number obtained above

$\sum\sum X_{i...k}^2$ is the result of adding up the squares of the sum of allele presences in each individual (e.g. Indiv.$_1$ in Pop.$_1$ will be $(0 + 0)^2$ + Indiv.$_2$ in Pop.$_1$ $(1 + 1)^2$ + Indiv. ...)

$\sum\sum\sum X_{ijk}^2$ is the sum of each value squared

SS is the sum of squares for effects a, b and w

An example of calculating SS:

$SS_a = \sum X..._k^2/ij - X...^2/ijk = [990/(15 \times 2)] - [2916/(15 \times 2 \times 3)] = 0.6$

MS are the mean squares for effects a, b and w

An example of calculating MS: $SS_a/df_a = 0.6/2 = 0.3$, where dfA refers to the degrees of freedom for effect a (populations).

An example of AMOVA (continued)

SV	df	SS	MS	EMS
Populations	2	0.6	0.3	$\sigma_w^2 + 2\sigma_b^2 + 2*15\sigma_a^2$
Indiv./pop.	42	11	0.26190476	$\sigma_w^2 + 2\sigma_b^2$
Within indiv.	45	10	0.22222222	σ_w^2

Estimates of variances and F statistics	
σ_a^2	0.0012698
σ_b^2	0.0198413
σ_w^2	0.2222222
σ^2	0.24333
F_{IT}	0.086758
F_{IS}	0.0819672
F_{ST}	0.0052185
$(1 - F_{IT})$	0.91324
$(1 - F_{IS})(1 - F_{ST})$	0.91324

Where,

SV = sources of variation
df = degrees of freedom
SS = sum of squares (see previous slide)
MS = mean squares (see previous slide)
σ^2 = total estimated variance
EMS = expected mean squares
$\sigma_w^2 = 0.2222222$
$\sigma_b^2 = (MS_b - MS_w)/2 = (0.26190476 - 0.22222222)/2 = 0.0198413$
$\sigma_a^2 = (MS_a - MS_b)/2 * 15 = (0.3 - 0.26190476)/2 * 15 = 0.0012698$
$\sigma^2 = \sigma_w^2 + \sigma_b^2 + \sigma_a^2 = 0.24333$ (total estimated variance)

Calculating the F statistics has already been explained in slide 22. For this particular example, they would be as follows:

$F_{IT} = (\sigma_a^2 + \sigma_b^2)/\sigma^2 = (0.0012698 + 0.0198413)/0.24333 = 0.086758$
$F_{ST} = \sigma_a^2/\sigma^2 = 0.0012698/0.24333 = 0.0052185$
$F_{IS} = \sigma_b^2/(\sigma_b^2 + \sigma_w^2) = 0.0198413/(0.0198413 + 0.222222) = 0.0819672$

The allele frequency differentiation among the three populations is very low ($F_{ST} = 0.0052185$) and is probably a result of many random matings. More loci need to be analysed to make a conclusion.

Quantifying genetic relationships: diversity and differentiation at the nucleotide level

- Using sequence data
  - Intrapopulation nucleotide diversity
  - Interpopulation nucleotide diversity
- Using restriction data
  - Variations in banding patterns
  - Intrapopulation nucleotide diversity
  - Interpopulation nucleotide diversity

For these calculations, the assumption is made that each nucleotide is a locus.

Using sequence data: intrapopulation nucleotide diversity

It measures the nucleotide diversity among several sequences in a given region of the genome within a population (π_X)

$$\pi_X = n/(n-1)\Sigma X_i X_j \pi_{ij}$$

Where,

n = number of sequences under analysis in the individuals of the population

X_i = estimated frequency of the i^{th} sequence in the population

X_j = estimated frequency of the j^{th} sequence in the population

π_{ij} = proportion of different nucleotides between sequences i and j

- The measure informs about the degree of nucleotide diversity among several sequences in a given region of the genome. It is equivalent to the measure of allelic diversity within a locus.
- It ranges from 0 to 1 ($0 \leq \pi_X \leq 1$).
- The factors limiting the use of this analytical tool are:

  Partial genomic sequences must be available

  The equation can only be applied to haploid data

This parameter informs about nucleotide sequences, and the model assumes haplotypes (haploid genotypes). Even if the study is based on diploid individuals, sequencing of each copy of the genome is needed.

Calculating intrapopulation nucleotide diversity

n	Sequence		Freq. X_i
5	Seq_1	TCC T CGAT T ATTC C CAGGGTGC C GATG A AT	5/10 = 0.5
2	Seq_2	TCC A CGAT T ATTC G CAGGGTGC C GATG A AT	2/10 = 0.2
1	Seq_3	TCC A CGAT C ATTC C CAGGGTGC A GATG G AT	1/10 = 0.1
2	Seq_4	TCC G CGAT T ATTC T CAGGGTGC G GATG A AT	2/10 = 0.2
10			

Π1,2 = 2/30, Π1,3 = 4/30, Π1,4 = 3/30, Π2,3 = 4/30, Π2,4 = 3/30, Π3,4 = 5/30

$\pi_X = 10/(10-1)\Sigma X_i X_j \pi_{ij}$

$= (10/9)[0.5 * 0.2 * (2/30) + 0.5 * 0.1 * (4/30) + ... + 1 * 0.2 * (5/30)]$

= 0.037

This example has 10 individuals in a population X. For each individual, we analyse one sequence of 30 nucleotides, and find that individual sequences differ at 5 nucleotide positions (blue). In total, 4 alternative sequences for those 30 nucleotides are present in the population. In the first column, we note the number of individuals (n) that have the particular sequence alternatives.

Then, we calculate the number of nucleotide differences in each sequence pair within the population. For example, Π1,2 = 2/30 means that between sequence 1 and 2 there are two nucleotide differences (T vs. A in position 4, and C vs. G in position 14).

Next, we calculate π_X for the entire population. The number obtained is 0.037, or 3.7% nucleotide diversity, based on the sequence analysed in the sample of 10 individuals.

Using sequence data: interpopulation nucleotide diversity

- V_{XY} measures population divergence based on the degree of sequence variation (1 sequence, 2 populations)

$$V_{XY} = d_{XY} - (\pi_X + \pi_Y)/2$$

- V_W measures average diversity in a population based on several sequences

$$V_W = (1/s)\Sigma\pi_X$$

- V_b measures the total differentiation in several populations

$$V_b = [1/(s(s-1))]\Sigma_X\Sigma_Y V_{XY}$$

- N_{ST} is the relative differentiation

$$N_{ST} = V_b/(V_b + V_W)$$

Where,

V_{XY} = divergence among populations X and Y

π_X = nucleotide diversity in population X

d_{XY} = the probability that two random nucleotides in populations X and Y be different

s = number of populations

- The measure informs about the level of differentiation among nucleotide sequences in populations.
- It requires sequence data in a sample of individuals for each population.
- It needs specific computer software that includes sequence alignment features.

  Some of these are CLUSTAL W, MALIGN and PAUP*.

Calculating interpopulation nucleotide diversity

Nucleotide divergence between X and Y

$V_{XY} = d_{XY} - (\pi_X \pi_Y)/2 = 0.14 - (0.037 + 0.09)/2 = 0.0765$

Total differentiation

$V_b = [1/(s(s-1))]\Sigma_X \Sigma_Y V_{XY} = [1/(2(2-1))]0.0765 = 0.03825$

Average diversity in each population

$V_W = (1/s)\Sigma\pi_X = ½(0.037 + 0.09) = 0.0635$

Relative differentiation

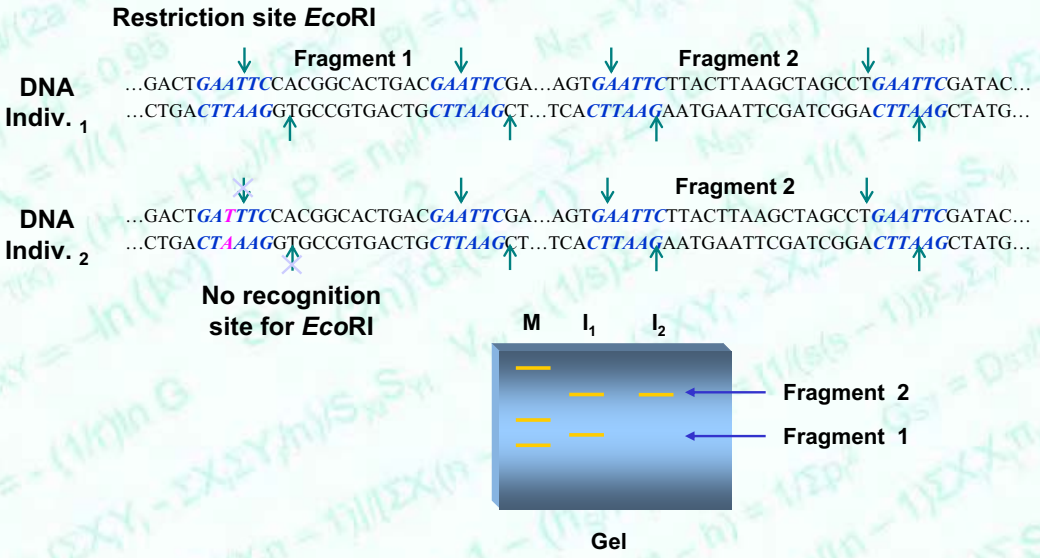
$N_{ST} = V_b/(V_b + V_W) = 0.03825/(0.03825 + 0.0635) = 0.3759$

Let's say that we have another population Y, in which the nucleotide diversity for the same sequence analysed in slide 31 is $\pi_Y = 0.09$.

We also know that the probability that two nucleotides as taken at random are different in X and Y is 0.14 (d_{XY}).

In this slide, we find the divergence between populations X and Y (V_{XY}), the total differentiation (V_b), the average diversity in each population (V_w) and the relative differentiation (N_{ST}).

Using restriction data: variations in banding patterns



The lack of fragment 1 in Individual$_2$ indicates that it carries a different DNA sequence at least in that restriction site. A small difference of just two nucleotides, in the drawing above, is sufficient to make the recognition site for the enzyme to disappear.

Using restriction data: intrapopulation nucleotide diversity

This measure (π) is based on the number of restriction fragments present in two samples

$$\pi = - (1/r) \ln G$$

(if π < 5%)

Where,

r = number of recognition nucleotides of a restriction enzyme

ln G = natural logarithm of the probability that there was no substitution in the restriction site. Its calculation is:

$$G = F(3 - 2G^{o})^{1/4}$$

$$F = [\Sigma X_i(X_i n - 1)]/[\Sigma X_i(n - 1)]$$

F = proportion of shared fragments

$G^{o} = F^{1/4}$

n = number of haploid genotypes in the population

X_i = estimated frequency of the i^{th} fragment in the population

- The measure estimates the diversity in restriction sites in a sample, because it relies on the nucleotide sequence of the recognition sites of a given restriction enzyme.
- It informs about the nucleotide substitution in restriction sites. It varies from 0 to 1 ($0 \leq \pi_X \leq 1$).
- The equations above can be used with haploid samples, mDNA, cpDNA or haplotypes.

Reference

Karp, A., P.G. Isaac and D.S. Ingram. 1998. Molecular Tools for Screening Biodiversity: Plants and Animals. Chapman & Hall, London.

Using restriction data: interpopulation nucleotide diversity

- This measure (V_{XY}) indicates the divergence or differentiation among populations based on restriction data

$$V_{XY} = d_{XY} - (\pi_X + \pi_Y)/2$$

- This measure is also used with RAPD marker data

Where,

V_{XY} = divergence or differentiation among populations X and Y
π_X = restriction diversity in population X
d_{XY} = fragment diversity among two populations = $-(2/r)\ln(G_{XY})$
$G_{XY} = F_{XY}(3 - 2G^{o}_{XY})^{1/4}$
$G^{o} = F_{XY}^{1/4}$
F_{XY} = proportion of shared alleles among populations X and Y
$= (2\Sigma X_{iX}X_{iY})/(\Sigma(X_{iX} + X_{iY}))$
X_{iX} = estimated frequency of the i fragment in population X

- It estimates diversity in the restriction sites of a sample of two or more populations. It informs about the nucleotide substitution in the restriction sites.
- Computer software such as BIOSYS and GENEPOP are useful. Data obtained are considered as belonging to haploid organisms.

If used with RAPDs, the value of 'r' is replaced by the primer length (r = 10).
In addition, some assumptions are taken:

The appropriate primers are used
Polymorphism due to insertion or deletion is rare
Similar size fragments in different populations belong to the same locus
Fragments must be identified without error

Software such as RAPDISTANCE and RAPDIS is typically used.

Calculating interpopulation nucleotide diversity

Population X	Seq.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Freq. Xi
	A_1	—	—				—	—					—	—								6/20 = 0.30
	A_2			—	—				—			—						—				5/20 = 0.25
	A_3					—				—	—				—	—	—		—	—	—	9/20 = 0.45
	$F = \frac{[0.30(0.30 * 3 - 1) + 0.25(0.25 * 3 - 1) + 0.45(0.45 * 3 - 1)]}{0.30(3-1) + 0.25(3-1) + 0.45(3-1)} = 0.0325$																					
	$G° = (0.0325)1/4 = 0.424591$													$G = 0.0325[3 - 2(0.424591)]1/4 = 0.039358$								
	$\pi X = -(1/6) \ln (0.039358) = 0.539176$																					

Population Y	Seq.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Freq. Xi
	A_1		—		—	—				—							—					5/20 = 0.25
	A_2	—		—			—		—		—	—	—	—	—			—	—	—	—	13/20 = 0.65
	A_3							—								—						2/20 = 0.10
	$F = \frac{[0.25(0.25 * 3 - 1) + 0.65(0.65 * 3 - 1) + 0.10(0.10 * 3 - 1)]}{0.25(3-1) + 0.65(3-1) + 0.10(3-1)} = 0.2425$																					
	$G° = (0.2425)1/4 = 0.701743$													$G = 0.2425[3 - 2(0.701743)]1/4 = 0.272587$								
	$\pi Y = -(1/6) \ln (0.272587) = 0.216633$																					

In each population, we detected three DNA fragments as a result of a restriction digest: A1, A2 and A3.

Nucleotide diversity in the regions analysed is larger in population X ($\pi_X = 0.5392$) than in population Y ($\pi_Y = 0.2166$); thus, X has more gene diversity than Y.

Between populations X and Y, the nucleotide differentiation based on restriction sites is 0.230766.

$$F = \frac{2[0.30*0.25+0.25*0.65+0.45*0.10]}{(0.30+0.25)+(0.25+0.65)+(0.45+0.10)} = 0.14125$$

$$G°XY = (0.14125)^{1/4} = 0.613052$$

$$GXY = 0.14125[3 - 2(0.613052)]^{1/4} = 0.163012$$

$$dXY = -(2/6)\ln(0.163012) = 0.604643$$

$$VXY = 0.604643 - \tfrac{1}{2}(0.539176 + 0.216633) = 0.226739$$

$$VW = \tfrac{1}{2}(0.539176 + 0.216633) = 0.377905$$

$$Vb = \tfrac{1}{2}(0.226739) = 0.11337$$

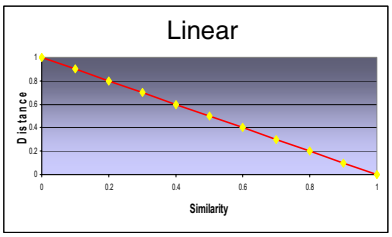
$$NST = \frac{0.11337}{0.11337 + 0.377905} = 0.230766$$

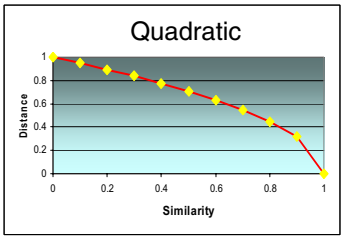
Quantifying genetic relationships: genetic distance

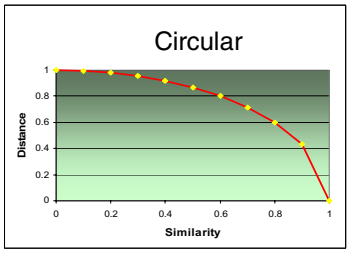
- The genetic distance between two samples is described as the proportion of genetic elements (alleles, genes, gametes, genotypes) that the two samples do not share
- D = 1 when, and only when, the two samples have no genetic elements in common

Depending on the similarities of individuals, three representation types of distance (D) are possible:

- D = 1 – S, known as linear distance, because it assumes that the relationship with similarity is linear.
- D = $\sqrt{(1 - S)}$, known as quadratic distance because it assumes that the relationship with similarity follows a quadratic function, so that, to make it linear, the square root must be calculated.
- D = $\sqrt{(1 - S^2)}$, known as circular distance.



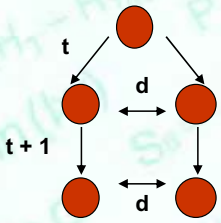




Distance models

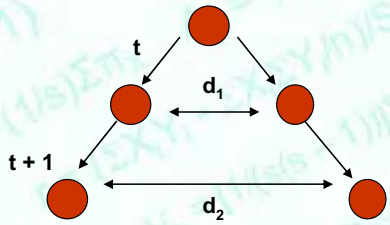
Calculation of distance, or dissimilarity, follows one of two possible models:

Equilibrium model



Distance remains constant over time (equilibrium exists between migration and genetic drift)

Disequilibrium model



Distance changes with time through migration and genetic drift

For our purposes, we will use the disequilibrium model. Two alternatives exist:

- Geometric distance
  - Does not take into account evolutionary processes
  - Based only on allele frequencies
  - Complex relationship exists between distance and divergence time
- Genetic distance
  - Does not take into account evolutionary processes
  - Distance increases from the time of separation from an ancestral population
  - A genetic model of evolution is needed

When should we use geometric or genetic distance?

- Geometric distance is used for diversity studies in which comparisons are made according to morphological or marker data gathered from the operative taxonomic units (OTUs). OTUs may be individuals, accessions or populations. It can be used with dominant markers (RAPDs, AFLPs) or codominant markers. Because evolutionary aspects are not considered, the dendrograms obtained cannot be interpreted as phylogenetic trees giving information about evolution or divergence among groups.
- In contrast, the genetic distance of any given OTU can be incorporated into phylogeny studies. The model considers allelic frequencies in OTUs and its mathematical foundation is different. It can be used with both codominant and dominant markers, although, with the latter, information is lost because only two alleles can be scored. Genetic distance with dominant markers, however, requires the examination of two generations of the same population to measure the segregation of loci (Lynch and Milligan, 1994).

Reference

Lynch, M. and B.G. Milligan. 1994. Analysis of population genetic structure with RAPD markers. Mol. Ecol. 3:91-99.

Disequilibrium models: geometric distance

- This measures the direct relationship between the similarity index (s) and distance (D = 1 – s)
- Different situations are possible, for example:
  - Binary variables
  - Quantitative variables
  - Mixed types of variables
  - P number of variables

(continued on next slide)

When analysing molecular data, we deal with binary variables (1,0). These will be discussed in the following slides.

In Appendix 4, you will find additional information for those cases where you also must deal with quantitative variables, mixed types of variables and a varied number of variables. In Appendix 5, an example of calculating geometric distances with quantitative variables has also been added.

Geometric distance (continued)

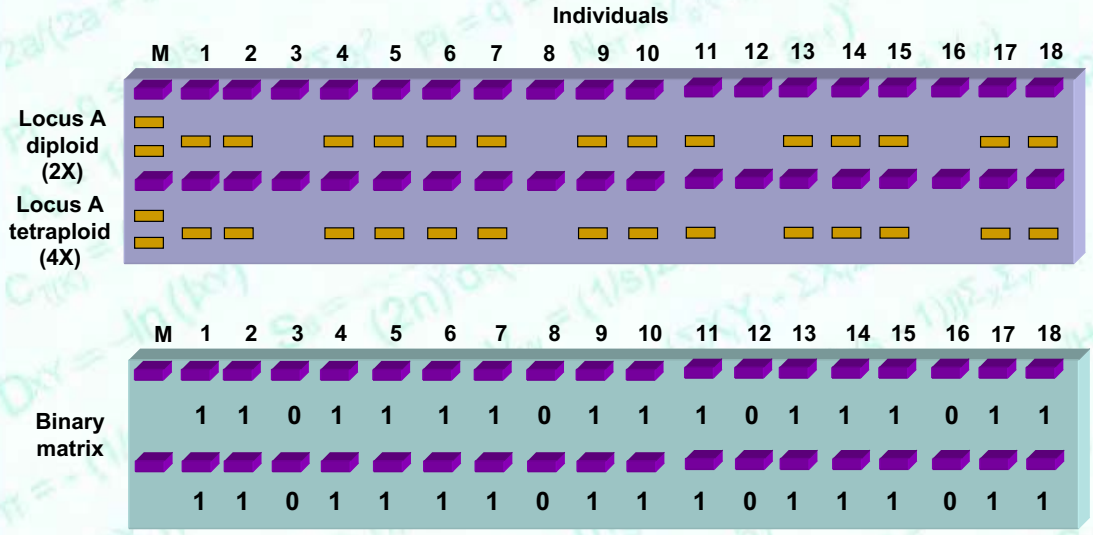
With binary variables:

- Multivariate analysis is used and similarity or differentiation matrices are formed between the possible pairs of individuals or operative taxonomic units (OTUs)
- Two similar individuals simultaneously have the minimum value of distance and the maximum value of similarity
- Distance and similarity are inversely related
- Similarity is estimated by the number of coincidence

When using molecular marker data and transforming them to binary data, the following should be considered:

- A species's ploidy number may mask the presence of allelic series in a locus. If this happens, genetic diversity will be underestimated when using dominant markers (presence/absence).
- If a marker is codominant, large samples are needed to permit detection of all possible genotypes, particularly if there are several alleles per locus.
- Segregation distortions are common in polyploid species.
- Most specialized computer software are designed to analyse diploid species. Therefore if used with polyploid species, biases may occur on estimating the various genetic diversity indices.
- The reproductive system of certain species has not been studied, so their inheritance type is not sufficiently known.
- The largest coverage (coding and non-coding regions) possible of the genome of the species under study should be sampled and analysed so that estimates of genetic diversity are reliable.

Calculating allele frequencies for diploids and tetraploids: dominant marker

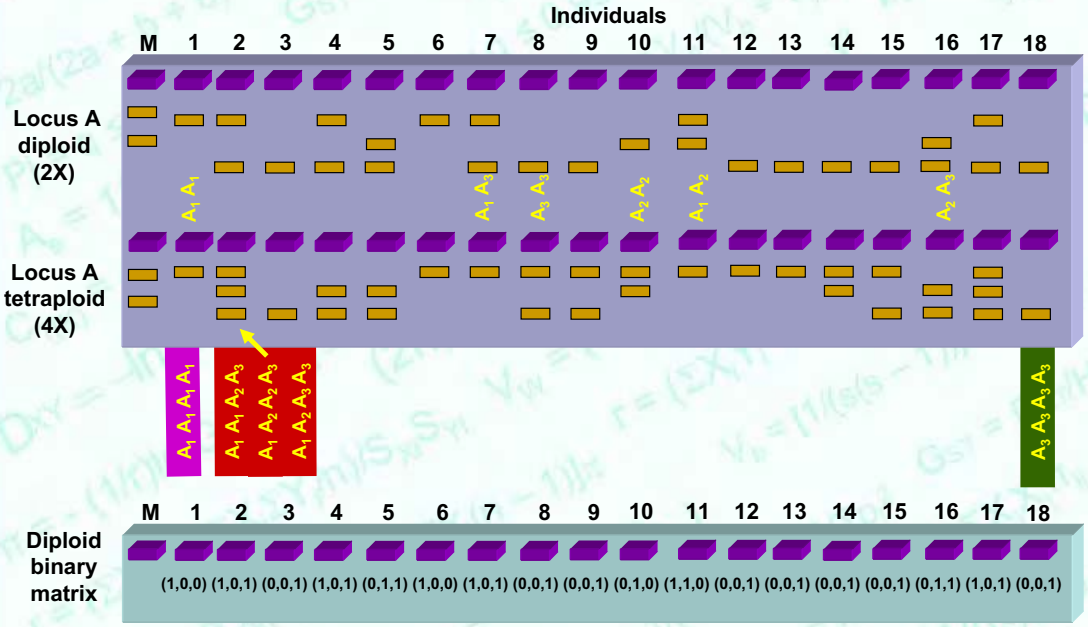


This example of 18 individuals from each of a diploid and a tetraploid species was analysed with a dominant marker. We are assuming that the banding patterns obtained are alike. Bands are converted to a binary table in both cases. The calculations of the frequencies are given in the table below. We can see that, in the tetraploid, genotype 1, for example, can be either AAAA, AAAa, AAaa or Aaaa; however the band will only be scored as present (1) the same as it will in the diploid (AA or Aa).

Locus	Genotypes				Allele freq.	
A (2X)	Diploid	AA, Aa	aa	Total	p	q
	Gen. freq. (e.)	$p^2 + 2pq$	q^2	1		
	Indiv. number	14	4	18		
	Gen. freq. (o.)	$P_1 = 0.78$	$P_2 = 0.22$	1	0.53	0.47
A (4X)	Tetraploid	AAAA, AAAa, AAaa, Aaaa	aaaa	Total	p	q
	Gen. freq. (e.)	$p^4 + 4p^3q + 6p^2q^2 + 4pq^3$	q^4	1		
	Indiv. number	14	4	18		
	Gen. freq. (o.)	$P_1 = 0.78$	$P_2 = 0.22$	1	0.31	0.69

Allele frequencies should be different in both cases; however, the information loss in the tetraploid individual is significant. Why? This is because, to estimate the frequency of the recessive allele a, heterozygotes AAAa, AAaa, Aaaa are not taken into account. This effect is larger when the ploidy number of the species under study is unknown.
(e. = expected value; o. = observed value.)

Calculating allele frequencies for diploids and tetraploids: codominant marker



In this example, we have 18 individuals from each of a diploid and tetraploid species and analysed with a codominant marker. One locus is detected (A) with three alleles in both situations (A_1, A_2 and A_3).

Calculating the allele frequencies in the diploid individual is not difficult (binary matrix, bottom of slide). For the tetraploid individual, however, conversion to binary data is hampered by the fact that individuals with alleles A_1 A_1 A_2 A_3 cannot be distinguished from those with other combinations such as A_1 A_2 A_2 A_3 or A_1 A_2 A_3 A_3. This situation can only be solved by inference based on estimating the DNA fragment copy number in the gel.

Genotype	A_1 A_1	A_1 A_2	A_1 A_3	A_2 A_2	A_2 A_3	A_3 A_3	Total	p	q	r
Gen. freq. (e.)	p^2	2pq	2pr	q^2	2qr	r^2	1			
Indivs. (no.)	2	1	4	1	2	8	18	0.25	0.15	0.60
Gen. freq. (o.)	P_{11} = 0.11	P_{12} = 0.06	P_{13} = 0.22	P_{22} = 0.06	P_{23} = 0.11	P_{33} = 0.44	1			

(e. = expected value; o. = observed value.)

Similarity coefficients for binary variables: examples

	Author	Expression	Example of the coefficient value if a = 3, b = 1, c = 3, d = 2
S1	Russel and Rao (1940)	a/n	0.333
S2	Simpson	a/min[(a + b),(a + c)]	0.750
S3	Braun-Blanquet	a/max[(a + b),(a + c)]	0.500
S4	**Dice (1945); Nei and Li (1979)**	**a/[a + (b + c)/2]**	**0.600**
S5	Ochiai (1957)	$a/[(a + b)(a + c)]^{1/2}$	0.612
S6	Kulczynski 2	(a/2)([1/(a+b)] + [1/(a+c)])	0.625
S7	**Jaccard (1900, 1901, 1908)**	**a/(a + b + c)**	**0.429**
S8	Sokal and Sneath 5 (1963)	a/[a +2(b + c)]	0.273
S9	Kulczynski 1 (1928)	a/(b + c)	0.750
S10	**Sokal and Michener (1958)**	**(a + d)/n**	**0.556**
S11	Rogers and Tanimoto (1960)	(a + d)/[a + d + 2(b + c)]	0.385
S12	Sokal and Sneath 1 (1963)	(a + d)/[a + d + (b + c)/2]	0.714
S13	Sokal and Sneath 3 (1963)	(a + d)/(b + c)	1.250

		Indiv.$_j$		
		1	0	
Indiv.$_i$	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	n

Where,

$$n = a + b + c + d$$

In the table above we see that:

Indices S1 to S9 give value only to the presence of information
Indices S10 to S13 give value to both presence and absence

Next, we will discuss three indices (in red on top table): Simple Matching (S10), Jaccard (S7) and Nei-Li (S4).

Indexes of geometric distance

Simple Matching Coefficient, or simple concordance coefficient:

$$(a + d)/(a + b + c + d)$$

Jaccard Coefficient:

$$a/(a + b + c)$$

Nei-Li Coefficient, or Dice:

$$2a/(2a + b + c)$$

These three indices differ in their approach for estimating the number of coincidences and differences.

The **Simple Matching Coefficient** considers that absence corresponds to homozygous loci. It can be used with dominant marker data (RAPD and AFLP), because absences could correspond to homozygous recessives. An example of application of the Simple Matching Coefficient for categorical variables is found in Appendix 6 (click here).

The **Jaccard Coefficient** only counts bands present for either individual (i or j). Double absences are treated as missing data. If false-positive or false-negative data occur, the index estimate tends to be biased. It can be applied with codominant marker data.

The **Nei-Li Coefficient** counts the percentage of shared bands among two individuals and gives more weight to those bands that are present in both. It considers that absence has less biological significance, and so this coefficient has complete meaning in terms of DNA similarity. It can be applied with codominant marker data (RFLP, SSR).

Disequilibrium models: genetic distance

- Measures the difference between two genes, proportional to the time of separation from a common ancestor
- Several models are possible:
  - Mutation of infinite alleles
    e.g. Nei's genetic distance
  - Stepwise mutation model
    e.g. Distance using microsatellites
  - Mutation in the nucleotide sequence

- Mutation of infinite alleles (i.e. isozymes)
  - Each mutation event gives rise to a new allele.
  - If 2 genes are the same, no mutation has occurred. If 2 genes are different, an unknown number of mutations occurred.
  - The average number of mutations since time t when they diverged from an ancestor is = $2t\mu$, where μ is the rate of mutation and is multiplied by 2 because we are dealing with 2 independent genes.
  - The probability that 2 genes are common by descent after time t is $P= e^{-2t\mu}$.

- Stepwise mutation model (i.e. SSRs)
  - Mutation is a progressive change so fragments that migrate similar distances have had few mutations.
  - In the case of SSRs, mutation is assumed to change the number of repeats, increasing or decreasing step by step. It can be shown that the square of the difference in the number of repeats between 2 microsatellites is proportional to the time of divergence from a common ancestor.

- Mutation in the nucleotide sequence
  - It indicates that the simplest substitution is the mutation of a single base.
  - The main limitation is the loss of information by not knowing the number of mutations that could have taken place at one site. To solve that problem, some methods assume the probability of transition (purine $\rightarrow$ purine or pyrimidine $\rightarrow$ pyrimidine) and transversion (purine $\rightarrow$ pyrimidine or pyrimidine $\rightarrow$ purine).

Calculating Nei's genetic distance

- The standard Nei's genetic distance is:

$$D_{XY} = -\ln (I_{XY})$$

- It is based on the concept of genetic identity (I_{XY}):

$$I_{xy} = \frac{J_{xy}}{\sqrt{(J_x J_y)}}$$

(continued on next slide)

Where,

J_X = average homozygosity in population X

J_Y = average homozygosity in population Y

J_{XY} = average interpopulation homozygosity

Such that,

I_{XY} = 1, if two populations have the same allele frequencies in all sampled loci

I_{XY} = 0, if two populations do not share the same allele frequencies in all sampled loci

- The value of D_{XY} varies from 0 (where populations have identical allele frequencies) to infinity (∞, where populations do not share any alleles).
- It assumes that the rate of substitution per locus is equal among all loci and populations.
- This distance estimates the codon differences per locus between two populations.

Calculating Nei's genetic distance (continued)

Locus	Alleles	Allele frequencies		
		Pop.$_1$	Pop.$_2$	Pop.$_3$
A	A_1	0.80	0.74	0.65
	A_2	0.20	0.26	0.35
Locus heterozygosity	h_{ijk}	0.3200	0.3848	0.4550
B	B_1	0.86	0.81	1.00
	B_2	0.01	0.10	0.00
	B_3	0.13	0.09	0.00
Locus heterozygosity	h_{ijk}	0.2434	0.3258	0.0000
D	D_1	0.00	1.00	0.30
	D_2	1.00	0.00	0.70
Locus heterozygosity	h_{ijk}	0.0000	0.00	0.4200
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	$\mathbf{J_{ii'}}$	$\mathbf{J_{1,2} = 0.8733}$	$\mathbf{J_{1,3} = 0.9346}$	$\mathbf{J_{2,3} = 0.8986}$
Genetic identity	$\mathbf{I_{ii'}}$	$\mathbf{I_{1,2} = 0.9183}$	$\mathbf{I_{1,3} = 0.9894}$	$\mathbf{I_{2,3} = 0.9570}$
Genetic distance	$\mathbf{D_{ii'}}$	$\mathbf{D_{1,2} = 0.0852}$	$\mathbf{D_{1,3} = 0.0107}$	$\mathbf{D_{2,3} = 0.0440}$

We have an example where i = 3 populations, j = 3 polymorphic loci, and there are 10 monomorphic loci with. Moreover there are different numbers (K) of alleles per locus (e.g. A and D have 2 alleles each and B has 3 alleles).

The table shows the results of calculating the allele frequencies in each population, as well as the locus heterozygosity. We then calculate the average heterozygosity and homozygosity (1 - heterozygosity) per population.

Lastly, we calculate inter-population homozygosity and genetic identity, so that we can estimate Nei's genetic distance. Calculations are as follows:

$j_{ii'jk} = \Sigma_{ii'j}\, p_{ijk}\, p_{i'jk}$, for example, $j_{1,2jk}$ = homozygosity among populations 1 and 2

$j_{1,2jk} = (0.8)(0.74) + (0.2)(0.26) + (0.86)(0.81) + (0.01)(0.10) + (0.13)(0.09) + (0.0)(1.0) + (1.0)(0.0) + 10 = 11.3533$

$J_{1,2}$ = average interpopulation homozygosity = $j_{1,2jk}/13 = 11.3533/13 = 0.8733$

$I_{1,2}$ = genetic identity among populations 1 and 2 = $J_{1,2}/\sqrt{(J_1 J_2)} = 0.8733/\sqrt{(0.9567 * 0.9453)} = 0.9183$

$D_{1,2}$ = genetic distance among populations 1 and 2 = $-\ln(I_{1,2}) = -\ln(0.9183) = 0.0852$

Because we have not yet explained clustering methods, we present the distance matrix and dendrogram of this example in Appendix 7 (click here).

Calculating intrapopulation distance, using microsatellites

- Intrapopulation distance is the average of the sum of squares of the differences in the number of repeats between alleles

$$S_{wi} = \frac{2}{2n(2n-1)} \sum_{i<i'} (a_{ii} - a_{i'i'})^2$$

- The average intrapopulation distance may be calculated for all analysed loci (d_s)

$$S_w = (1/d_s) \sum_j S_{wj}$$

Where,

a_{ij} = size of the allele of the i^{th} copy (i = 1, 2, …, 2n) in the j^{th} population (j = 1, 2, …, d_s)

n = number of individuals in the sample

Two considerations:

The calculation of distance between two alleles is a transformation of the number of repeats.

One difficulty in using SSRs to estimate genetic distances is their high rate of mutation.

Calculating interpopulation distance, using microsatellites

This is the interpopulation component for the average distance among all allele pairwise comparisons

$$S_B = \frac{2}{(2n)^2 d_s (d_s - 1)} \sum_{j<j'} \sum_{i<i'} (a_{ij} - a_{i'j'})^2$$

The global distance is the weighted average among the component intra- and interpopulations

$$\hat{S} = \frac{2n-1}{(2nd_s - 1)} S_w + \frac{2n(d_s - 1)}{(2nd_s - 1)} S_B$$

These coefficients represent the probability of choosing two different copies of one locus in the same population and between two populations.

Useful computer software: MICROSAT, BIOSYS, GENEPOP, GDA and POPGENE.

Displaying relationships: classification or clustering

Is the process of grouping (or clustering) objects in categories or classes based on their common attributes or relationships. Grouping can be:

- Hierarchical:
  - Essentialist, which tries to unveil the true nature or form
  - Cladistic, which is based on genealogy or phylogeny
  - Evolutionary, which is based on phylogeny and the quantity of evolutionary changes
  - Phenetic, which is based on the highest number of traits of an organism and its life cycle
- Nonhierarchical
- Overlapping

- Hierarchical: a major class that contains minor classes termed ‘branches’.
- Nonhierarchical: each individual is assigned to a unique group by comparing it with the initial classes so that its positioning is the most appropriate.
- Overlapping: individuals may belong to more than one group.

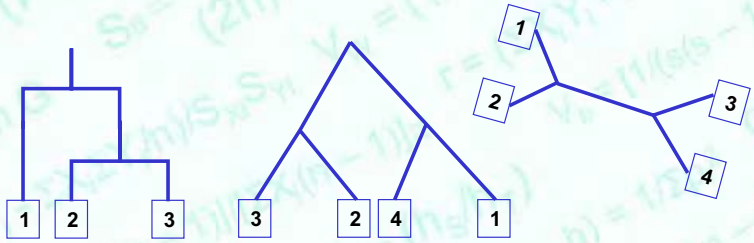
Classification types refer to procedures of cataloguing objects, organisms, etc., and are used in several fields of knowledge. In our case, we use the hierarchical classification because of the nature of relationships between individuals, that is, individual, population, accession, variety, etc., are units that cannot be assigned to two different groups simultaneously.

Reference

García, J.A., M.C. Duque, J. Tohme, S. Xu and M. Levy. 1995. SAS for Classification Analysis; Agrobiotecnology Course, October 1995. Working document. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Phenetic classification

- Shows the relationships among samples by using a similarity index
- A grouping method or distance is selected so that a tree diagram (dendrogram) or a phenogram (if the similarity matrix contains phenotypic data) can be drawn



In this example of hierarchical grouping, all characters are given the same weight in the grouping process.

Total similarity between two groups is the sum of similarity for each character.

It does not consider genealogy.

Phenetic refers to any character used in the classification procedure, whether morphological, physiological, ecological, molecular or cytological.

Clustering methods

- Clustering steps:
  - Proximity is defined
  - Each grouping is estimated according to distance
  - The branches of the dendrogram are built in each cycle
- Three main methods are:
  - Simple linkage (or 'nearest neighbour')
  - Complete linkage (or 'farthest neighbour')
  - Average linkage (or UPGMA)

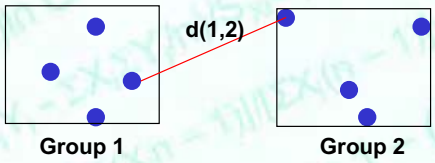
Other methods for clustering are available such as:

- Unweighted pair-group method using the centroid (UPGMC). It is based on the distance between the mean value for each group.
- Weighted pair-group method using the centroid (WPGMC). It takes the OTUs' median value in the groups.
- Ward. It works with the sum of the squared distances for pairs of OTUs. It is also known as the method of minimal variance because, while taking the squared values, it becomes a very sensitive method (different OTUs will look more dissimilar and similar OTUs will look even closer). It may be used with Euclidian distances and molecular data when a high number of DNA bands is available.

In the next few slides, we discuss in more detail the three methods listed on the slide above and show an example for each.

Simple linkage

- Or ‘nearest neighbour’
- It minimizes the inter-group distance by taking the distance to the neighbour with the highest similarity
- It works with regular and compact groups, but is highly influenced by distant individuals. This is inconvenient when there are different groups that are not well distributed in space



d(1,2) = minimum distance between two OTUs

Simple linkage: an example

(1)

	A	B	C	D
A	0			
B	0.30	0		
C	0.43	0.35	0	
D	0.28	0.60	0.40	0

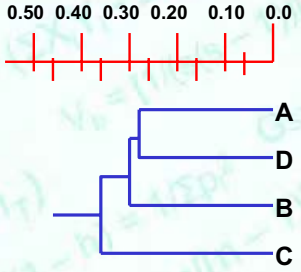
(2)

	B	C	AD
B	0		
C	0.35	0	
AD	0.30	0.40	0

(3)

	C	ADB
C	0	
ADB	0.35	0

(4)



1. The distance matrix is formed first; then, in a first cycle, the shortest distance is selected $d_{AD} = 0.28$.

2. A new matrix is formed by grouping individuals A and D and calculating the combined distances:

   $d_{B(AD)} = \min (d_{BA}; d_{BD}) = \min (0.30; 0.60) = 0.30$
   $d_{C(AD)} = \min (d_{CA}; d_{CD}) = \min (0.43; 0.40) = 0.40$

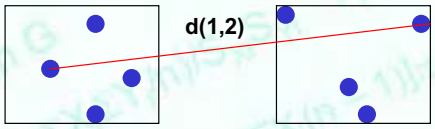
3. A new matrix is formed by grouping individual B with group (AD) and calculating the combined distances

   $d_{C(ADB)} = \min (d_{AC}; d_{CD}; d_{CB}) = \min (0.43; 0.40; 0.35) = 0.35$

4. The dendrogram is drawn.

Complete linkage

- Or ‘farthest neighbour’
- It minimizes the inter-group distance by taking the distance to the individual with minimal similarity
- It works well with regular and compact groups but, again, it is influenced by distant individuals



d(1,2) = major distance between two OTUs

Complete linkage: an example

(1)

	A	B	C	D
A	0			
B	0.30	0		
C	0.43	0.35	0	
D	0.28	0.60	0.40	0

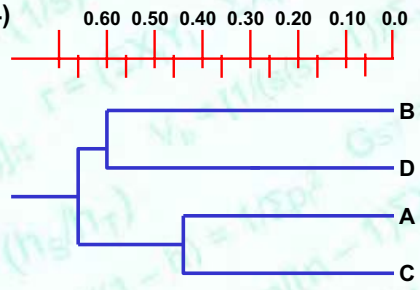
(2)

	A	C	BD
A	0		
C	0.43	0	
BD	0.30	0.40	0

(3)

	AC	DB
AC	0	
DB	0.40	0

(4)



1. The distance matrix is formed first; then, in a first cycle, the longest distance is selected, $d_{BD} = 0.60$.

2. A new matrix is formed by grouping individuals B and D and calculating the combined distances:

$$d_{A(BD)} = \max(d_{BA}; d_{AD}) = \max(0.30; 0.28) = 0.30$$
$$d_{C(BD)} = \max(d_{CB}; d_{CD}) = \max(0.35; 0.40) = 0.40$$

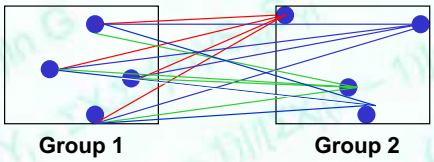
3. The new matrix is formed with groups AC and BD, and the combined distances calculated:

$$d_{(AC)(DB)} = \max(d_{AD}; d_{AB}; d_{CD}; d_{CB}) = \max(0.28; 0.30; 0.40; 0.35) = 0.40$$

4. The dendrogram is drawn.

Average linkage

- Or ‘unweighted pair-group method using the arithmetic average’ (UPGMA)
- It minimizes the inter-group distance by taking the average pairwise distance among all individuals of the sample
- Most used method



$d(1_i, 2_j)$ = average distance between OTU_i and OTU_j of groups 1 and 2

Average linkage: an example

(1)

	A	B	C	D
A	0			
B	0.30	0		
C	0.43	0.35	0	
D	0.28	0.60	0.40	0

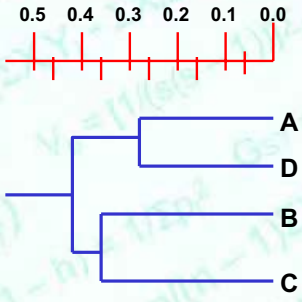
(2)

	B	C	AD
B	0		
C	0.35	0	
AD	0.45	0.415	0

(3)

	BC	AD
BC	0	
AD	0.42	0

(4)



1. The distance matrix is formed first; then, in a first cycle, the shortest distance is selected, $d_{AD} = 0.28$

2. Next, a matrix is formed by grouping individual A with D and calculating the combined distances:

$$d_{B(AD)} = (d_{BA} + d_{BD})/2 = (0.30 + 0.60)/2 = 0.45$$
$$d_{C(AD)} = (d_{CA} + d_{CD})/2 = (0.43 + 0.40)/2 = 0.415$$

3. A new matrix is formed by grouping the individuals with the shortest distance B with C, and calculating the combined distances:

$$d_{(AD)\,(BC)} = (d_{AB} + d_{AC} + d_{BD} + d_{BC})/4 = (0.30 + 0.43 + 0.60 + 0.35)/4 = 0.42$$

Choosing a clustering method

- First, gather knowledge of the species under study such as its diversity, reproduction system, ploidy number and levels of heterozygosity
- Carefully select the genetic characters to analyse
- Test different clustering methodologies and assess the level of agreement obtained with each of them

In addition, it will always be important to combine as much information as possible. An example may be found in Appendix 8 (click here), in which both morphological and molecular data are available, and the use of separate data sets is compared with the use of combined data.

Validating the clustering analysis

- External validation
- Internal validation
- Relative validation
- Bootstrapping

External validation:
: The matrix distance is compared with other information not used in the grouping calculations (e.g. genealogy).

Internal validation:
: This technique quantifies the distortion due to the grouping method used. It builds a new similarity or distance matrix, the 'co-phenetic matrix', directly from the dendrogram. Validation is calculated by means of a correlation coefficient between similarity or distance data from the original matrix and those from the new co-phenetic matrix. Whether the original distances are maintained is assessed after the grouping exercise (Sokal and Rohlf, 1994).

Relative validation:
: Similarity between methods is compared.

Bootstrapping:
: This is a re-sampling method by replacement with the same data matrix. It allows calculation of standard deviations and variances, and is useful for those situations in which the number of samples or resources (e.g. time, budget) is limited.

Examples of applying the co-phenetic correlation and bootstrapping methods are shown next.

Reference

Sokal, R. and J. Rohlf. 1994. Biometry: The Principles and Practice of Statistics in Biological Research (3rd edn.). Freeman & Co, NY.

Co-phenetic correlation: an example

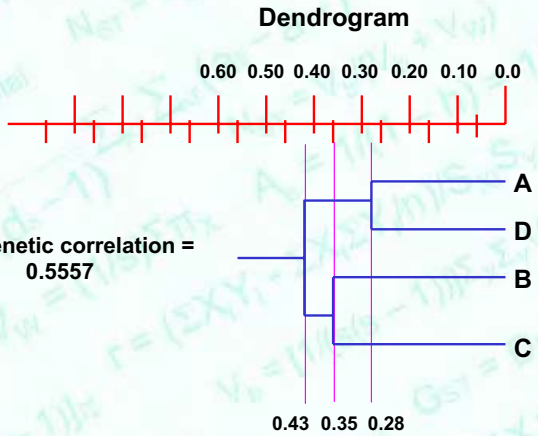
	A	B	C	D
A	0			
B	0.30	0		
C	0.43	0.35	0	
D	0.28	0.60	0.40	0

Original distance matrix

	A	B	C	D
A	0			
B	0.43	0		
C	0.43	0.35	0	
D	0.28	0.43	0.43	0

Co-phenetic matrix

Co-phenetic correlation = 0.5557



To construct the co-phenetic matrix, we look at the dendrogram previously built with the original matrix (this example comes from slide 58). We see that the distance between D and C in the dendrogram is 0.43, so we fill that cell in the co-phenetic matrix. Distance between B and C is 0.35, and so on.

Calculations for the co-phenetic correlation are based on the correlation coefficient:

$$r = (\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i / n) / S_{Xi} S_{Yi}$$

Where,

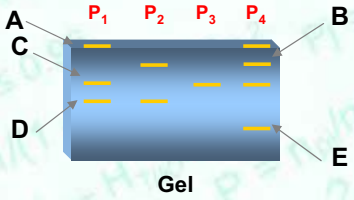
X_i and Y_i are the similarity or distance values of the original and co-phenetic matrix, respectively

S_{Xi} and S_{Yi} are the standard deviations for each variable

If the correlation value is high, we can conclude that the dendrogram does indeed reflect the distances in the original matrix and that therefore there is no distortion due to the grouping method. In the above example, we obtained a value 0.5557. This is an average value that could indicate that the dendrogram distances do not reflect the distance data in the original matrix, and so distortion exists because of the method used. However, in building the example, we used very few data; nor were they the real results of an experiment, thus explaining the value obtained.

Bootstrapping validation: an example

(1)



Gel

(2)

	P_1	P_2	P_3	P_4
L_1	1	0	0	1
L_2	0	1	0	1
L_3	1	0	1	1
L_4	1	1	0	0
L_5	0	0	0	1

Data matrix

(3)

	P_1	P_2	P_3	P_4
P_1	1			
P_2	0.400	1		
P_3	0.600	0.400	1	
P_4	0.400	0.200	0.400	1

Similarity matrix

(continued on next slide)

In the gel above (top left corner), we have 4 individuals (P_i) and 5 loci (L_j). We will suppose we perform the validation in three samples with replacement.

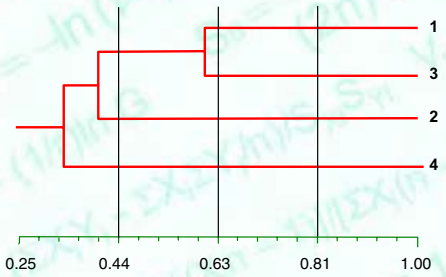
First, we score the marker data in the individuals (data matrix) and then, we calculate the average similarity (simple matching) and its interval.

Bootstrapping validation: an example (continued)

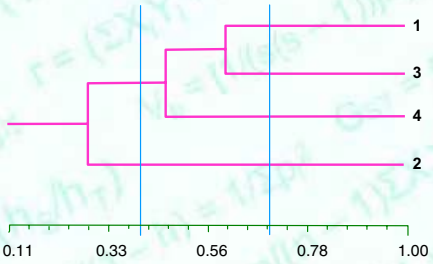
Average similarity matrix with standard deviations

	P_1	P_2	P_3	P_4
P_1	1			
P_2	0.267 ± 0.115	1		
P_3	0.600 ± 0.000	0.400 ± 0.200	1	
P_4	0.533 ± 0.115	0.200 ± 0.000	0.400 ± 0.200	1

Dendrogram before replacement



Dendrogram with replacement



For each individual, the value for each locus is taken, one by one, with replacement and a sample formed of equal size to the number of loci. The possibility exists that a locus is selected one or more times. For the example:

M_1: L_1 L_1 L_2 L_3 L_5 (locus L_4 was not drawn)
M_2: L_1 L_2 L_3 L_4 L_3
M_3: L_3 L_1 L_5 L_2 L_4

In each sample a similarity matrix is calculated.

Average similarities and their standard deviations are estimated for each individual pair (1 & 2, 1 & 3, 2 & 3, and so on), and the average similarity matrix is created.

A new dendrogram is built, using the average similarity matrix.

For real situations, more than 100 replacement samples should be created.

Displaying relationships: ordination

- Ordination is the arrangement or 'ordering' of sample units along coordinate systems
- The purpose of ordination, as well as classification methods, is to interpret patterns in the composition of samples

Ordination is a multivariate method that complements clustering, and is usually considered to be an approach that is closer to biological reality.

With ordination methods, we want to represent the relationships of samples in a simple way by reducing the real situation to a 'low dimensional space' (Gauch, 1982). In doing so, the sample's composition is studied as a whole, the statistical power of the analysis is improved because redundancy is somehow eliminated or reduced, and the relative importance of different gradients can be determined. Most of all, we get graphical representations that help us intuitively interpret the relationships of the different groups of samples.

In principle, ordination is both an exploratory and hypothesis-testing tool. In any case, the results obtained with ordination methods should always be contrasted with the available knowledge of the sample being studied and as much as possible with additional information related to the biological question being addressed in the research.

Reference

Gauch, H.G., Jr. 1982. Multivariate Analysis and Community Structure. Cambridge University Press, UK.

Useful ordination methods for molecular marker data

- Principal coordinates analysis (PCoA)
- Nonmetric multidimensional scaling (NMDS)
- Correspondence analysis (CA)

Many ordination techniques exist—some are based on distance data or on the calculations of the so-called Eigenvalues (the sum of all variances for each character in each component). However, those techniques based on continuous variables (e.g. principal component analysis or PCA) are not appropriate for use with marker data. Hence, we discuss only briefly the three listed in the slide above. More details on the basics of these methods would require a deeper mathematical understanding of the algorithms involved than what we expect from the average user of this module. We therefore encourage our readers who want to know more about these methods to search for ordination methods through the Web. For an overview, check the site <http://www.okstate.edu/artsci/botany/ordinate/overview.htm>

Principal coordinates analysis (PCoA) attempts to represent distances between samples and may accommodate matrices from different dissimilarity measures. It maximizes the linear correlation between sample distances. When used with Euclidean distances, the results are identical to PCA.

Nonmetric multidimensional scaling (NMDS) works by maximizing the rank order correlation and attempting to find the best shape to accommodate the data. This technique uncovers the basic configuration from the dissimilarity sample matrix. With NMDS, only the pattern of points is relevant, not the origin, and the representation may be rotated.

Correspondence analysis (CA) repeats the averages of sample scores and finds spots where all samples falling in the same spot are as similar as possible and, simultaneously, samples at different spots are as different as possible.

Appendices

2. Analysis of molecular variance: example 1
3. Analysis of molecular variance: example 2
4. Geometric distance
5. Transforming data from quantitative variables: an example
6. Applying the simple matching coefficient for morphological characters (categorical variables)
7. Calculating Nei's genetic distance
8. Morphological and molecular similarities

In summary

- The analysis of genetic diversity and structure of populations involves:
  - The quantification of diversity and the relationships within and between populations and/or individuals
  - The display of relationships
- Molecular data are usually handled as binary data
- Molecular data can be usefully complemented with morphological or evaluation data. To do so, there types of variables can be transformed to binary variables

By now you should know …

- The basic steps involved in measuring genetic diversity
- The major ways to describe genetic diversity within and among populations
- The correct selection of distance calculation to assess relationships in the sample of interest
- The differences between alternative clustering methods
- The options available to validate grouping
- The basic notions underlying the concept of ordination
- The similarities and differences between clustering and ordination

References

Cavalli–Sforza, L.L. and W.F. Bodmer. 1981. Genética de las Poblaciones Humanas. Ed. Omega, Barcelona.

García, J.A., M.C. Duque, J. Tohme, S. Xu and M. Levy. 1995. SAS for Classification Analysis; Agrobiotecnology Course, October 1995. Working document. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Gauch, H.G., Jr. 1982. Multivariate Analysis and Community Structure. Cambridge University Press, UK.

Karp, A., P.G. Isaac and D.S. Ingram. 1998. Molecular Tools for Screening Biodiversity: Plants and Animals. Chapman & Hall, London.

Lynch, M. and B.G. Milligan. 1994. Analysis of population genetic structure with RAPD markers. Mol. Ecol. 3:91-99.

Sokal, R. and J. Rohlf. 1994. Biometry: The Principles and Practice of Statistics in Biological Research (3rd edn.). Freeman & Co, NY.

Next

- Software programs for analysing genetic diversity
- Glossary