Genetic diversity analysis with molecular marker data: Learning module

M.C. de Vicente (IPGRI),
C. Lopez (Universidad Nacional Agraria ‘La Molina’, Peru) and
T. Fulton (Institute for Genomic Diversity, Cornell University)
Foreword

*Genetic Diversity Analysis with Molecular Marker Data: Learning Module* is the second of two learning modules produced with the goal of promoting the educated application of molecular techniques in biodiversity studies. It complements the first module, *Using Molecular Marker Technology in Studies on Plant Genetic Diversity: Learning Module*, and likewise aims to spread knowledge about molecular marker technology for assessing genetic diversity, thus providing a basis for understanding data analysis and interpretation beyond the recurring use of technologies just for the sake of being fashionable.

The motive that triggered the production of this learning module was a training course on molecular technologies organized by IPGRI and carried out in China with the participation of scientists involved in genetic diversity of tropical fruits from China, India, Indonesia, Malaysia, the Philippines, Sri Lanka and Thailand. At the end of the course, it was realized that teaching the techniques had made very good laboratory technicians. However, if scientists competent in genetic resources management and use are to be formed, then IPGRI’s partners must be provided with data analysis and interpretation tools.

The learning module is especially directed to scientists who are interested in assessing genetic diversity, and who have a basic knowledge of biology and genetics, are familiar with molecular technologies, and need guidance on how to tackle experimental planning, and analyse and interpret their results.

The authors hope the users of these learning modules will not only find the information provided enlightening but also practical, and that the use of molecular marker technologies can be an exciting and viable approach for anyone planning research on genetic diversity analysis.
Objectives of the module

Our goals for the module are that users:

- Understand the scientific concepts of genetic diversity through the fundamentals of population genetics
- Become familiar with the mathematical expressions used to describe genetic diversity and be able to perform the indispensable calculations based on molecular marker data
- Acquire the basic knowledge to apply molecular technologies for assessing genetic diversity and to interpret molecular data accordingly
Acknowledgements

The production of *Genetic Diversity Analysis with Molecular Marker Data: Learning Module* was made possible through a collaborative effort between the International Plant Genetic Resources Institute (IPGRI) and the Institute for Genomic Diversity (IGD) of Cornell University. The authors especially thank the following contributors:

The IPGRI Asia, the Pacific and Oceania Regional Office who, through its Tropical Fruit Trees Project, itself funded by the Asian Development Bank (ADB), partly funded the development of this material.

Félix Alberto Guzmán (IPGRI Americas Regional Office), for his help in simplifying some of the information provided and for adapting the final version with the corrections received by different colleagues and the editor.

Amanda Garris (IGD) and Steve Tanksley (Department of Plant Breeding) both at Cornell University, for allowing us to use information and figures from their slides.

Joanne Labate for ‘virtual discussions’ and Alexandra Casa and Martha Hamblin of the IGD for advice on which programs to discuss.

Humberto Gómez Paniagua (Centro Internacional de Agricultura Tropical [CIAT], Colombia), for his helpful comments to improve both the content and didactic presentation of this product.

Myriam Cristina Duque (CIAT), for her help in preparing the outline for this module and clarifying concepts on measures of genetic diversity. Marilyn Warburton (Applied Biotechnology Center, Centro Internacional para Mejoramiento de Maíz y Trigo [CIMMYT], Mexico) for her suggestions on how to improve the slides on clustering methods.

The staff at the Genetic Resources Unit, CIAT (Colombia) for sharing the picture of beans for use as background in one of the submodules.

Professor Heiber Cárdenas (Biology Department, Universidad del Valle, Cali, Colombia), for testing the learning module with five of his students (Iván Andrés González, María Fernanda Castillo, Carmenza Montoya, Fernando Rondón González and Diego Mauricio Villamarín Miranda), and for providing valuable comments for the module’s improvement.

Luigi Guarino, Ramanatha Rao, Issiaka Zoungrana, Margarita Baena, Zongwen Zhang, Mikkel Grum, Jan Engels and Elisabeth Goldberg (different IPGRI offices), for their comments and suggestions on how to improve this learning module and hopefully make it more useful to our partners.

Elizabeth L. McAdam, for her help in manuscript editing and her good suggestions to improve the format of this product, and Lynn Menéndez (CIAT), for coordinating the editing of the manuscript and also for suggesting several very helpful modifications.

Gladys Rodríguez (CIAT), for her help in laying out this product.
Genetic diversity analysis with molecular marker data: Learning module

About this module
Objectives
I  Introduction
II  Basic concepts of population genetics
III  Measures of genetic diversity
IV  Software programs for analysing genetic diversity
V   Glossary
VI  Feedback form
What you should know about the module

When planning research, it is crucial that the questions you address are well defined, even if other questions come along and you want to accommodate changes into the original plan as you progress. You must also know some of the principles of setting up experiments to ensure convincing results. In the context of your work with genetic resources, you need to be familiar with the foundations of genetic diversity and the available tools for examining your data so you can correctly interpret it.

Some of the concepts and tools we discuss include:

- Sampling strategies
- Basics of population genetics
- The mathematical measures used to describe genetic diversity, genetic distance indices and the methods used to express relatedness among samples
- Software and available Internet resources

The calculations involved in the key concepts of population genetics, measures of genetic diversity, indices and clustering methods are illustrated with examples that were prepared precisely for this publication. The user can then immediately see how mathematical expressions are applied. In addition, we want to show that, even though these calculations are often performed with the help of a computer program, most can be done manually. Even if a computer program is used, we believe the scientist should understand what the computer does and be able to develop his/her own criteria for deciding what method to use.

This learning module is intended to help those who want to analyse genetic diversity with the aid of molecular data. As such, it is not a comprehensive tool for learning about or teaching population genetics. However, we have listed bibliographic references to support significant notions of population genetics, to back up the mathematical expressions and provide for better comprehension of how the methods are applied in real-life research situations. The idea is that the learning module can be followed as a stand-alone tool to facilitate the learning process, in particular to help students, especially undergraduate students, understand how to choose amongst the applied molecular techniques in genetic diversity studies and to be able to apply them in their research projects. It can also be used for teaching and preparation of classroom lectures or as a reference guide by working professionals who need to apply molecular techniques, statistical methods and data analysis necessary to carry out research.

Users of this module must have basic knowledge of genetics. In addition, if they are not familiar with molecular marker technologies, we strongly recommend beginning the training by undertaking the first module Using Molecular Marker Technology Effectively in Plant Genetic Resources: Learning Module. Once the basic principles of these techniques are understood, the second module can be followed with a better basis to understand mathematical algorithms.

We have organized the learning module into complementary and independent submodules, so that the user may select at any point the section of interest. In some cases, we have given extra information, added as appendixes so not to
complicate the essential, yet providing further mathematical expressions or examples for those who would find them beneficial for a thorough understanding. This learning module is presented in such a way that either only slides may be used in a presentation format or both slides and accompanying notes.

Feedback on this publication is extremely important for us because we are convinced that it can always be enhanced. To respond effectively to our partners and other users’ needs, we would greatly appreciate your giving us feedback on the organization, content and usefulness of this tool at the following addresses:

M. Carmen de Vicente
International Plant Genetic Resources Institute (IPGRI)
Via dei Tre Denari 472/a
00057 Maccarese
Rome, Italy
E-mail: cdevicente@cgiar.org

César López
Universidad Nacional Agraria ‘La Molina’
Av. La Universidad s/n. Apdo. 456 Lima 12
Lima, Peru
E-mail: cflb@lamolina.edu.pe

Theresa Fulton
Institute for Genomic Diversity
130 Biotechnology Building
Cornell University
Ithaca, New York 14583
E-mail: tf12@cornell.edu

We sincerely hope that this module complements the previous module, Using Molecular Marker Technology in Studies on Plant Genetic Diversity: Learning Module, and that the two learning modules together offer our partners, especially those in developing countries with limited access to state-of-the-art technologies and comprehensive scientific literature or instruction, a chance to conduct advanced research in plant genetic diversity, thereby contributing to the world's knowledge of these valuable resources.

M. Carmen de Vicente
IPGRI

César López
Univ. ‘La Molina’

Theresa Fulton
IGD, Cornell University
Genetic diversity analysis with molecular marker data: Learning module

Introduction
Contents

▶ The scientific method
  • Defining the biological question
  • Developing a hypothesis and the experimental design
    – Sampling within populations
    – Sampling within the genome
  • Conducting experiments
  • Analysing and interpreting data

▶ Levels of biological diversity

▶ Measuring genetic variation

▶ Relationships between phenotype and genotype
The scientific method

- Defining the biological question
- Developing a hypothesis and the experimental design
- Conducting experiments
- Analysing and interpreting data

The scientific method starts with defining a biological question, that is, the reason why the research is being undertaken. After that, an iterative process takes place, going through several steps that guide towards result analysis. This process may be done several times before the final results are obtained. Results should provide evidence to prove or reject the hypothesis that supported the experiment’s design. If the hypothesis is proved, then a satisfactory answer has been obtained to the opening question. If the hypothesis is rejected, then a new hypothesis must be postulated and another experiment designed. The process then starts all over again, and so on in iteration until a satisfactory interpretation is reached.

Sometimes, when the research interest is in assessing genetic diversity, the initial question is replaced with a need for description. In this case, a hypothesis does not exist and the experiment is designed to gather information that will account for the existing variation.

In the following slides, each step mentioned above is discussed.
Defining the biological question

- What is the problem?
- Gathering state-of-the-art information on the subject:
  - Has it been studied already?
  - Does an explanation exist?

The first step is to define the biological question underlying our research interest. We may look for either an answer to a question or a description. In genetic diversity studies, we often start with seeking a description, for example, How much variation is there? How is the variation organized? The results we eventually obtain will most likely lead us to questions, which must already have been implicit in the descriptive analysis.

Searching the literature will help us define our subject of interest, gather information on a broader context and, consequently, identify clues for the hypotheses we may need to propose.
What kind of questions?

Ask questions on:

- Conservation strategies
- Use for crop breeding
- Evolution and domestication

As mentioned previously, many questions may lead to the need for research. Examples of questions are those:

- Related to conservation:
  - How is diversity represented in nature and collections?
  - What are the priorities for conservation?
  - Using only finite resources, how do we insure for the future?
  - How can we quantify current diversity for future reference for studies on genetic erosion?
  - Does a pattern of distribution exist that can be used to guide our collection activities?
  - Can we ensure that our samples are distinct?
  - Do these apparently different samples belong to different taxa?

- Related to crop breeding:
  - How can our sample be used to support crop improvement?
  - Are these genetic resources likely sources for allelic diversity?
  - Are they good sources for desirable agronomic traits?

- Related to evolution:
  - Where did the crop originate?
  - Which is the progenitor of a particular crop?
  - Has introgression occurred among population samples of different origins?
An example of two questions:  
How much variation is there?  
How much of it can we afford to lose?

<table>
<thead>
<tr>
<th>Loci</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>A₁</td>
<td>B₁</td>
<td>C₁</td>
<td>D₁</td>
<td>E₁</td>
<td>F₁</td>
<td>G₁</td>
<td>H₁</td>
<td>I₁</td>
</tr>
<tr>
<td></td>
<td>A₂</td>
<td>B₂</td>
<td>C₂</td>
<td>D₂</td>
<td>E₂</td>
<td>F₂</td>
<td>G₂</td>
<td>H₂</td>
<td>I₂</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C₃</td>
<td></td>
<td></td>
<td>E₃</td>
<td></td>
<td>G₃</td>
<td></td>
<td>I₃</td>
</tr>
<tr>
<td>Genotype</td>
<td>A₁A₁</td>
<td>B₁B₁</td>
<td>C₁C₁</td>
<td>D₁D₁</td>
<td>E₁E₁</td>
<td>F₁F₁</td>
<td>G₁G₁</td>
<td>H₁H₁</td>
<td>I₁I₁</td>
</tr>
<tr>
<td></td>
<td>A₂A₂</td>
<td>B₁B₂</td>
<td>C₁C₂</td>
<td>D₁D₂</td>
<td>E₁E₂</td>
<td>F₁F₂</td>
<td>G₁G₂</td>
<td>H₁H₂</td>
<td>I₁I₂</td>
</tr>
<tr>
<td></td>
<td>A₂A₂</td>
<td>B₂B₂</td>
<td>C₁C₂</td>
<td>D₁D₂</td>
<td>E₁E₂</td>
<td>F₂F₂</td>
<td>G₂G₂</td>
<td>H₁H₂</td>
<td>I₁I₂</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>etc.</td>
</tr>
</tbody>
</table>
| Total number of genotypes | 3\(K(K + 1)/2\)
| Genotypes (no.) | 3 | 3 | 6 | 1 | 10 | 6 | 1 | 6 | 58320 |

How much variation is there? and How much of it can we afford to lose?

These two questions are examples of what we would ask to account for the amount of variation present in a sample and, as a result, estimate the amount of variation we could lose. The slide shows a simplified situation of an organism for which we analysed 9 loci, each with different number of alleles in the sample (loci A and B each has 2 alleles, locus C has 3 alleles, locus D only 1 allele and so on). The second part of the table shows the genotypes for each locus: A₁A₁, A₁A₂, A₂A₂, B₁B₁ and so on.

The calculation is provided for the total number of genotypes per locus, where K = the number of alleles per locus. For each locus, the number of genotypes is calculated (e.g. locus A = 3, locus B = 3, locus C = 6, locus D = 1, ...). Then, the total number of genotypes possible in this organism is calculated based on these 9 loci (3 x 3 x 6 x 1 x 10 x 3 x 6 x 1 x 6 = 58 320 genotypes).

In the last two rows of the table, 2 genotypes of the 58 320 possible are shown. In a conservation strategy, if only 2 genotypes were selected at random, a high percentage of genetic variation would be lost.
Developing a hypothesis and the experimental design

- How can the question be best approached?
- Are there alternative strategies?
- Considerations of cost, time and available resources
- Sampling strategies
  - Populations versus individuals
  - Genome

A good hypothesis is an assumption about an observation and should be amenable to experimental validation, which later serves to build broad conclusions. Defining a good hypothesis for studying genetic diversity requires a basic knowledge of, for example, reproduction systems, spatial and temporal distribution, and interspecific interactions. For instance, if we want to analyze a mainly self-pollinating species, our hypothesis might be based on the fact that the species will have low intrapopulation polymorphism and high polymorphism among populations. Or, if individuals of a given species are highly extended in a region, our hypothesis may assume that populations in that region will have similar allelic frequencies.

Once the initial question is defined, the hypothesis will help us design the experiment. Normally, we could use one of several strategies. The strategy finally chosen will depend on the resources available, that is, infrastructure, expertise, time, and funds. Also critically important is to think about sampling strategies: (1) sampling individuals in our target populations and (2) sampling the genome. In studies with molecular markers, sampling the genome implies to decide on the number of points along the DNA strands (DNA sequence or chromosomes) to ensure a data set that would adequately respond to our original question (i.e. how much precision do we need?).

Next, we will discuss sampling in more detail.

For more details on some of these issues, look at the training module Using Molecular Marker Technology in Studies on Plant Genetic Diversity (www.ipgri.cgiar.org).
Sampling within populations

- How many populations and/or individuals do we need to reach the objective of our research?
- Is the objective of our research realistic?

Each approach to measuring variation and its structure in populations may require a decision on a specific, optimal, sampling strategy.

Two major criteria help identify the optimal sampling procedure, regardless of the research question being addressed (Brown and Weir 1983): (1) the strategy of choice may be considered optimal if the sampling variance is minimized per unit of experimentation; and (2) ease of operation and available resources.

While no comprehensive publication dealing with questions for all situations exists, treatment of some situations can be found in the literature. For example, Brown and Weir (1983) on sampling to estimate allele frequency, number of alleles per locus and gene diversity. They, in turn, mention basic literature that may be relevant to the subject. Another example, Gregorius (1980) gives ideas about sampling procedures and presents a table for calculating minimum sample sizes needed to ensure that all alleles, with frequencies higher than a defined threshold, are detected with a given probability.

In all cases, however, it will be important to assess and balance the intended results of the research against the available resources. For more information, see the section Final Considerations in the training module Using Molecular Marker Technology in Studies on Plant Genetic Diversity (http://www.ipgri.cgiar.org/publications/pubfile.asp?ID_PUB=912).

References


Sampling within the genome

How many loci should be assayed?

Two strategies for selecting are possible:

- Select a few highly informative markers
- Select numerous, poorly informative, markers randomly distributed within the genome

In terms of genome sampling, the ideal situation would be one where the distribution of genetic polymorphism within the genome is known. However, this is not usually the case and, as such, no a priori knowledge is available on the distribution of polymorphism. Consequently, studies are either performed unthinkingly by reproducing similar experiments already published, or they must be based on simulations. In all cases, comparing the advantages and disadvantages of different marker types is important because they have typically different distributions within the genome.

Mariette et al. (1999) present the results of simulating an experiment in which the power of two marker types with different numbers of loci was assayed: (1) an AFLP experiment with 200 loci scored either as dominant or as codominant data, and (2) a microsatellite experiment with two different situations—one with 5 loci and another with 50 loci. The results were presented in a chart, which showed a trade-off between the number of loci analysed and the amount and accuracy of information obtained. The chart illustrated that both number and information content (dominant inheritance versus codominant) of the marker type used are important for ensuring proper sampling of the genome.

Reference

### Problems of bad sampling

1. If the sample is too small, we get biased estimates of allele frequencies

<table>
<thead>
<tr>
<th>Indivs. (no.)</th>
<th>A₁ A₁</th>
<th>A₁ A₂</th>
<th>A₂ A₂</th>
<th>Allele freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>p=0.75, q=0.25</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>16</td>
<td>10</td>
<td>p=0.4, q=0.6</td>
</tr>
<tr>
<td>100</td>
<td>27</td>
<td>39</td>
<td>34</td>
<td>p=0.465, q=0.535</td>
</tr>
<tr>
<td>1000</td>
<td>299</td>
<td>395</td>
<td>306</td>
<td>p=0.4965, q=0.5035</td>
</tr>
<tr>
<td>1 000 000</td>
<td>300 000</td>
<td>400 000</td>
<td>300 000</td>
<td>p=0.5, q=0.5</td>
</tr>
</tbody>
</table>

(continued on next slide)

The table above shows an example where, for just one gene (A) with two alleles, allele frequencies are calculated for samples of increasing numbers of individuals. When the sample size increases, the chance that all genotypes would be equally represented becomes higher. Hence, the estimate of allele frequencies becomes closer to the actual situation, and the conclusions of the experiment more reliable.
Problems of bad sampling (continued)

2. Some alleles escape sampling, so they are scored as absent:

<table>
<thead>
<tr>
<th>Initial population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of A₁ (p) = 25/96 = 0.260</td>
<td>Frequency of A₁ (p) = 8/36 = 0.222</td>
</tr>
<tr>
<td>Frequency of A₂ (q) = 68/96 = 0.708</td>
<td>Frequency of A₂ (q) = 28/36 = 0.778</td>
</tr>
<tr>
<td>Frequency of A₃ (s) = 3/96 = 0.031</td>
<td>Frequency of A₃ (s) = 0/36 = 0.000 (lost)</td>
</tr>
</tbody>
</table>

The example above shows a situation in which a subsample is taken from a population in such a way that allele A₃, present in the original population, is omitted. The subsample does not contain all the alleles and is therefore not representative of the wider population. The description of the genetic diversity of the smaller sample, as an estimator of that of the bigger population, will not be equivalent or legitimate.
**Conducting experiments**

- Use the appropriate tools to test the hypothesis
- Include quality control measures for the experiment and data collection

To set up an experiment, a researcher must take into account that molecular tools can offer greater depth to diversity studies and that they provide a common ground for measuring and analysing diversity. However, molecular data are often complementary to other characterization data (e.g. morphology, pathology) and the combined analysis of these data may offer a more comprehensive ground for interpretation.

On selecting molecular technologies, attention should be given to:

- Its potential to respond to the biological question
- Options for sampling within the genome
- Type and quality of plant material
- Measures of data quality
- Measures to ensure reproducibility
- Expertise required for data collecting

For more details on the issues mentioned above, see the training module *Using Molecular Marker Technology in Studies on Plant Genetic Diversity* ([www.ipgri.cgiar.org](http://www.ipgri.cgiar.org)).
Analysing and interpreting data

► Does the data support the hypothesis?
► Which analytical tools should we use?
► What we also need:
  • Scientific concepts underlying the notion of diversity, e.g.:
    – Levels of biological diversity
    – Variation between and variation within
    – Relationships between phenotype and genotype
  • Basic concepts of population genetics

Once the data are collected, analysis becomes the next step of the experiment. Note that results are based on a given sample, that is, the values obtained will only be estimators of population parameters.

At this point, different analytical tools, as well as different software, may be available. These will be discussed in the last two sections of this module. Simultaneously, we need to be familiar with the basic scientific concepts behind genetic diversity and population genetics (which is that part of genetics dealing with genetic variation). The following slides and section give a brief overview of these subjects.
Levels of biological diversity

1. Intrapopulation diversity (= genetic diversity)

2. Species diversity

3. Ecosystem diversity

Biodiversity can be divided into categories that describe different aspects of living systems, which scientists measure in different ways.

It is now common practice to define biodiversity in terms of genes (intrapopulation diversity or genetic diversity), species and ecosystems, corresponding to three fundamental and hierarchically related levels of biological organization.

‘Genetic diversity’ refers to the variation of genes within species, that is to the mix of genes contained within individuals. This covers distinct populations of the same species or genetic variation within a population. Ultimately, genetic diversity resides in changes in the sequence of the four base-pairs of the DNA that constitutes the genetic code. New genetic variation is created by gene and chromosome mutations and, in organisms with sexual reproduction, gets spread by recombination. Other types of genetic diversity may be caused by the amount of DNA per cell, or by variants in chromosome structure and number. Selection acts on the pool of current genetic variation and, consequently, it facilitates evolution and artificial selective breeding. Genetic diversity allows populations to adapt to changing environmental conditions.

The species level is generally regarded as the most natural for considering whole-organism diversity. Species are also the primary focus of evolution, and the origin and extinction of species are the main agents of biological diversity in most senses. Diversity at the species level can be measured in several ways. ‘Species richness’ is the number of species in a region. ‘Taxonomic diversity’ also measures the relationship of species to each other.

The definition and classification of ecosystems are loose, because their boundaries are vague. Thus, in practice, it is difficult to assess ecosystem diversity other than on a local or regional basis and then only largely in terms of vegetation. Ecosystems differ from genes and species not only in their composition (they include abiotic components, and are partly determined by soils and climate), but also in their structure and function.
**Measuring genetic variation**

- Variation or polymorphism can be assessed at different levels of organization.
- The distribution of polymorphism is observed for different hierarchical levels within the organization (areas, regions, populations, subpopulations, individuals).

Genetic differences are found between individuals within a population, and differences are also found in allelic frequencies between populations. In all, the relative amount of variation depends on the species, history and environment.

The presence of variants (polymorphism) in a sample may be assessed by its genotypes, alleles, haplotypes or nucleotides. Samples can be hierarchically divided at the level of the species, population or within the population. Sampling size should be sufficient when the goal is to study genetic variation (previous slides). Sampling strategies and size may depend on the organization of the species:

- Few individuals (a newly introduced species)
- Many introduced individuals
- Individuals of different geographical origins
### Relationships between phenotype and genotype

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁, A₂</td>
<td>DNA</td>
</tr>
<tr>
<td>Chromosome</td>
<td>RNA</td>
</tr>
</tbody>
</table>

- **Phenotype**
  - Trait-specific
- **Molecular markers**
  - Neutral diversity
  - DNA or protein
- **DNA sequence**
  - Allelic diversity
- **Expression**
  - RNA level

Analysing genetic variation with molecular technologies gives information at the DNA level. It can be neutral diversity, identified along the DNA sequence in regions whose function is unknown, such as when we use anonymous types of markers (e.g. AFLPs, RAPDs). Or, the diversity can be based on known genes, that is, within the coding regions of the DNA sequence. This diversity affects the expression of those genes and, consequently, the RNA—the nucleic acid in charge of translating the information of the genetic code into proteins. Proteins, in their turn, are the elements that make up the structure of organisms, which means they are responsible for what we see, the phenotype. Hence, genotype and phenotype are closely associated. Phenotypic measures of diversity can also be used and, if correctly taken, they may reflect the molecular constitution of a given individual.
In summary

- Undertaking research for genetic diversity analysis should follow the scientific method.
- To assess genetic diversity, the biological question to be addressed must be defined.
- Guidelines must be followed when sampling, not only for individuals and within populations, but also within the genome.
- Interpreting results entail having a basic knowledge of the concepts behind biological diversity and population genetics.
By now you should know …

- The essential steps for setting up an experiment
- The importance of identifying the key question for your research
- The main points to consider when sampling
- The basic concepts underlying the notion of genetic diversity
References


Next

- Basic concepts of population genetics
- Measures of genetic diversity
- Software programs for analysing genetic diversity
- Glossary
Genetic diversity analysis with molecular marker data: Learning module

Basic concepts of population genetics
Contents

- Definitions
- The Hardy-Weinberg principle
- Examples of calculating allele frequencies
- Reproduction and mating systems
- Forces shaping genetic diversity
- Appendix 1: Critical values of the chi-square distribution
‘Population genetics’ is defined in many ways. In general, we can say that population genetics is the study of the application of Mendel’s laws and other genetic principles to entire populations of organisms instead of just to individuals. Population genetics is also the study of changes in gene frequencies and, as such, is closely related to evolutionary genetics because evolution depends heavily on changes in gene frequencies. A short introduction to the main factors possibly causing changes in genetic diversity may be found in slides 33 to 42 of this section.

Although inspecting all genetic variables present in a population is virtually impossible, we can examine a population through variation in individual phenotypes (description of certain morphological or physiological traits) and genotypes (molecular markers).
Phenotype is …

- The description of all traits of an individual as they concern its morphology, physiology, ecological relationships and behaviour.
- At any given time, the phenotype is the result of the genes of the individual interacting with the environment.

Phenotypic differences can be either qualitative (present or absent) or can be quantitative. Qualitative traits can be classified. Quantitative traits are measured.

Individuals share the same phenotype if they look alike. Some genotypes may have the same phenotype. The distinction between genotype and phenotype is important in those traits that are influenced by the environment: two individuals with the same genotype may result in different phenotypes because of the environment.
Natural populations are phenotypically diverse. The amount of phenotypic diversity is extraordinary and obvious, even at the most spontaneous observation.

A population of closely related individuals will show low variability. This is especially critical if environmental conditions change and that population does not have the variation to cope with the change. The population could rapidly move towards extinction.

Population genetics deals with phenotypic diversity, especially where it is due to differences in the genotypic composition of individuals.
Gene and allele

- A gene is the basic physical and functional unit of heredity, passing information from one generation to the next.
- An allele is any of the alternative forms of a gene that can exist at a single locus.

Not all of a DNA sequence is made up of genes. Genes are those sections of DNA that have a known function. They include a transcribed section and a regulatory element that allows its transcription. The existence of genes is deduced by observing the segregation of variants in the progeny of crosses, whether naturally or artificially produced. This observation was the basis on which Gregor Mendel defined the laws of inheritance at the end of the 19th century.

Genes may have two or more alleles. Indeed, a gene may have so many alleles as to constitute an allelic series for that gene. Alleles belonging to a series may show different patterns of dominance to each other. For example, an allele may show a dominant effect, which means that it expresses its phenotype, even if accompanied by a recessive allele. A recessive allele is one where its phenotype is not expressed in an heterozygous individual. If an allele is codominant, its phenotypic effect will be intermediate in the heterozygote relative to the effect of an homozygous dominant and that of an homozygous recessive.
The term ‘gene’ generally refers to the physical entity that is transmitted from parents to offspring during the reproductive process that influences hereditary traits. An individual’s genotype, then, is the sum of all genes inherited from its parents. Genes determine the composition of proteins and also influence external traits and behaviour.
Genetic variation

- Genetic variation deals with the concept of genotype
- Genetic variation affecting traits exists in most natural populations, but these traits are influenced by the alleles of many genes in addition to the effects of the environment
- It is difficult to trace phenotypic differences to the effects of particular genes

Hidden genetic variation is even more extensive than that observed through the phenotype, so much so that it is virtually impossible for two individuals in a population to have the same genotype at all loci. This genetic variation may be detected through molecular technologies, which reveal polymorphisms, which are then useful as genetic markers. However, even molecular tools are limited and, except for comparisons of entire DNA sequences, most methods are limited to a certain number of genes or loci. Even so, sufficient variation is usually found in samples of genes to make assessing the genetic variation possible for most populations.

As mentioned earlier, because only a small portion of the genome is usually explored in genetic variation studies, questions arise about the reliability of results for extrapolation to natural populations. This becomes an important reason to carefully plan the experiments and pay particular attention to the sampling of both individuals and the loci to be screened.
What is polymorphism?

- ‘Presence of many forms’
- In genetic terms, it refers to the coexistence of two or more alternative phenotypes in a population or among populations. In general, these diverse phenotypes are caused by alternative alleles of one gene
- At the molecular level, polymorphism refers to the coexistence of alternative banding patterns or DNA variants when revealed by a given detection method

A gene or a phenotypic trait is said to be polymorphic if more than one form of the gene or trait exists in a population. Genetic variation, which may cause evolutionary change, is ever-present.

More information about polymorphism in general, and molecular polymorphism in particular, is given throughout the training module Using Molecular Marker Technology in Studies on Plant Genetic Diversity (www.ipgri.cgiar.org/publications).
A population is …

- **Ecologically:**
  A group of individuals of the same species living within a restricted geographical area that allows any two individuals to interbreed

- **Genetically:**
  A group of individuals who share a common gene pool and have the potential to interbreed

Populations are extremely complex entities. In population genetics, the focus is on the local interbreeding unit of a larger population because changes in allele frequencies occur within such limited units and may result in the evolution of adaptive traits. These local interbreeding units are usually called local populations, subpopulations or simply populations. Normally, in a population, members of a species are unevenly distributed. Subdivision of populations are often due to accidents in the environment where they are present. In principle, population size is not infinitely large and does not remain constant.
Three levels of population structure are identified:

- Individual organisms
- Subpopulations
- Total population

A population may be considered as a single unit. However, in many species and circumstances, populations are subdivided in smaller units. Such subdivision may be the result of ecological (habitats are not continuous) or behavioural factors (conscious or unconscious relocation). If a population is subdivided, the genetic links among its parts may differ, depending on the real degree of gene flow taking place.

A population is considered structured if (1) genetic drift is occurring in some of its subpopulations, (2) migration does not happen uniformly throughout the population, or (3) mating is not random throughout the population. A population’s structure affects the extent of genetic variation and its patterns of distribution.

See the following slides and the Glossary for more details on these new concepts (e.g. gene flow, migration).
Gene flow is the passage and establishment of genes typical of one population in the genepool of another by natural or artificial hybridization and backcrossing.

In the drawing above, population Y has a higher frequency of allele a (q = 0.90). Pollinators flying from that population will transport more copies of allele a when they move towards another population X. The resulting effect of gene flow is observed in subsequent generations of population X as an increase in the frequency of the migrant allele a.
Allele frequency

- Allele frequency is the concept used to quantify genetic variation.
- It is defined as a measure of the commonness of a given allele in a population, that is, the proportion of all alleles of that gene in the population that are specifically this type.

An allele is an alternative form of a gene. If a gene corresponds to a specific sequence of nucleotides along a DNA molecule, the alleles represent the different sequences of nucleotides that are possible for that particular locus.

Often the term ‘gene’ is used synonymously with ‘allele’ and, hence, ‘gene frequency’ is sometimes used synonymously with ‘allele frequency’. Allelic differences at a single locus in a population indicate genetic variation. This genetic variation needs to be quantified for different genes and for different individuals or populations.
Calculating the allele frequency

\[ P(A) = \frac{[2(AA) + (Aa)]}{2n} \]

- Twice the number of homozygous genotypes with that allele (because homozygotes carry two copies each of the same allele),
- plus the number of heterozygous genotypes with that allele (because heterozygotes carry only one copy of a particular allele),
- divided by two times the total number of individuals in the sample (because each individual carries two alleles per locus)

Note that any result obtained with this formula will only be an estimate of the total allele frequency in the population, because only a sample of individuals is usually studied. However, if the sampling of individuals is well done, that is, the size of the sample is sufficiently large, then it can be assumed that our calculation is close to the true allele frequency. As a rule of thumb, allele frequency estimates should be performed, where possible, on samples of 100 individuals or more.
Genotype frequency

- This is the frequency of a given genotype in a population
- The frequencies of various types of breeding systems determine the mathematical relationship between the allele and genotype frequencies

The natural breeding system of individuals can be examined through studies of the frequencies with which alternative genotypes occur in a population. When a population undergoes mating that is random relative to the alleles of interest, certain patterns of genotype frequencies may be expected.

Genotype frequencies are also used to estimate the amount of self-pollination occurring in populations of individuals that have this or mixed type of reproduction. The mating or breeding system, therefore, has a significant effect on the frequency of occurrence of alternative genotypes in a given population.
The Hardy-Weinberg principle

- A population with random mating results in an equilibrium distribution of genotypes after only one generation, so that the genetic variation is maintained.

- When the assumptions are met, the frequency of a genotype is equal to the product of the allele frequencies: $A A = p^2$, $A a = 2pq$, $aa = q^2$.

The H-W equilibrium states that sexual reproduction does not reduce genetic variation generation after generation; on the contrary, the amount of variation remains constant if there are no disturbing forces acting against it. It establishes the relationship for calculating genotype frequencies under random mating and, in doing so, provides the foundation for many studies in population genetics.

This principle describes the expectations for allele frequencies in an idealized situation where,

- The organism is diploid
- Reproduction is sexual
- Generations are not overlapping
- Mating is random
- Population size is very large
- Migration is negligible
- Mutations can be ignored
- Natural selection does not affect the alleles under consideration

Note that most crop plants violate at least one of these assumptions!
Demonstrating the H-W principle

The starting point is generation 0. We have a gene with two alleles, A_1 and A_2. The frequency of allele A_1 is p and the frequency of allele A_2 is q. The genotype frequencies in generation 0 are for A_1 A_1 = p^2, for A_1 A_2 = 2pq and for A_2 A_2 = q^2. If random mating occurs, the probability of any allele from the female plant meeting any allele from the male plant will be the same. The table to the right of the slide depicts the four possible genotypes for next generation. The frequency of occurrence of each genotype is given by the product of the frequency of each allele in the genotype (e.g. for A_1 A_1 is p x p = p^2). If the results in the table are summarized, as in the blue inset at the bottom of the slide, we see that the genotype frequencies in generation 1 remain the same as in the previous generation.
An example of three populations in Hardy-Weinberg equilibrium

<table>
<thead>
<tr>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes of (G_0)</td>
</tr>
<tr>
<td>Popul.</td>
</tr>
<tr>
<td>Pop. 1</td>
</tr>
<tr>
<td>Pop. 2</td>
</tr>
<tr>
<td>Pop. 3</td>
</tr>
</tbody>
</table>

Population genotype frequencies are given in rows. Generations (\(G_0\) and \(G_1\)) are given in columns. Again, we have one gene with two alleles, \(A_1\) and \(A_2\). The frequency of allele \(A_1\) is \(p\) and the frequency of allele \(A_2\) is \(q\). Genotype frequencies are different for each population in generation 0 (e.g. the frequency of \(A_1 A_1\) in population 1 is 0.6; in population 2, 0.49; and in population 3, 0.4, and so on for the other genotypes). We note, though, that the allele frequencies in the three populations are similar in \(G_0\) (\(p = 0.7\) and \(q = 0.3\)). In the next generation, \(G_1\), if all requirements of the H-W principle are met, the genotype frequencies in the three populations balance (now the frequency of \(A_1 A_1\) is 0.49 in the three populations, and the same happens with the frequencies of \(A_1 A_2\) and \(A_2 A_2\)). The allele frequencies are kept.
The chi-square test

This hypothesis test is useful for determining whether the allelic frequencies are in H-W equilibrium.

The procedure is as follows:

- Define $H_0$ (and $H_a$)
- Define significance level $\alpha$
- Perform the statistical test
- Apply the decision criteria

$H_0$ = the hypothesis that says the allele frequencies for trait Q in a given population are in H-W equilibrium.

$H_a$ = an alternative hypothesis that says the allele frequencies for trait Q are not in H-W equilibrium.

We choose a significance level to give us a certain percentage of confidence in our results. The statistical test follows the formula:

$$ST = \frac{\sum [O_i - E_i]^2}{E_i} = \chi^2_{k - mdf}$$

Where,

- $ST$ = statistical test
- $k$ = number of genotypic classes
- $O_i$ = observed frequencies
- $E_i$ = expected frequencies
- $m$ = number of alleles
- $df$ = degrees of freedom

If our sample allows for only 1 degree of freedom, then the difference in frequencies is reduced by 0.5, a correction factor, such as:

$$ST = \left[\frac{\sum [O_i - E_i]^2}{E_i} - 0.5\right]$$

The decision criteria is applied as follows:

$$\left[\frac{[O_i - E_i] - 0.5}{E_i}\right]^2$$

If $\chi^2_{cal} \leq \chi^2_{tab}$ then $H_0$ is accepted; and, if $\chi^2_{cal} > \chi^2_{tab}$ then $H_0$ is rejected.

Where,

- $cal$ = the result of calculating $ST$ with the data obtained in our sample
- $tab$ = the value identified in the table (a $\chi^2$ table may be found in Appendix 1, click here).
Applying the $\chi^2$ test: An example

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. observed</td>
<td>169</td>
<td>520</td>
<td>311</td>
</tr>
<tr>
<td>No. expected</td>
<td>250</td>
<td>500</td>
<td>250</td>
</tr>
<tr>
<td>$\chi^2$ calculation</td>
<td>$(169-250-0.5)^2/250$</td>
<td>$(520-500-0.5)^2/500$</td>
<td>$(311-250-0.5)^2/250$</td>
</tr>
<tr>
<td>$\chi^2$ value</td>
<td>25.921</td>
<td>0.760</td>
<td>14.641</td>
</tr>
</tbody>
</table>

Decision criteria:

$$\chi^2_{\text{cal}} (41.322) > \chi^2_{\text{tab}} (3.8)$$

$H_0$ is rejected

In this example, let’s say that allele frequencies were 0.429 for $A_1$ and 0.571 for $A_2$. Each genotypic class was represented as in the table on the slide above.

The hypotheses being tested are:

$H_0 =$ this population is in H-W equilibrium for its allele frequencies

$H_a =$ this population is not in H-W equilibrium for its allele frequencies

Because the number of genotypic classes is 3 and we therefore have only 1 degree of freedom, we apply the correction factor in our calculation of the $\chi^2$ elements.

The $\chi^2$ calculated is 41.322. With this figure and with an error margin of 0.05%, the statistical evidence rejects $H_0$, which means that this population is not in H-W equilibrium for the trait under study.
Calculating allele frequencies: Examples

- Calculating allele frequencies with a codominant marker
- Calculating allele frequencies with a dominant marker
- Calculating the allele frequencies with a codominant gene having multiple alleles

In the next few slides we show examples of calculations of allele frequencies for results obtained with different marker types. The examples are given with figures that emulate real gels and the bands obtained through the application of molecular markers. For details on molecular marker technologies and the interpretation of bands, see the training module Using Molecular Marker Technology in Studies on Plant Genetic Diversity (http://www.ipgri.cgiar.org/publications/pubfile.asp?ID_PUB=912).
With a codominant marker, the genotypes of the three genotypic classes can be observed for the two homozygotes and the heterozygote. In the drawing above, top centre, we see a gel image with the banding pattern of a codominant marker for a single locus of a diploid organism. We need to score the bands in the gel and convert them to numbers. To do so, each of the band sizes (the band in the same row) is scored and transformed to a 1 if it is present or to a 0 if it is absent. We can do it by band, as in the table at the bottom left, or by genotype, as at the bottom right corner. In the table below we can see the calculations of the expected and observed genotype frequencies, as well as the allele frequencies. (M = size marker.)

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>A₁ A₁</th>
<th>A₁ A₂</th>
<th>A₂ A₂</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype frequency (expected)</td>
<td>p²</td>
<td>2pq</td>
<td>q²</td>
<td></td>
</tr>
<tr>
<td>Number of individuals</td>
<td>n₁₁ = 40</td>
<td>n₁₂ = 20</td>
<td>n₂₂ = 140</td>
<td>n = 200</td>
</tr>
<tr>
<td>Genotype frequency (observed)</td>
<td>P₁₁ = n₁₁/n = 0.20</td>
<td>P₁₂ = n₁₂/n = 0.10</td>
<td>P₂₂ = n₂₂/n = 0.70</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
p = \frac{2n_{11}}{2n} + \frac{n_{12}}{2n} = \frac{P_{11}}{1} + \frac{1}{2} \frac{P_{12}}{1} = 0.20 + \frac{1}{2} (0.10) = 0.25
\]

\[
q = \frac{2n_{22}}{2n} + \frac{n_{12}}{2n} = \frac{P_{22}}{1} + \frac{1}{2} \frac{P_{12}}{1} = 0.70 + \frac{1}{2} (0.10) = 0.75
\]
The example in this slide is similar to that in the previous slide, but with 10 individuals and three segregating loci (A, B and D). For ease of presentation, only one scoring method is used (bottom of slide). (M = size marker.)

Note that a gel, like the one in the example, can be obtained only by multiplexing, that is, loading different marker reactions into the same well.

Below, we see a table with the calculations of genotype and allele frequencies. (exp. = expected values; obs. = observed values.)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Data analysis</th>
<th>Allele freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotypes</td>
<td>A₁ A₁</td>
</tr>
<tr>
<td>A</td>
<td>Genotype freq. (exp.)</td>
<td>p²</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁₁ = 0.2</td>
</tr>
<tr>
<td>B</td>
<td>Genotypes</td>
<td>B₁ B₁</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (exp.)</td>
<td>p²</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁₁ = 0.8</td>
</tr>
<tr>
<td>D</td>
<td>Genotypes</td>
<td>D₁ D₁</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (exp.)</td>
<td>p²</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁₁ = 0.2</td>
</tr>
</tbody>
</table>
With a dominant marker, only two genotypic classes can be observed: AA + Aa and aa, that is, one of the homozygote classes is confounded with the heterozygote. The gel image with the banding pattern of a dominant marker for a single locus will show either one band or no band for each individual. The bands are scored in a way similar to that for the codominant marker, where bands are converted to a score of 1 if present or 0 if not. (M = size marker.)

The calculations of frequencies are performed as shown in the table below. (p, q = allele frequencies.)

<table>
<thead>
<tr>
<th>Phenotypes</th>
<th>A _</th>
<th>aa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes</td>
<td>AA</td>
<td>Aa</td>
<td>aa</td>
</tr>
<tr>
<td>Phenotype frequencies (expected)</td>
<td>p² + 2pq</td>
<td>q²</td>
<td>1</td>
</tr>
<tr>
<td>Number of individuals</td>
<td>n₁ = 84</td>
<td>n₂ = 16</td>
<td>n = 100</td>
</tr>
<tr>
<td>Phenotype frequencies (observed)</td>
<td>P₁ = n₁/n = 0.84</td>
<td>P₂ = n₂/n = 0.16</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ q = \sqrt{n₂/n} = \sqrt{P₂} = \sqrt{0.16} = 0.4 \]

This is a biased estimate because it does not take into account the recessive alleles in the heterozygotes.

\[ p = (1 - q) = 0.6 \]
Here we have an example similar to that in the previous slide but with 10 individuals and three segregating loci (A, B, and D). (M = size marker.)

Non-segregating bands (monomorphic) are not scored and thus are not included in the analysis.

Below, we see the table with calculations of genotype and allele frequencies:

<table>
<thead>
<tr>
<th>Locus</th>
<th>Data analysis</th>
<th>Allele freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotypes</td>
<td>A₁₁ A₂ A₂</td>
</tr>
<tr>
<td>A</td>
<td>Genotype freq. (exp.)</td>
<td>p² + 2pq</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁ = 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>Genotypes</td>
<td>B₁₁ B₂ B₂</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (exp.)</td>
<td>p² + 2pq</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁ = 0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>Genotypes</td>
<td>D₁₁ D₂ D₂</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (exp.)</td>
<td>p² + 2pq</td>
</tr>
<tr>
<td></td>
<td>Number of indivs.</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Genotype freq. (obs.)</td>
<td>P₁ = 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

We cannot distinguish heterozygotes, but we can estimate the expected number of heterozygotes in a population. For example, if sample size = 1000, then:
For locus A, no. expected heterozygotes = 2pqN = 2(0.55)(0.45)(1000) = 495
For locus B, no. expected heterozygotes = 2pqN = 2(0.16)(0.84)(1000) = 269
and so on ...
... with a codominant gene having multiple alleles

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>(A_1) (A_1)</th>
<th>(A_1) (A_2)</th>
<th>(A_1) (A_3)</th>
<th>(A_2) (A_2)</th>
<th>(A_2) (A_3)</th>
<th>...</th>
<th>(A_n) (A_n)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype frequencies (expected)</td>
<td>(p_{11}) ({}^2)</td>
<td>(2p_1p_2)</td>
<td>(2p_1p_3)</td>
<td>(p_{22}) ({}^2)</td>
<td>(2p_2p_3)</td>
<td>...</td>
<td>(p_{nn}) ({}^2)</td>
<td>1</td>
</tr>
<tr>
<td>Number of individuals</td>
<td>(n_{11})</td>
<td>(n_{12})</td>
<td>(n_{13})</td>
<td>(n_{22})</td>
<td>(n_{23})</td>
<td>...</td>
<td>(n_{nn})</td>
<td>(n)</td>
</tr>
<tr>
<td>Genotype frequencies (observed)</td>
<td>(P_{11}) = (n_{11}/n)</td>
<td>(P_{12}) = (n_{12}/n)</td>
<td>(P_{13}) = (n_{13}/n)</td>
<td>(P_{22}) = (n_{22}/n)</td>
<td>(P_{23}) = (n_{23}/n)</td>
<td>...</td>
<td>(P_{nn}) = (n_{nn}/n)</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
p_1 = P_{11} + \frac{1}{2} \sum_{j=1}^{n} P_{1j}
\]
\[
p_2 = P_{22} + \frac{1}{2} \sum_{j=2}^{n} P_{2j}
\]
\[
p_3 = P_{33} + \frac{1}{2} \sum_{j=3}^{n} P_{3j}
\]
\[
p_4 = P_{44} + \frac{1}{2} \sum_{j=4}^{n} P_{4j}
\]
\[
p_n = P_{nn} + \frac{1}{2} \sum_{j=n}^{n} P_{nj}
\]

(continued on next slide)

This is the situation encountered when markers such as microsatellites are used.

We have a locus \(A\) with \(n\) alleles \(A_1, A_2, A_3, \ldots, A_n\) and allele frequencies \(p_1, p_2, p_3, \ldots, p_n\), respectively, such that \(A_1 = A_2 = A_3 = \ldots = A_n\)
Again, with a codominant marker, the genotypes of the three genotypic classes can be observed. In the drawing above, top centre, we see a gel image with the banding pattern of a codominant marker with three alleles (A₁, A₂ and A₃) in a diploid sample. We score each band (each row) independently, and transform them to a score of 1 if present or a score of 0 if not. We can do it by band (bottom left of slide) or by genotype (bottom right corner). In the table below, we can see the calculations of the expected and observed genotype frequencies, as well as the allele frequencies (p₁, p₂ and p₃). (M = size marker.)

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>A₁ A₁</th>
<th>A₁ A₂</th>
<th>A₁ A₃</th>
<th>A₂ A₂</th>
<th>A₂ A₃</th>
<th>A₃ A₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype frequency (exp.)</td>
<td>p₁²</td>
<td>2p₁p₂</td>
<td>2p₁p₃</td>
<td>p₂²</td>
<td>2p₂p₃</td>
<td>p₃²</td>
<td>1</td>
</tr>
<tr>
<td>Number of individuals</td>
<td>n₁₁ = 4</td>
<td>n₁₂ = 6</td>
<td>n₁₃ = 0</td>
<td>n₂₂ = 10</td>
<td>n₂₃ = 2</td>
<td>n₃₃ = 2</td>
<td>n = 24</td>
</tr>
<tr>
<td>Genotype frequency (obs.)</td>
<td>P₁₁ = n₁₁/n = 0.17</td>
<td>P₁₂ = n₁₂/n = 0.25</td>
<td>P₁₃ = n₁₃/n = 0</td>
<td>P₂₂ = n₂₂/n = 0.42</td>
<td>P₂₃ = n₂₃/n = 0.08</td>
<td>P₃₃ = n₃₃/n = 0.08</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
p₁ = P_{11} + \frac{1}{2}P_{12} + \frac{1}{2}P_{13} = P_{11} + \frac{1}{2}\sum_{j \neq 1} P_{1j} = 0.17 + \frac{1}{2}(0.25 + 0.00) = 0.30
\]

\[
p₂ = P_{22} + \frac{1}{2}P_{21} + \frac{1}{2}P_{23} = P_{22} + \frac{1}{2}\sum_{j \neq 2} P_{2j} = 0.42 + \frac{1}{2}(0.25 + 0.08) = 0.59
\]

\[
p₃ = P_{33} + \frac{1}{2}P_{31} + \frac{1}{2}P_{32} = P_{33} + \frac{1}{2}\sum_{j \neq 3} P_{3j} = 0.08 + \frac{1}{2}(0.00 + 0.08) = 0.12
\]
Reproduction and mating systems

Outbreeding, inbreeding or asexual reproduction

They influence:

• The degree of genetic relatedness among mates
• The organization of genes into genotypes

In principle, outbreeding occurs as random mating, and inbreeding and asexual reproduction both involve nonrandom mating.

Outcrossing species, compared with inbreeders, may retain large numbers of deleterious recessives because the dominance situation hides them. The recessives undergo frequent recombination, resulting in new gametic types.

Dominance refers to situations where, under heterozygous conditions, one allele has a phenotypic effect that is strong enough to conceal the presence of the other (recessive) allele. In a situation of dominance, only two phenotypes can be observed: the dominant phenotype that is a mixture of the homozygous dominant and the heterozygous, and the recessive phenotype.

In cross-pollinating species, selfing leads to inbreeding depression because it increases the proportion of homozygotes, thus permitting rare recessive alleles to become visible. Heterozygotes in cross-pollinating species have a more favourable effect.

In organisms with asexual reproduction, a variety of consequences result, depending on its type. Although asexual reproduction may be a constant mode of reproduction, it may be combined with cycles of sexual reproduction, which allow recombination of the current variation and, as such, the generation of new forms or combinations. If only asexual reproduction occurs in the population, genotypic frequencies cannot be changed.
Random mating

- Mating that takes place at random, that is, the chances of individual A mating with individual B do not depend on the genotypes of either

- If random mating occurs, the chance that an individual mates with a given genotype is equal to the frequency of that genotype in the population

Random mating is typically found in many outbreeding populations. For example, we may have a population in which genotype AA is present 10% of times, Aa 58% of times and aa 32% of times. If mating is random, then the chances of an individual AA mating with another AA is 10/100, Aa 58/100, or aa 32/100.
Nonrandom mating occurs when individuals that are more closely (inbreeding) or less closely related mate more often than would be expected by chance for the population.

Self-pollination or inbreeding is similar to mating between relatives. It increases the homozygosity of a population and its effect is generalized for all alleles. Inbreeding per se does not change the allelic frequencies but, over time, it leads to homozygosity by slowly increasing the two homozygous classes.
**Inbreeding coefficient**

- It compares the actual proportion of heterozygous genotypes with those expected under random mating

\[ F = \frac{(H_0 - H)}{H_0} \]

- \( F \) is the inbreeding coefficient and measures the reduction of heterozygosity

\( H = \) actual frequency of heterozygotes in the population.
\( H_0 = \) expected number of heterozygotes under random mating.

The inbreeding coefficient indicates the degree of inbreeding in a population.
Selfing is a powerful inbreeding system that allows the attainment of high levels of homozygosity within few generations. Simultaneously, heterozygosity decreases. The chart on the slide shows the phenomenon and, in the table below, we can see the changing values of homozygosity and heterozygosity in 9 generations (G₀ to G₈).

<table>
<thead>
<tr>
<th>Generation</th>
<th>Selfed genotypes ratio/generation</th>
<th>Homozygosity (%)</th>
<th>Heterozygosity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G₀</td>
<td>Aa</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>G₁</td>
<td>1AA, 2Aa, 1aa</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>G₂</td>
<td>6AA, 4Aa, 6aa</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>G₃</td>
<td>28AA, 8Aa, 28aa</td>
<td>87.5</td>
<td>12.5</td>
</tr>
<tr>
<td>G₄</td>
<td>120AA, 16Aa, 120aa</td>
<td>93.75</td>
<td>6.25</td>
</tr>
<tr>
<td>G₅</td>
<td>496AA, 32Aa, 496aa</td>
<td>96.875</td>
<td>3.125</td>
</tr>
<tr>
<td>G₆</td>
<td>2016AA, 64Aa, 2016aa</td>
<td>98.4375</td>
<td>1.5625</td>
</tr>
<tr>
<td>G₇</td>
<td>8128AA, 128Aa, 8128aa</td>
<td>99.21875</td>
<td>0.78125</td>
</tr>
<tr>
<td>G₈</td>
<td>32640AA, 256Aa, 32640aa</td>
<td>99.60938</td>
<td>0.390625</td>
</tr>
</tbody>
</table>
Forces shaping genetic diversity

- Mutation
- Migration
- Recombination
- Selection
- Drift

If, for any reason, a population becomes homogeneous, evolution would not occur. Thus, constant change essentially depends on new variation.

A genetic population is the aggregate of allelic frequencies of all genes in that population. Populations change or evolve because their gene frequencies experience change. Several factors can produce changes in fitness—the capability of an individual to survive until reproduction is achieved. If the fitness of an individual in a population changes, the genotypes in the subsequent generation will not be directly related to the gene frequencies of the former population, thus driving the population to evolve.

Because changes in populations require changes in gene frequencies, we must understand how these frequencies can change. In the next slides, we discuss the primary causes of change: mutation, migration, recombination, selection and drift.
Mutation

- It is the ultimate source of variation and may be caused by:
  - Errors in DNA replication
  - Damage by radiation

- Mutation increases diversity but, because spontaneous mutations are rare, the rate of change in gene frequency is very low

- Consequently, mutation alone does not drive the evolution of populations and species

The simplest mutation is that which produces a change in a single nucleotide in the DNA sequence of a gene. A mutation can cause one allele to change to another allele already present in the population (from a dominant to a recessive) or it could create a completely new allele. Mutations may be adverse or favourable. Many will be adverse and disappear. But, if they are good for the individual, then frequencies of that allele will increase from generation to generation. Moreover, this mutation may migrate to other populations and spread.

*Note:* Genomes may go through a process known as gene duplication. This event helps the individual to endure an unfavourable mutation in a copy of the gene without major difficulties, because the other copy of the gene still may function appropriately. More changes may affect the mutated gene and determine different types of adaptation for the individual.
Migration

- It is the movement of individuals or any form of introduction of genes from one population to another.
- Migration increases diversity and its rate can be large, causing significant changes in frequency.
- The change in gene frequency is proportional to the difference in frequency between the recipient population and the average of the donor populations.

From a genetic perspective, migration implies not only the movement of individuals into new populations but that this movement introduces new alleles into the population (gene flow). Changes in gene frequencies will occur through migration either because more copies of an allele already present will be brought in or because a new allele arrives.

Several factors affect migration in crop species:

- Breeding system
- Sympathy with wild and/or weedy relatives
- Pollinators
- Seed dispersal

The immediate effect of migration is to increase a population’s genetic variability and, as such, helps increase the possibilities of that population to withstand environmental changes. Migration also helps blend populations and prevent their divergence.
Recombination

- It is the process whereby a cell generates new chromosomal combinations, compared with that cell or with those of its progenitors.
- It does not create new diversity but generates new combinations of the existing diversity.
- Genetic variance through recombination, given that segregating alleles exist at different loci, can occur very quickly.

Genetic diversity through recombination is the result of reshuffling the genetic components of the original type. Mechanisms exist to generate allelic diversity (intragenic recombination) and genomic diversity (new multigene combinations).
Selection

It is the inherited ability of organisms to survive and reproduce. It acts in such a manner that, with time, superior genotypes increase their frequency in the population.

As a result of mutation, new forms develop. These forms, as explained, may be favourable or deleterious to the individual’s ability to survive. If changes are advantageous, then the new alleles will tend to prevail by being selected in the population.

The effect of selection on diversity may be:

- Directional, where it decreases diversity
- Balancing, where it increases diversity. Heterozygotes have the highest fitness, so selection favours the maintenance of multiple alleles
- Frequency dependent, where it increases diversity. Fitness is a function of allele or genotype frequency and changes over time
Genetic drift refers to fluctuations in allele frequencies that occur by chance (particularly in small populations) as a result of random sampling among gametes.

Drift decreases diversity within a population because it tends to cause the loss of rare alleles, reducing the overall number of alleles.

In the example above, population size is constant in each generation (8 individuals). Each individual may produce thousands of gametes but, from the total genepool in each generation, only 2N gametes per individual are needed (16 in our example). This situation is similar to extracting small samples from two boxes, one containing a million white balls and the other a million red balls. In each extracting experiment, we may take a different number of white and red balls. We simulate that, in generation G₀, 10 gametes of the thousands possible carried allele A and only 6 carried allele a. In G₁, from those gametes that participated in the constitution of the zygotes for next generation, 5 carried allele A and 11 allele a, and so on. These values vary randomly. In G₃ all individuals are formed by allele a (homozygotes) and allele A is lost.
Effective population size

- $N_e$ is the number of parents responsible for the genetic composition of the next generation

- $N_e$ is generally less than $N$ because of:
  - Variation of population size from generation to generation
  - Unequal sex ratio
  - Overlapping generations
  - Geographic dispersion of populations

How big is the population?

- The actual number of individuals in a population is called the census number ($N$). This number is often an imprecise representation of the population size from a genetic standpoint.

- The effective population size ($N_e$) describes the size of an ideal population that shows the same rate of loss of genetic variation due to genetic drift as for the population of interest.
Consequences of a decreasing population size

- Genetic drift, with random variation of allele frequencies
- Inbreeding
- Homozygosity: fixation and loss of alleles
- Subpopulation differentiation

Events that reduce population size:

- Domestication
- Subpopulations (inbreeding, clonal reproduction)
- Long-range dispersal (founder effect)
- Regeneration of genetic resources collections
A bottleneck develops when the population size drops sharply.

The founder effect occurs when a few individuals colonize and become established in a new environment.

Small populations are highly vulnerable to extinction because the surviving sample may not be representative of the pre-crash genepool.

Both effects depend on the number of survivors (or colonizers) and the rate of population growth.

The graphic in the slide shows a bottleneck effect. To the left is a population in Hardy-Weinberg equilibrium with allele frequencies of 0.5. If a sudden reduction occurs, then the original size restored, the result may be a loss of alleles and the fixation of others. In the example (right), only $A_2 A_2$ survived and allele $A_1$ was lost.
With bottleneck and founder effects ...

- Heterozygosity declines at the rate of:
  \[ H_1 = (1 - 1/2N)H_0 \]

- Alleles are lost at the rate of:
  \[ P = p^{2N} + q^{2N} \]

• Heterozygosity decline

  Where,
  
  $H_1$ = final heterozygosity
  $H_0$ = initial heterozygosity
  $N$ = population size

  Then:
  
  If $N = 100$ and $H_0 = 0.25$, then $H_1 = 0.24875$
  If $N = 40$ and $H_0 = 0.25$, then $H_1 = 0.24685$

• Allele loss

  Where,
  
  $P$ = allele loss
  $p$ and $q$ = allele frequencies
  $2N$ = total number of alleles in the population

  Then:
  
  If $N = 100$, $p = 0.90$ and $q = 0.1$, then $P = 7.05508 \times 10^{-10}$
  If $N = 15$, $p = 0.90$ and $q = 0.1$, then $P = 0.0423911$
  If $N = 10$, $p = 0.90$ and $q = 0.1$, then $P = 0.12157665$
Appendix 1

Appendix 1. Critical values of the chi-square distribution
In summary

To analyse and interpret genetic diversity data we must be acquainted with:

- A few basic definitions of population genetics
- The Hardy-Weinberg principle
- Calculating allele and genotype frequencies
- The primary causes of changes in genetic diversity: mutation, migration, recombination, selection and drift
By now you should know …

- The basic definitions used in population genetics
- The Hardy-Weinberg principle
- How to calculate allele frequencies with marker data
- The mating system’s effect on a population’s diversity
- The main sources of variation and their consequences


- Measures of genetic diversity
- Software programs for analysing genetic diversity
- Glossary
Appendix 1 to:

*Basic Concepts of Population Genetics*

Critical values of the chi-square distribution

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.84</td>
<td>6.64</td>
</tr>
<tr>
<td>2</td>
<td>5.99</td>
<td>9.21</td>
</tr>
<tr>
<td>3</td>
<td>7.81</td>
<td>11.34</td>
</tr>
<tr>
<td>4</td>
<td>9.49</td>
<td>13.28</td>
</tr>
<tr>
<td>5</td>
<td>11.07</td>
<td>15.09</td>
</tr>
<tr>
<td>6</td>
<td>12.59</td>
<td>16.81</td>
</tr>
<tr>
<td>7</td>
<td>14.07</td>
<td>18.48</td>
</tr>
<tr>
<td>8</td>
<td>15.51</td>
<td>20.09</td>
</tr>
<tr>
<td>9</td>
<td>16.92</td>
<td>21.67</td>
</tr>
<tr>
<td>10</td>
<td>18.31</td>
<td>23.21</td>
</tr>
<tr>
<td>11</td>
<td>19.68</td>
<td>24.72</td>
</tr>
<tr>
<td>12</td>
<td>21.03</td>
<td>26.22</td>
</tr>
<tr>
<td>13</td>
<td>22.36</td>
<td>27.69</td>
</tr>
<tr>
<td>14</td>
<td>23.68</td>
<td>29.14</td>
</tr>
<tr>
<td>15</td>
<td>25.00</td>
<td>30.58</td>
</tr>
<tr>
<td>16</td>
<td>26.30</td>
<td>32.00</td>
</tr>
<tr>
<td>17</td>
<td>27.59</td>
<td>33.41</td>
</tr>
<tr>
<td>18</td>
<td>28.87</td>
<td>34.80</td>
</tr>
<tr>
<td>19</td>
<td>30.14</td>
<td>36.19</td>
</tr>
<tr>
<td>20</td>
<td>31.41</td>
<td>37.57</td>
</tr>
<tr>
<td>21</td>
<td>32.67</td>
<td>38.93</td>
</tr>
<tr>
<td>22</td>
<td>33.92</td>
<td>40.29</td>
</tr>
<tr>
<td>23</td>
<td>35.17</td>
<td>41.64</td>
</tr>
<tr>
<td>24</td>
<td>36.41</td>
<td>42.98</td>
</tr>
<tr>
<td>25</td>
<td>37.65</td>
<td>44.31</td>
</tr>
<tr>
<td>26</td>
<td>38.88</td>
<td>45.64</td>
</tr>
<tr>
<td>27</td>
<td>40.11</td>
<td>46.96</td>
</tr>
<tr>
<td>28</td>
<td>41.34</td>
<td>48.28</td>
</tr>
<tr>
<td>29</td>
<td>42.56</td>
<td>49.59</td>
</tr>
<tr>
<td>30</td>
<td>43.77</td>
<td>50.89</td>
</tr>
</tbody>
</table>
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
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, Ind₁, Ind₂, … 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 categorical 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

\[ P_j = q \leq 0.95 \quad \text{or} \quad P_j = q \leq 0.99 \]

Where,

\[ P_j = \text{rate of polymorphism} \]
\[ q = \text{allele frequency} \]

- 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

Proportion of polymorphic loci

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

\[ P = \frac{n_{pj}}{n_{total}} \]

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

\[ n = \left( \frac{1}{K} \right) \sum_{i=1}^{K} n_i \]

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

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

Where,

- $p_i =$ frequency of the $i^{th}$ allele in a locus
- $h = 1 - \sum p_i^2 =$ 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.
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: $h_j = 1 - \sum p_i^2$
  - Average for several loci: $H = \Sigma h_j / L$

- 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 subtracting 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

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)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Data analysis</th>
<th>Allele freq.</th>
<th>$h_i = (1 - p^2 - q^2)$</th>
<th>$H_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>Genotypes</td>
<td>$A_1$, $A_2$</td>
<td>$A_1$, $A_2$</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>$p^2$</td>
<td>$2pq$</td>
<td>$q^2$</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>2</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_{11} = 0.07$</td>
<td>$P_{12} = 0.13$</td>
<td>$P_{22} = 0.80$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.13</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>Genotypes</td>
<td>$B_1$, $B_1$</td>
<td>$B_1$, $B_2$</td>
<td>$B_2$, $B_2$</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>$p^2$</td>
<td>$2pq$</td>
<td>$q^2$</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>7</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_{11} = 0.23$</td>
<td>$P_{12} = 0.10$</td>
<td>$P_{22} = 0.67$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td>Genotypes</td>
<td>$E_1$, $E_1$</td>
<td>$E_1$, $E_2$</td>
<td>$E_2$, $E_2$</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>$p^2$</td>
<td>$2pq$</td>
<td>$q^2$</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>15</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_{11} = 0.50$</td>
<td>$P_{12} = 0.27$</td>
<td>$P_{22} = 0.23$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.63</td>
</tr>
</tbody>
</table>

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:

   $$ H_o = \frac{16(0/5) + 13(1/5) + 1(2/5))}{30} = 0.1 $$

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:

   $$ H_i = \frac{(0.23 + 0.41 + 0.46)/5} = 0.22 $$
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)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Data analysis</th>
<th>Allele freq.</th>
<th>$h_i = (1 - p^2 - q^2)$</th>
<th>$H_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotypes</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Aa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>p^2</td>
<td>2pq</td>
<td>q^2</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>6</td>
<td>24</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_1 = 0.20$</td>
<td>$P_2 = 0.80$</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>BB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BB</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>p^2</td>
<td>2pq</td>
<td>q^2</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_1 = 0.33$</td>
<td>$P_2 = 0.67$</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>EE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>p</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (exp.)</td>
<td>p^2</td>
<td>2pq</td>
<td>q^2</td>
</tr>
<tr>
<td></td>
<td>Individuals (no.)</td>
<td>23</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Gen. freq. (obs.)</td>
<td>$P_1 = 0.77$</td>
<td>$P_2 = 0.23$</td>
<td>1</td>
</tr>
</tbody>
</table>

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 ($h_i$) 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})

\[ g_{ST} = 1 - \left( \frac{h_S}{h_T} \right) \]

\( h_S \) = population diversity
\( h_T \) = total diversity

Where,

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

$$G_{ST} = \frac{D_{ST}}{H_T}$$

Where,

- $H_T$ = total genic diversity = $H_S + D_{ST}$
- $H_S$ = intrapopulation genic diversity
- $D_{ST}$ = interpopulation diversity

$$\left(\frac{H_T}{H_T}\right) = \left(\frac{H_S}{H_T}\right) + \left(\frac{D_{ST}}{H_T}\right) = 1$$

- $G_{ST}$ 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, $G_{ST} = \left(\frac{D_{ST}}{H_T}\right) \times 100 = \left(\frac{0.061}{0.263}\right) \times 100 = 23.19\%$. 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.
The equation for the genetic structure of populations is:

\[(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})\]

\[
F_{IT} = 1 - \left(\frac{H_I}{H_T}\right)
\]

\[
F_{IS} = 1 - \left(\frac{H_I}{H_S}\right)
\]

\[
F_{ST} = 1 - \left(\frac{H_S}{H_T}\right)
\]

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:

0 (no genetic divergence) to 1 (fixation for alternate alleles in different subpopulations)

When $F_{ST}$ 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

<table>
<thead>
<tr>
<th>Pop.</th>
<th>Genotype frequency</th>
<th>( p_i )</th>
<th>( q_i )</th>
<th>2pq_i</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( A_1 A_1 )</td>
<td>0.40</td>
<td>0.55</td>
<td>0.4950</td>
<td>0.3939</td>
</tr>
<tr>
<td></td>
<td>( A_1 A_2 )</td>
<td>0.30</td>
<td>0.45</td>
<td>0.4200</td>
<td>0.5238</td>
</tr>
<tr>
<td></td>
<td>( A_2 A_2 )</td>
<td>0.30</td>
<td>0.375</td>
<td>0.4688</td>
<td></td>
</tr>
</tbody>
</table>

\[ H_T = 2(0.625)(0.375) = 0.4688 \]
\[ p_o = (0.55 + 0.70)/2 = 0.625 \]
\[ H_I = (0.3 + 0.2)/2 = 0.25 \]
\[ q_o = (0.45 + 0.30)/2 = 0.375 \]
\[ H_S = (0.495 + 0.420)/2 = 0.4575 \]

\[ F_{IT} = 1 - (0.25/0.4688) = 0.4667 \]
\[ F_{IS} = 1 - (0.25/0.4575) = 0.4536 \]
\[ F_{ST} = 1 - (0.4575/0.4688) = 0.0241 \]

(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)

<table>
<thead>
<tr>
<th>Pop.</th>
<th>Genotype frequency</th>
<th>p_i</th>
<th>q_i</th>
<th>2p_iq_i</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A_1 A_1 0.25 0.50 0.25</td>
<td>0.50</td>
<td>0.50</td>
<td>0.500</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>A_1 A_2 0.80 0.10 0.10</td>
<td>0.85</td>
<td>0.15</td>
<td>0.255</td>
<td>0.6078</td>
</tr>
<tr>
<td></td>
<td>A_2 A_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H_T(0.675)(0.325) = 0.4388</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H_I(0.5 + 0.1)/2 = 0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H_S(0.500 + 0.255)/2 = 0.3775</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ F_{IT} = 1 - \left(\frac{0.30}{0.4388}\right) = 0.3163 \]
\[ F_{IS} = 1 - \left(\frac{0.30}{0.3775}\right) = 0.2053 \]
\[ F_{ST} = 1 - \left(\frac{0.3775}{0.4388}\right) = 0.1397 \]

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 \((F_{IS} = 0.2053)\).
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

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 ($X_{ijk}$). We have $i = 15$ individuals (effect b), $j = 2$ alleles (effect w), $k = 3$ populations (effect a).

Where,

- $X_{i,k}$ is the result of summing up all the band presences (1) in the individuals per population
- $X_{i,k}^2$ 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:

$$SS_a = \sum X_{i,k}^2/ij - X_{i,k}^2/ijk = [990/(15 \times 2)] - [2916/(15 \times 2 \times 3)] = 0.6$$

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).

<table>
<thead>
<tr>
<th>Ind.</th>
<th>Pop. 1</th>
<th>Pop. 2</th>
<th>Pop. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_1$</td>
<td>$A_2$</td>
<td>$A_1$</td>
</tr>
<tr>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>$A_2$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

SS | 0.6 | MS | 0.3 |
SS | 11  | MS | 0.26190476 |
SS | 10  | MS | 0.22222222 |

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

(continued on next slide)
An example of AMOVA (continued)

<table>
<thead>
<tr>
<th>SV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Populations</td>
<td>2</td>
<td>0.6</td>
<td>0.3</td>
<td>(\sigma_w^2 + 2\sigma_b^2 + 2*15\sigma_a^2)</td>
</tr>
<tr>
<td>Indiv./pop.</td>
<td>42</td>
<td>11</td>
<td>0.26190476</td>
<td>(\sigma_w^2 + 2\sigma_b^2)</td>
</tr>
<tr>
<td>Within indiv.</td>
<td>45</td>
<td>10</td>
<td>0.22222222</td>
<td>(\sigma_w^2)</td>
</tr>
</tbody>
</table>

Estimates of variances and F statistics

| \(\sigma_a^2\) | 0.0012698 |
| \(\sigma_b^2\) | 0.0198413 |
| \(\sigma_w^2\) | 0.2222222 |
| \(\sigma^2\) | 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)
\(\sigma^2\) = total estimated variance
EMS = expected mean squares
\(\sigma_w^2\) = 0.22222222
\(\sigma_b^2\) = \((\text{MS}_b - \text{MS}_w)/2 = (0.26190476 – 0.22222222)/2 = 0.0198413\)
\(\sigma_a^2\) = \((\text{MS}_a - \text{MS}_b)/2 * 15 = (0.3 – 0.26190476)/2 * 15 = 0.0012698\)
\(\sigma^2\) = \(\sigma_w^2 + \sigma_b^2 + \sigma_a^2\) = 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:

\(F_{IT} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2} = \frac{(0.0012698 + 0.0198413)/0.24333 = 0.086758}{0.24333}\)
\(F_{ST} = \frac{\sigma_b^2}{\sigma^2} = \frac{0.0012698/0.24333 = 0.0052185}{0.24333}\)
\(F_{IS} = \frac{\sigma_b^2}{\sigma_a^2 + \sigma_b^2} = \frac{0.0198413/(0.0198413 + 0.2222222) = 0.0819672}{0.24333}\)

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.
It measures the nucleotide diversity among several sequences in a given region of the genome within a population ($\pi_X$)

$$\pi_X = n/(n - 1)\sum X_iX_j\pi_{ij}$$

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
- $\pi_{ij}$ = 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

<table>
<thead>
<tr>
<th>n</th>
<th>Sequence</th>
<th>Freq. $X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Seq₁ TCC T CGAT T ATTC C CAGGGTGC C GATG A AT</td>
<td>5/10 = 0.5</td>
</tr>
<tr>
<td>2</td>
<td>Seq₂ TCC A CGAT T ATTC G CAGGGTGC C GATG A AT</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>1</td>
<td>Seq₃ TCC A CGAT C ATTC C CAGGGTGC A GATG G AT</td>
<td>1/10 = 0.1</td>
</tr>
<tr>
<td>2</td>
<td>Seq₄ TCC G CGAT T ATTC T CAGGGTGC G GATG A AT</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\Pi_{1,2} = 2/30$, $\Pi_{1,3} = 4/30$, $\Pi_{1,4} = 3/30$, $\Pi_{2,3} = 4/30$, $\Pi_{2,4} = 3/30$, $\Pi_{3,4} = 5/30$

\[ \pi_x = \frac{10}{(10 - 1)} \sum X_i \pi_{ij} \]
\[ = \frac{(10/9)(0.5 \times 0.2 \times (2/30) + 0.5 \times 0.1 \times (4/30) + \ldots + 1 \times 0.2 \times (5/30))}{10} \]
\[ = 0.037 \]

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, $\Pi_{1,2} = 2/30$ 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 $\pi_x$ 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

  \[ V_W = (1/s)\sum \pi_X \]

- $V_b$ measures the total differentiation in several populations

  \[ V_b = [1/(s(s - 1))]\sum_X \sum_Y V_{XY} \]

- $N_{ST}$ is the relative differentiation

  \[ N_{ST} = V_b/(V_b + V_W) \]

Where,

- $V_{XY}$ = divergence among populations X and Y
- $\pi_X$ = 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*.
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 (\( N_{ST} \)).
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 ($\pi$) is based on the number of restriction fragments present in two samples

$$\pi = - \frac{1}{r} \ln G$$

(if $\pi < 5\%$)

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:

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

$$F = \frac{\sum X_i(X_i - 1)}{\sum X_i(n - 1)}$$

- $F$ = proportion of shared fragments
- $G^2 = F^{1/4}$
- $n$ = 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

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 $Y$

$\pi_X$ = restriction diversity in population $X$

$d_{XY}$ = fragment diversity among two populations = $- (2/r) \ln (G_{XY})$

$G_{XY} = F_{XY}(3 - 2G^0_{XY})^{1/4}$

$G^0 = F_{XY}^{1/4}$

$F_{XY}$ = proportion of shared alleles among populations $X$ and $Y$

$$= (2\Sigma X_i X_j)/(\Sigma(X_i + X_j))$$

$X_i$ = 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

| Seq. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | Freq. Xi |
|------|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-------|
| A_X |   |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    | 6/20 = 0.30 |
| A_Y |   |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    | 5/20 = 0.25 |
| A_Z |   |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    | 9/20 = 0.45 |

\( F = \frac{[0.30(3-1) + 0.25(3-1) + 0.45(3-1)]}{0.30(3-1) + 0.25(3-1) + 0.45(3-1)} = 0.0325 \)

\( G^0_{XY} = (0.0325)^{1/4} = 0.701743 \)

\( G_{XY} = 0.0325[3 - 2(0.039358)]^{1/4} = 0.272587 \)

\( d_{XY} = -(2/6)\ln(0.272587) = 0.216633 \)

\( V_{XY} = 0.604643 - \frac{1}{2}(0.539176 + 0.216633) = 0.226739 \)

\( V_W = \frac{1}{2}(0.539176 + 0.216633) = 0.377905 \)

\( V_b = \frac{1}{2}(0.226739) = 0.11337 \)

\( N_S T = \frac{0.11337}{0.11337 + 0.377905} = 0.230766 \)

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.
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)

\[ d_{t+1} = d_t \]

**Disequilibrium model**
- Distance changes with time through migration and genetic drift

\[ d_{t+1} = d_t + \delta \]

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

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.
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 AAAAA, AAAa, AAa or Aaaa; however the band will only be scored as present (1) the same as it will in the diploid (AA or Aa).

<table>
<thead>
<tr>
<th>Locus</th>
<th>Genotypes</th>
<th>Allele freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p</td>
</tr>
<tr>
<td>A (2X) Diploid</td>
<td>AA, Aa</td>
<td>aa</td>
</tr>
<tr>
<td>Gen. freq. (e.)</td>
<td>p² + 2pq</td>
<td>q²</td>
</tr>
<tr>
<td>Indiv. number</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Gen. freq. (o.)</td>
<td>P₁ = 0.78</td>
<td>P₂ = 0.22</td>
</tr>
<tr>
<td>A (4X) Tetraploid</td>
<td>AAAAA, AAAa, AAa, Aaaa</td>
<td>aaaa</td>
</tr>
<tr>
<td>Gen. freq. (e.)</td>
<td>p⁴ + 4p³q + 6p²q² + 4pq³</td>
<td>q⁴</td>
</tr>
<tr>
<td>Indiv. number</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Gen. freq. (o.)</td>
<td>P₁ = 0.78</td>
<td>P₂ = 0.22</td>
</tr>
</tbody>
</table>

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, AAa, 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.)
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 (A₁, A₂ and A₃).

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 A₁ A₁ A₂ A₃ cannot be distinguished from those with other combinations such as A₁ A₂ A₂ A₃ or A₁ A₂ A₃ A₃. This situation can only be solved by inference based on estimating the DNA fragment copy number in the gel.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>A₁ A₁</th>
<th>A₁ A₂</th>
<th>A₁ A₃</th>
<th>A₂ A₂</th>
<th>A₂ A₃</th>
<th>A₃ A₃</th>
<th>Total</th>
<th>p</th>
<th>q</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen. freq. (e.)</td>
<td>p²</td>
<td>2pq</td>
<td>2pr</td>
<td>q²</td>
<td>2qr</td>
<td>r²</td>
<td>1</td>
<td>0.25</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td>Indivs. (no.)</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gen. freq. (o.)</td>
<td>P₁₁ = 0.11</td>
<td>P₁₂ = 0.06</td>
<td>P₁₃ = 0.06</td>
<td>P₂₂ = 0.06</td>
<td>P₂₃ = 0.11</td>
<td>P₃₃ = 0.44</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Similarity coefficients for binary variables: examples

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

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:

\[ \frac{a + d}{a + b + c + d} \]

**Jaccard Coefficient**:

\[ \frac{a}{a + b + c} \]

**Nei-Li Coefficient**, or Dice:

\[ \frac{2a}{2a + b + c} \]

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 $\mu$ 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 \(I_{XY}\):

\[ I_{XY} = \frac{J_{xy}}{\sqrt{(J_xJ_y)}} \]

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 \((\infty\), 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.

(continued on next slide)
Calculating Nei’s genetic distance (continued)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>Allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pop.1</td>
</tr>
<tr>
<td>A</td>
<td>A_1</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>A_2</td>
<td>0.20</td>
</tr>
<tr>
<td>Locus heterozygosity</td>
<td>h_{11}</td>
<td>0.3200</td>
</tr>
<tr>
<td>B</td>
<td>B_1</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>B_2</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>B_3</td>
<td>0.13</td>
</tr>
<tr>
<td>Locus heterozygosity</td>
<td>h_{11}</td>
<td>0.2434</td>
</tr>
<tr>
<td>D</td>
<td>D_1</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>D_2</td>
<td>1.00</td>
</tr>
<tr>
<td>Locus heterozygosity</td>
<td>h_{11}</td>
<td>0.0000</td>
</tr>
<tr>
<td>Average heterozygosity</td>
<td>J_i</td>
<td>0.0433</td>
</tr>
<tr>
<td>Average homozygosity</td>
<td>J_i</td>
<td>0.9567</td>
</tr>
<tr>
<td>Average interpop. homozygosity</td>
<td>J_{12}</td>
<td>J_{12} = 0.8733</td>
</tr>
<tr>
<td>Genetic identity</td>
<td>I_i</td>
<td>I_{12} = 0.9183</td>
</tr>
<tr>
<td>Genetic distance</td>
<td>D_i</td>
<td>D_{12} = 0.0852</td>
</tr>
</tbody>
</table>

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:

\[ j_{i|jk} = \sum p_{i|k} p_{i|j} \] for example, \( j_{1,2|k} \) = homozygosity among populations 1 and 2

\[ j_{1,2|k} = (0.8)(0.74) + (0.2)(0.26) + (0.86)(0.81) + (0.01)(0.10) + (0.13)(0.09) + (0.0)(1.0) + (1.0)(0.0) + 10 = 11.3533 \]

\[ J_{1,2} = \text{average interpopulation homozygosity} = j_{1,2|k}/13 = 11.3533/13 = 0.8733 \]

\[ I_{1,2} = \text{genetic identity among populations 1 and 2} = J_{1,2}/\sqrt{(J_1J_2)} = 0.8733/\sqrt{(0.9567 \times 0.9453)} = 0.9183 \]

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

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

\[ S_{wi} = \frac{2}{2n(2n-1)} \sum_{i<j} (a_{ii} - a_{jj})^2 \]

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

\[ S_w = \frac{1}{d_s} \sum_j S_{wj} \]

Where,

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

\[ S_B = \frac{2}{(2n)^2 d_s (d_s - 1)} \sum_{i<j} \sum_{j<k} (a_{ij} - a_{kj})^2 \]

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

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

![Diagram of simple linkage](image)

\[ d(1,2) = \text{minimum distance between two OTUs} \]
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

$d(1,2) = \text{major distance between two OTUs}$
Complete linkage: an example

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

\[ d(i, j) = \text{average distance between OTU}_i \text{ and OTU}_j \text{ of groups 1 and 2} \]
Average linkage: an example

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
  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

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:

\[ r = \frac{\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i/n}{S_{X_i} S_{Y_i}} \]

Where,

- \( X_i \) and \( Y_i \) are the similarity or distance values of the original and co-phenetic matrix, respectively.
- \( S_{X_i} \) and \( S_{Y_i} \) 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.
In the gel above (top left corner), we have 4 individuals (P₁) and 5 loci (L₁). 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)

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:

\[ M_1: L_1 L_1 L_2 L_3 L_5 \text{ (locus L}_4 \text{ was not drawn)} \]
\[ M_2: L_1 L_2 L_3 L_4 L_3 \]
\[ M_3: L_3 L_1 L_5 L_2 L_4 \]

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

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.
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


Software programs for analysing genetic diversity

Glossary
Appendix 2 to:  
*Measures of Genetic Diversity*

**Analysis of molecular variance: Example 1**

This model, also called AMOVA, measures gene diversity among populations with specific reference to areas of a region in a continent (situation 3, slide 26).

We have: \(i = \text{individuals}, j = \text{alleles}, k = \text{populations}\)

\[
Y_{ijkl} = Y + a_k + b_{ki} + w_{ijkl}
\]

Where,
\begin{align*}
Y_{ijkl} & \quad \text{a value between 0 and 1 for the } j^{\text{th}} \text{ allele of individual } i^\text{th} \text{ of the } k^\text{th} \text{ population} \\
a_k & \quad \text{the effect of the } k^\text{th} \text{ population, with variance } \sigma^2 a \\
b_{ki} & \quad \text{the effect of the } i^\text{th} \text{ individual within the } k^\text{th} \text{ population, with variance } \sigma^2 b \\
w_{ijkl} & \quad \text{the effect of the } j^\text{th} \text{ locus of individual } i^\text{th} \text{ of the } k^\text{th} \text{ population, with variance } \sigma^2 w \\
n & \quad \text{the product of } i, j \text{ and } k, \text{ that is, the total number of observations}
\end{align*}

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among populations</td>
<td>((k - 1))</td>
<td>(\sum X_{ij} - \sum \Sigma X_{ijk} / n)</td>
<td>(MS_a)</td>
<td>(\sigma^2 w + 2\sigma^2 b + 2n\sigma^2 a)</td>
</tr>
<tr>
<td>Among individuals/pop.</td>
<td>((k - 1))</td>
<td>(\sum X_{i} - \sum \Sigma X_{i} / k)</td>
<td>(MS_b)</td>
<td>(\sigma^2 w + 2\sigma^2 b)</td>
</tr>
<tr>
<td>Within individuals</td>
<td>((i - 1))</td>
<td>(\sum X_{ijk} - \sum \Sigma X_{i} / j)</td>
<td>(MS_w)</td>
<td>(\sigma^2 w)</td>
</tr>
<tr>
<td>Total</td>
<td>((ij - 1))</td>
<td>(\sum X_{ij} - \sum \Sigma X_{ij} / i)</td>
<td>(MS_{ij})</td>
<td>(\sigma^2 w)</td>
</tr>
</tbody>
</table>

**Variances and F statistics estimates**

\[
\begin{align*}
\sigma^2 a &= F_{St} / \sigma^2 \\
\sigma^2 b &= (F_{St} - F_{S}) / \sigma^2 \\
\sigma^2 w &= (1 - F_{S}) / \sigma^2 \\
\sigma^2 &= \sigma^2 w + \sigma^2 b + \sigma^2 a
\end{align*}
\]

Where,
\begin{align*}
\sigma^2 a & \quad \text{the parametric value of the variance between populations that carry identical alleles. It is estimated by } (MS_a - MS_b) / 2n \\
\sigma^2 b & \quad \text{the parametric value of the variance between individuals within each population. It is estimated by } (MS_b - MS_w) / 2 \\
\sigma^2 w & \quad \text{the parametric value of the variance within individuals or the measure of the probability that alleles within loci be different. It is estimated by the mean square within individuals (CMW)}
\end{align*}
Appendix 3 to:

*Measures of Genetic Diversity*

Analysis of molecular variance: Example 2

As described in Appendix 2, this model (AMOVA) measures gene diversity among populations, this time with specific reference to populations within an area of a region in a continent (situation 4, slide 26). It has a new hierarchical level (region), with its respective parametric values and estimators of the mean squares.

We have: i = individuals, j = alleles, k = populations, l = regions

$$Y_{ijkl} = Y + r_i + a_{lk} + b_{ki} + w_{ijkl}$$

Where,

- $Y_{ijkl}$ is a value between 0 and 1 for the $j^{th}$ allele of the $i^{th}$ individual of the $k^{th}$ population in the $l^{th}$ region
- $r_i$ is the effect of the $l^{th}$ region, with variance $\sigma^2 r$
- $a_{lk}$ is the effect of the $k^{th}$ population within the $l^{th}$ region, with variance $\sigma^2 a$
- $b_{ki}$ is the effect of the $i^{th}$ individual within the $k^{th}$ population in the $l^{th}$ region, with variance $\sigma^2 b$
- $w_{ijkl}$ is effect of the $j^{th}$ locus within individual $i$ of the $k^{th}$ population of the $l^{th}$ region, with variance $\sigma^2 c$
- $n$ is the product of i, j, k and l, that is the total number of observations

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between regions</td>
<td>l – 1</td>
<td>MS_r</td>
<td>$\sigma^2 w + 2\sigma^2 b + 2n\sigma^2 a + 2nl\sigma^2 r$</td>
</tr>
<tr>
<td>Between pops. within reg.</td>
<td>l(k – 1)</td>
<td>MS_a</td>
<td>$\sigma^2 w + 2\sigma^2 b + 2n\sigma^2 a$</td>
</tr>
<tr>
<td>Between indiv./pop./reg.</td>
<td>lki(i – 1)</td>
<td>MS_b</td>
<td>$\sigma^2 w + 2\sigma^2 b$</td>
</tr>
<tr>
<td>Within individuals</td>
<td>lkij(j – 1)</td>
<td>MS_c</td>
<td>$\sigma^2 w$</td>
</tr>
<tr>
<td>Total</td>
<td>lkij – 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total variance (%)</td>
<td></td>
<td></td>
<td>$% \sigma^2 r = (\sigma^2 r / \sigma^2) \times 100$</td>
</tr>
<tr>
<td>$\sigma^2 = \sigma^2 r + \sigma^2 w + \sigma^2 b + \sigma^2 a$</td>
<td></td>
<td>$% \sigma^2 a = (\sigma^2 a / \sigma^2) \times 100$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$% \sigma^2 b = (\sigma^2 b / \sigma^2) \times 100$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$% \sigma^2 w = (\sigma^2 w / \sigma^2) \times 100$</td>
<td></td>
</tr>
</tbody>
</table>

$\sigma^2 r$ is the parametric value of the variance between regions and is estimated by (CMA – CMB)/2nl.

In the variance estimates, the sign ‘%’ is added because we can express the variance accounted for each source (region, population within a region, individuals within a population) as a function of the total variance and, as such, we can determine which of the variation components is the most important. For example, if the value of the variation due to regions was high and those of the other sources was low, we could conclude that populations within regions have homogeneous allelic frequencies, but populations from different regions differ notably in their allelic frequencies.
Appendix 4 to:

*Measures of Genetic Diversity*

**Geometric distance**

**Quantitative variables**

Geometric distance, also known as the taxonomic distance (Sokal, 1961), is measured by Euclidian distances, following the formula below:

\[ d_{ij} = [\sum k(X_{ik} - X_{jk})^2]^{1/2} \]

Where,

- \( X_{ik} \) = the value of the k\(^{th} \) variable of individual i

See Appendix 5 for an example of calculation.

**Mixed variables**

If there are mixed variables, they should be transformed or standardized first by following the formula:

\[ X_{ij\text{stand}} = \frac{X_i - \bar{X}_i}{s_i} \]

Where,

- \( X_i \) = the i\(^{th} \) character value in individual j
- \( \bar{X}_i \) = the average for the i\(^{th} \)character
- \( s_i \) = the standard deviation for character i\(^{th} \)

**P number of variables**

If there are P numbers of variables, distance must be weighted to become independent of the number of variables, as follows:

\[ d_{ij}^2 = \frac{\sum_k \left[ (X_{ik} - X_{jk}) / \sigma_k \right]^2}{P} \]

Reference

Appendix 5 to:

*Measures of Genetic Diversity*

**Transforming data from quantitative variables:**

**An example**

We have three characters in four individuals:

- Plant height (m)
- Seed weight (g)
- Diameter of the pollen grain (µ)

Before calculating distances, we first need to standardize the data by using the following formula:

\[ m_{\text{stand}} = m - \bar{m}/\sigma \]

After standardization, the measurement units are lost.

<table>
<thead>
<tr>
<th>Individ.</th>
<th>m</th>
<th>(m_{\text{stand}})</th>
<th>g</th>
<th>(g_{\text{stand}})</th>
<th>(\mu)</th>
<th>(\mu_{\text{stand}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiv. 1</td>
<td>1.50</td>
<td>0.35</td>
<td>0.02</td>
<td>0.00</td>
<td>80.00</td>
<td>-0.15</td>
</tr>
<tr>
<td>Indiv. 2</td>
<td>1.20</td>
<td>-1.41</td>
<td>0.03</td>
<td>1.00</td>
<td>70.00</td>
<td>-1.32</td>
</tr>
<tr>
<td>Indiv. 3</td>
<td>1.45</td>
<td>0.06</td>
<td>0.01</td>
<td>-1.00</td>
<td>90.00</td>
<td>1.02</td>
</tr>
<tr>
<td>Indiv. 4</td>
<td>1.60</td>
<td>0.94</td>
<td>0.02</td>
<td>0.00</td>
<td>85.00</td>
<td>0.44</td>
</tr>
<tr>
<td>Average (X)</td>
<td>1.44</td>
<td>---</td>
<td>0.02</td>
<td>---</td>
<td>81.25</td>
<td>---</td>
</tr>
<tr>
<td>Deviation (s_i)</td>
<td>0.17</td>
<td>---</td>
<td>0.01</td>
<td>---</td>
<td>8.54</td>
<td>---</td>
</tr>
</tbody>
</table>

Distances may now be calculated for each pair of individuals by applying the formula we already know:

\[ d_{ij} = \sqrt{(X_{ij} - X_{kj})^2} \]

\[ d_{12} = \sqrt{(0.35 - (-1.41))^2 + (0.0 - 1.0)^2 + (-0.15 - (-1.32))^2} = 2.34 \]

\[ d_{11} = 0 \]

\[ d_{13} = \sqrt{(0.35 - 0.06)^2 + (0.0 - (-1.0))^2 + (-0.15 - (-1.02))^2} = 1.57 \]

\[ d_{14} = \sqrt{(0.35 - 0.94)^2 + (0.0 - 0.0)^2 + (-0.15 - 0.44)^2} = 0.83 \]

\[ d_{22} = 0 \]

\[ d_{23} = \sqrt{(-1.41 - 0.06)^2 + (1.0 - (-1.0))^2 + (-1.32 - (-1.02))^2} = 3.41 \]

\[ d_{24} = \sqrt{(-1.41 - 0.94)^2 + (1.0 - 0.0)^2 + (-1.32 - 0.44)^2} = 3.10 \]

\[ d_{33} = 0 \]

\[ d_{34} = \sqrt{(0.06 - 0.94)^2 + (-1.0 - 0.0)^2 + (1.02 - 0.44)^2} = 1.45 \]

Once we have the pairwise distances, we proceed to find the groups by using the UPGMA method (for details, see slides 58 and 59 in the module).
First, we organize our calculated distance values in a symmetrical table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.34</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.57</td>
<td>3.41</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>3.10</td>
<td>1.45</td>
<td>0</td>
</tr>
</tbody>
</table>

In the first cycle, we choose the shortest distance, which, in our case, is $d_{1,4} = 0.83$. A new matrix may therefore be formed by grouping Individual 1 with Individual 4, and calculating the combined distances:

$$d_{2(1,4)} = \frac{d_{1,2} + d_{2,4}}{2} = \frac{2.34 + 3.10}{2} = 2.72$$

$$d_{3(1,4)} = \frac{d_{1,3} + d_{3,4}}{2} = \frac{1.57 + 1.45}{2} = 1.51$$

We see that the shortest distance now is between $I_{1,4}$ and $I_3$. Then, in a new cycle, a new matrix is formed, grouping Individual 2 with group $I_{3(1,4)}$, and calculating the combined distance $d_{2(1,4)3} = 3.07$

Based on the results above, we can proceed to draw the dendrogram, relating the four individuals of the example:
Appendix 6 to:  
*Measures of Genetic Diversity*

**Applying the simple matching coefficient for morphological characters (categorical variables)**

We have three characters:

- Leaf hairiness: rare (1), common (2), abundant (3)
- Petal colour: white (1), yellow (2), red (3)
- Petiole length: short (1), medium (2), long (3)

We first transform the measurement data to binary data. Note that the three original characters are transformed into 9 binary characters. This could overweight these characters in detriment to others used in the analysis.

<table>
<thead>
<tr>
<th>Character 1</th>
<th>Character 2</th>
<th>Character 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>OTU 2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>OTU 3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>OTU 4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

**Character (binary code)**

<table>
<thead>
<tr>
<th>OTU 1</th>
<th>Rare</th>
<th>Common</th>
<th>Abundant</th>
<th>White</th>
<th>Yellow</th>
<th>Red</th>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>OTU 2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>OTU 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OTU 4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We then apply the simple matching coefficient to calculate the pairwise distances between individuals:

**Pairwise comparisons for all characters**

<table>
<thead>
<tr>
<th>OTU 1 vs. 2</th>
<th>1</th>
<th>0</th>
<th>OTU 1 vs. 3</th>
<th>1</th>
<th>0</th>
<th>OTU 1 vs. 4</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a = 1</td>
<td>b = 2</td>
<td>1</td>
<td>a = 0</td>
<td>b = 3</td>
<td>1</td>
<td>a = 0</td>
<td>b = 3</td>
</tr>
<tr>
<td>0</td>
<td>c = 2</td>
<td>d = 4</td>
<td>0</td>
<td>c = 3</td>
<td>d = 3</td>
<td>0</td>
<td>c = 3</td>
<td>d = 3</td>
</tr>
<tr>
<td>OTU 2 vs. 3</td>
<td>1</td>
<td>0</td>
<td>OTU 2 vs. 4</td>
<td>1</td>
<td>0</td>
<td>OTU 3 vs. 4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>a = 0</td>
<td>b = 3</td>
<td>1</td>
<td>a = 1</td>
<td>b = 2</td>
<td>1</td>
<td>a = 1</td>
<td>b = 2</td>
</tr>
<tr>
<td>0</td>
<td>c = 3</td>
<td>d = 3</td>
<td>0</td>
<td>c = 2</td>
<td>d = 4</td>
<td>0</td>
<td>c = 2</td>
<td>d = 4</td>
</tr>
</tbody>
</table>
Now we can proceed with the methodology to find groups and draw the corresponding phenogram:

<table>
<thead>
<tr>
<th></th>
<th>O1</th>
<th>O2</th>
<th>O3</th>
<th>O4</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O2</td>
<td>0.56</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td>0.33</td>
<td>0.33</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>O4</td>
<td>0.33</td>
<td>0.56</td>
<td>0.56</td>
<td>0</td>
</tr>
</tbody>
</table>

Phenogram

Coefficient

0.25  0.44  0.63  0.81  1.00
Appendix 7 to:

*Measures of Genetic Diversity*

**Calculating Nei’s genetic distance**

First, with the data obtained in the example (see slide 48), we make the distance matrix as follows:

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>0.0852</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>0.0107</td>
<td>0.0440</td>
<td>0</td>
</tr>
</tbody>
</table>

In the first cycle, we choose the shortest distance: \(d_{1,3} = 0.0107\)

Then, in the second cycle, a new matrix is formed by grouping Individual 1 with Individual 3, and calculating the combined distances:

\[
d_{21,3} = (d_{172} + d_{273})/2 = (0.0852 + 0.044)/2 = 0.0646
\]

\[
\begin{array}{c|c|c}
 & P1,3 & P2 \\
\hline
P1,3 & 0     \\
P2  & 0.0646 & 0
\end{array}
\]

We can now draw the dendrogram:
Appendix 8 to:  
*Measures of Genetic Diversity*

**Morphological and molecular similarities**

Let's say we have three individual roses (1, 2, 3). Numbers 2 and 3 look morphologically alike, whereas number 1 looks different.

If we look at the DNA fragments, supposedly generated with a molecular marker, we see that Individuals 1 and 2 seem more similar. So, what happened? This shows the importance of studying genetic diversity at all possible levels. The combination of information from different types of markers—that is, those that refer to functional genes and those that show polymorphism in the genomic regions—will give the best approximation to knowledge on present genetic variation. The same would apply if we could combine morphological and molecular data.

In this Appendix, we show the type of errors we may incur if conclusions are based on only one type of marker data.
Based on the DNA band profile obtained in the gel for the three individuals, the pairwise distances are calculated, using the Jaccard coefficient:

\[ J_{1,2} = \frac{11}{11+1+2} = 0.786 \]
\[ J_{1,3} = \frac{6}{6+6+3} = 0.400 \]
\[ J_{2,3} = \frac{5}{5+8+4} = 0.294 \]

Based on the DNA band profile obtained in the gel for the three individuals, the pairwise distances are calculated, using the Jaccard coefficient:

\[ J = \frac{a}{a + b + c} \]

We then make the distance matrix and draw the dendrogram:
This dendrogram results from the molecular data. We can now compare it with another dendrogram, shown below, that was developed from morphological observations, and note differences. According to the molecular dendrogram, Individuals 1 and 2 are closer together, even though the morphological data suggest that Individuals 2 and 3 are closer.

But, we can use both molecular and morphological data, combining them and redoing the process with both data simultaneously.
The combined dendrogram shows grouping distances that differ from either molecular or morphological dendrogram alone. We may therefore assume that the information provided by combining the data is closer to the reality of this situation.

\[
J_{1,2} = \frac{11}{11 + 2 + 2} = 0.733 \\
J_{1,3} = \frac{6}{6 + 7 + 3} = 0.375 \\
J_{2,3} = \frac{6}{6 + 8 + 4} = 0.333
\]

\[
S_{(1,2)3} = \frac{S_{1,3} + S_{2,3}}{2} = 0.554
\]
Genetic diversity analysis with molecular marker data: Learning module

Software programs for analysing genetic diversity
Contents

- Main characteristics
  - A summary table
  - Some software programs, their authors, Web sites …
  - … and other features
- Five software programs in detail
  - Arlequin
  - PowerMarker
  - DnaSP
  - PAUP*
  - MEGA
- Internet resources
- Appendix 9: References to software programs
Numerous software programs are available for assessing genetic diversity. Most are freely available through Internet. Many perform similar tasks, with the main differences being in the user interface, type of data input and output, and platform. Thus, choosing which to use depends heavily on individual preferences. In this section, we describe some of the programs available, noting specific options that users may find preferable.
Joanne Labate (2000) wrote an excellent review of six programs: TFPGA (Miller, 1997), Arlequin (Schneider et al., 1997), GDA (Lewis and Zaykin, 1999), GENEPOP (Raymond and Rousset, 1995), GeneStrut (Constantine et al., 1994), and POPGENE (Yeh et al., 1997). Her review includes the particular options of each program, a table of functions available in each, and Web sites where they can be downloaded. To avoid redundancy, we have included only the Arlequin, which is possibly the most widely used program of the six.

For full references to these six programs and selected others, see Appendix 9.

**Reference**

We review other programs, selecting them for their wide use and giving priority to those that are available for no charge (except for PAUP*). Web sites listing and linking many other available programs are also given in this and following slides. We include information on authors, costs, platform specificities and Web sites. Note that, while these programs are sometimes made specifically for only one platform (usually Windows or Macintosh), with the recent advent of ‘emulators’ (such as SoftWindows, VirtualPC), most programs can be run on any computer, regardless of platform. Although we note the cases where a program has been successfully used with one of these emulators, we do not say that all listed programs can be used across platforms; simply that we know for sure that these have been successfully used. Sometimes, using emulators can cause the program to run more slowly or create other problems. Where possible, prefer using the platform for which the program was designed.

Programs listed in **bold face** in the slide are discussed in this paper. Web sites were available as of 28 February, 2003.
<table>
<thead>
<tr>
<th>Name</th>
<th>Windows</th>
<th>Macintosh</th>
<th>Other</th>
<th>Cost (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlequin</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>DnaSP</td>
<td>X</td>
<td>c</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>PowerMarker</td>
<td>X</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>MEGA2</td>
<td>X</td>
<td>c</td>
<td>c</td>
<td>0</td>
</tr>
<tr>
<td>Arlequin2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>100</td>
</tr>
<tr>
<td>TFPGA</td>
<td>X</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>GDA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>GENEPOP</td>
<td></td>
<td>X</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>NTSYSpc</td>
<td>X</td>
<td></td>
<td></td>
<td>230-300</td>
</tr>
<tr>
<td>structure</td>
<td>X</td>
<td></td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>GeneStrut</td>
<td></td>
<td>X</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>POPGENE</td>
<td>X</td>
<td>c</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>MacClade</td>
<td></td>
<td>X</td>
<td></td>
<td>125</td>
</tr>
<tr>
<td>PHYLIP</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>SITES</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>CLUSTAL W</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>MALIGN</td>
<td></td>
<td></td>
<td>X</td>
<td>0</td>
</tr>
</tbody>
</table>

*c = program runs well with an emulator such as SoftWindows or VirtualPC.

To save space, references to the programs discussed are given in Appendix 9.
Five software programs were selected to show detail. The choice was made after informally surveying users on the programs they use most and their opinions as to the most useful or representative. Users included graduate students, postdoctorates and research associates, as well as faculty. A list was compiled of the most-mentioned programs. In cases of doubt, those that were freely available or seemed more widely used were chosen. To be more representative, an additional criterion was to choose those programs that did different things.
Arlequin

- <http://lgb.unige.ch/arlequin>
- Windows, Mac, and Linux versions available, all free of charge
- Can use different types of data (but not yet dominant markers)
- Missing or ambiguous data can be included
- Data can be entered as DNA sequences, RFLP haplotypes, microsatellite profiles or multilocus haplotypes
- Data can be imported from files created for other programs

(continued on next slide)

Released in 1997, Arlequin (current version 2.001) is still very popular. It is ‘an exploratory population genetics software environment able to handle large samples of molecular data (RFLPs, DNA sequences, microsatellites), while retaining the capacity of analysing conventional genetic data (standard multi-locus data or mere allele frequency data).’

Arlequin can use many different types of data, such as molecular data and genotype or haplotype frequencies, including codominant or recessive data but not yet dominant data. Molecular data can be entered as DNA sequences, RFLP haplotypes, microsatellite profiles, or multilocus haplotypes. The data format is specified in an input file. The user can create a data file from scratch, using a text editor and appropriate keywords, or use the ‘Project Outline Wizard’. Data can be imported from files created for other programs, including MEGA, BIOSYS, GENEPOP, and PHYLIP. Missing or ambiguous data can be included. A very detailed user manual is available, which includes a large amount of theoretical information, formulae, and references. A large number of data can be analysed, and a Batch Files option is available.

Authors: Laurent Excoffier, Stefan Schneider and David Roessli, University of Geneva, Switzerland.

Reference

Arlequin (continued)

- Advantages:
  - Good support provided through a detailed manual, which includes large amounts of theoretical information, formulae, and references; and a well-organized Web site with such features as ‘Frequently Asked Questions’
  - The graphic interface is very user-friendly

- Disadvantages:
  - Numerous features and options to learn
  - Setting up a data file can be complex and must be formatted correctly (but good examples are given in the manual)
PowerMarker

- [http://www.powermarker.net](http://www.powermarker.net)
- Designed for use with SSR/SNP data in population genetics analyses
- Available options include summary statistics, consensus trees, population structure, Mantel’s test, triangle plotting and visualization of linkage disequilibria results

(continued on next slide)

PowerMarker is a new program, with the first official version released in January 2004. It was designed specifically for the use of SSR/SNP data in population genetics analyses. Data can be imported from Excel or other formats, making data set-up very easy. Data can also be exported to NEXUS and Arlequin formats. It includes a ‘2D viewer’ for linkage disequilibrium visualization. The user can edit graphics within PowerMarker or export them for publication. The program has been tested extensively for accuracy and efficiency. Full documentation is included. Several new modules for association study are included in the package. Several demonstration datasets available to get started. The program is free, but requires having PHYLIP, TreeView and the Microsoft.net framework system (all freely available) and Excel 2000 (not free). Another disadvantage is that it is available only for Windows 98 and above (not for Macintosh or other systems). Email support (registered user only): powermarker@hotmail.com

Author: Kejun (Jack) Liu, North Carolina State University.

Reference

PowerMarker (continued)

Advantages:
- Allows importing data from Excel, which makes data management very easy
- The graphical interfaces are very user-friendly
- Graphics are ‘guaranteed’ to have publication quality

Disadvantages:
- Very new program with possible bugs still to be ironed out
- Not a stand-alone program, needs downloading of several other software programs and buying Excel
DnaSP

- <http://www.ub.es/dnasp>
- It uses DNA sequence data to perform population genetics analyses
- It performs a very large number of analyses, including measures of polymorphism, divergence between populations (including a measure of gene flow), synonymous and non-synonymous substitutions, linkage disequilibrium, recombination, and many statistical tests (Hudson, Kreitman and Aguade’s, Fu and Li’s, Tajima’s, and the McDonald and Kreitman test)

DnaSP, for DNA Sequence Polymorphism, uses DNA sequence data. This program is widely used for sequence analysis because it does all the necessary analyses and at the same time is easy to use. It was written exclusively for the Windows operating system, but can be run on a Macintosh using SoftWindows or VirtualPC software emulators. DnaSP can import and export several types of data formats, including FASTA and NEXUS, which is very convenient, and can handle large numbers of long sequences, depending on your computer’s memory. The authors are currently working on version 4. It is freely available, downloadable from the Web site. Although no manual is available, a Help file is incorporated into the program. In addition, the Web site includes much explanatory material, as well as many references. The authors have several publications about the program (e.g. see citations below).

Authors: Julio and Ricardo Rozas

References


DnaSP (continued)

Advantages:
- The Windows user-interface makes the program very easy to use
- A very large number of analyses are possible

Disadvantages:
- Currently only available for Windows users (and Macintosh with the use of a Windows emulator)
- No manual or tutorials are yet available
- Support is, apparently, not readily available
PAUP* is widely used for inferring and interpreting evolutionary trees. It originally meant Phylogenetic Analysis Using Parsimony, but now has many other options. PAUP* is available from Sinauer Associates, Sunderland, MA, at <http://www.sinauer.com/detail.php?id=8060>. Although not free, it is relatively inexpensive (US$100 at writing). A new version, 4.0 beta, has been released as a provisional version. Macintosh, PowerMac, Windows and Unix/OpenVMS versions are available; the Mac version has some extra features. PAUP* is closely compatible with MacClade (another program available from Sinauer), since they use a common data format (NEXUS, Maddison et al. 1997).

Author: David Swofford, Laboratory of Molecular Systematics, National Museum of Natural History, Smithsonian Institution, Washington, DC.

Reference

PAUP* (continued)

- **Advantage:** Good support available, both online (including a long list of ‘Frequently Asked Questions’) and via e-mail.

- **Disadvantage:** It is not free.
MEGA (Molecular Evolutionary Genetics Analysis) software has been widely used since its creation in 1993; MEGA2 has since come out. It uses DNA sequence, protein sequence, evolutionary distance or phylogenetic tree data. The authors’ goal was to take advantage of advances in computer power and graphic user interfaces to make available a ‘flexible and easy-to-use genetic data analysis workbench’. Although it was designed for the Windows platform, it runs well on Macintosh with a Windows emulator, Sun workstation (with SoftWindows95) or Linux (with Windows by VMWare). The newest version, 2.1, has many important additions, such as the ability to import data from NEXUS or CLUSTAL W, unlimited dataset sizes, and many others.

A book by the software authors Nei and Kumar (2000) includes theoretical information about statistical analyses and how to interpret results from both their software and other software programs. Online, a thorough manual is available (although the format is not easy to page through), together with a bulletin board for users to interact with each other.

Authors: Sudhir Kumar, Koichiro Tamura, Ingrid Jakobsen and Masatoshi Nei.

References


MEGA (continued)

- Advantages:
  - As this software has been available since 1993, it’s likely that most of the bugs have already been discovered
  - There is good support available, including an online manual and a published book

- Disadvantage:
  It is not the easiest program for beginners to learn
In this and the next slide, we give a sample list of Internet resources that you may find useful for locating information related to, for example, genetic diversity analysis, population genetics, other software available, and links to useful extra information. For each resource presented, we briefly describe their contents in the notes below.

- **The European Molecular Biology Laboratory–European Bioinformatics Institute (EBI) site**
  <http://www.ebi.ac.uk/services/>

- **Biological software page from the Institut Pasteur in France**
  <http://bioweb.pasteur.fr/intro-uk.html>

- **Phylogeny Programs (listed by Joe Felsenstein at the University of Washington)**

(continued on next slide)
Dr. Ed Buckler’s Maize Genetics site
<http://www.maizegenetics.net/bioinformatics/index.htm>

Kent Holsinger’s site at the University of Connecticut
<http://darwin.eeb.uconn.edu/archives.html>

Claire Constantine’s site at Murdoch University
<http://wwwvet.murdoch.edu.au/vetschl/imgad/GSLinks.htm>

Software page of the Institute of Forest Genetics and Forest Tree Breeding, University of Göttingen, Germany
<http://www.uni-forst.gwdg.de/forst/fg/index.htm>

- Dr. Ed Buckler’s Maize Genetics site: contains freely available software programs developed by his laboratory group. Although some information is specific to maize, the site also contains useful information about genomics, links to many journals, and PDF versions of Dr. Buckler’s publications.

- Kent Holsinger’s site at the University of Connecticut: has links to many software programs, including ones for biology, programming and statistics. The author does not post things he does not use regularly, so some guarantee of good quality exists. At writing, this was also the most recently updated.

- Claire Constantine’s site at Murdoch University: although not updated recently, this site contains links to the most used programs for population genetics analyses. Even more useful, it includes a comparison table of the kinds of statistics available in each of 7 commonly used programs (Arlequin, GENEPOP, POPGENE, GDA, GeneStrut, DnaSP and SITES).

- Software page of the Institute of Forest Genetics and Forest Tree Breeding, University of Göttingen, Germany: this site contains just 4 software programs, developed in-house, but they are freely available, include good descriptions, and the page is regularly updated.
Appendix 9. References to software programs
In summary

- Many computer programs exist for analysing molecular data for genetic diversity.
- Most programs perform similar tasks and their main differences should be evaluated, depending on resources available and/or individual preferences.
- Nowadays, in addition to freely available computer programs, plenty of resources are also found on Internet to help us obtain both basic and more specialized information on methods.
By now you should …

▶ Be familiar with:
  • The contrasting features of available software for genetic diversity analysis
  • The advantages and disadvantages of some of the major computer programs

▶ Have become acquainted with some Internet resources that may assist you with the task of studying genetic diversity
References


Glossary
Appendix 9 to:

*Software Programs for Analysing Genetic Diversity*

References to software programs

**Arlequin**


**CLUSTAL W**


**DnaSP**


**GDA**

Lewis, P.O. and D. Zaykin. 1999. Genetic Data Analysis: Computer Program for the Analysis of Allelic Data, Version 1.0 (d12). Distributed by the authors.

**GENEPOP**


**GeneStrut**


**MacClade**

MALIGN


MEGA2


NTSYSpc


PAUP*


PHYLIP


POPGENE


PowerMarker


SITES


Structure

TFPGA

Genetic diversity analysis with molecular marker data:
Learning module

Glossary

**AFLP:** Amplified fragment length polymorphism. A highly sensitive method for detecting polymorphisms in DNA. DNA first undergoes restriction enzyme digestion, and a subset of DNA fragments is then selected for PCR amplification and visualization.

**Allele:** One of the alternative forms of a gene that can exist at a single locus.

**Allogamy:** Transfer of pollen (i.e. pollination) from the anther of the flower of one plant to the stigma of the flower of a genetically different plant. Also called cross-breeding, outcrossing, and xenogamy. *See also* Outbreeding.

**AMOVA:** The analysis of molecular variance is a method for studying molecular variation within a species.

**Asexual reproduction:** The formation of new individuals from the cell(s) of a single parent. It does not entail recombination or mixing of parental forms.

**Autogamy:**
- Transfer of pollen (pollination) from the anther of a flower to the stigma of the same flower or sometimes to that of a genetically identical flower (as of the same plant or clone)
- The ability of many plant species to naturally and successfully fertilize within one individual. Also called self-pollination.

**Biodiversity:** The totality of genes, species, and ecosystems in a given region, be it a microhabitat or the world. Also called biological diversity.

**Breeding:** The propagation and genetic manipulation by hybridization or deliberate self-crossing of plants, for the purpose of selecting improved offspring.

**Breeding system:** The system by which a species reproduces. There are several natural systems in plants, for example, *see also* Outbreeding and Inbreeding.

**Characterization:** Assessment of plant traits that are highly heritable, easily seen by the eye, equally expressed in all environments, and usable for distinguishing phenotypes.
Collection (of plant genetic resources):

- The gathering together of domesticates (landraces, old and modern cultivars and breeding lines), and related wild or weedy species.
- The material gathered by the act of collecting, is termed a collection.

Conservation: The management of human use of the biosphere so that it may yield the greatest sustainable benefit to current generations while maintaining its potential to meet the needs and aspirations of future generations. Thus, conservation is positive, embracing preservation, maintenance, sustainable use, restoration, and enhancement of the natural environment.

Diploid: A full set of genetic material, consisting of paired chromosomes, with one chromosome from each parental set. Most animal cells, except the gamete, have a diploid set of chromosomes. (cf. Haploid)

Domestication: The evolution of plants or animals either naturally or through artificial selection, to forms more useful to man, e.g. nonshattering seed.

ex situ conservation:

- A conservation method that entails removing germplasm resources (seeds, pollen, sperm, individual organisms) from their original habitat or natural environment.
- Keeping components of biodiversity alive outside their original habitat or natural environment. (cf. in situ conservation)

Gene: The basic physical and functional unit of heredity, which passes information from one generation to the next. It is a segment of DNA that includes a transcribed section and a regulatory element, which allows its transcription.

Gene flow: The exchange of genetic material between populations. This may be used in the sense of plant reproduction (i.e. due to the dispersal of gametes and zygotes) or due to human influences, such as the introduction of new crop varieties by farmers.

Genetic distance: The degree of relatedness between subgroups or populations as measured by various statistics.

Genetic diversity: Variation in the genetic composition of individuals within or among species; the heritable genetic variation within and among populations.

Genetic drift: The unpredictable changes in allele frequency that occur in small populations.

Genetic erosion: Loss of genetic diversity between and within populations of the same species over time, or reduction of the genetic base of a species.

Genetic marker: An allele, a band in a gel or trait that serves experimentally as a probe to identify an individual or one of its characteristics.

Genome: The entire complement of genetic material in an organism.
**Haploid:** A single set of chromosomes (half the full set of genetic material), present in each egg and sperm cell of animals and in each egg and pollen cell of plants (*Gk. haploos, single*).

(cf. **Diploid**)

**Haplotype:** A specific allelic constitution at a number of loci within a defined linkage block.

**Hardy-Weinberg equilibrium:** The stable frequency distribution of genotypes, AA, Aa, and aa, in the proportions $p^2$, $2pq$, and $q^2$, respectively (where $p$ and $q$ are the frequencies of the alleles, A and a), that is a consequence of random mating in the absence of mutation, migration, natural selection or random drift.

**Heterozygote:** A diploid individual that has different alleles at one or more genetic loci (*Gk. heteros, other*)

(cf. **Homozygote**)

**Homozygote:** A diploid individual that has identical alleles at one or more genetic loci (*Gk. homos, same*)

(cf. **Heterozygote**)

**Inbreeding:** The mating of genetically related individuals or between relatives. Breeding through a succession of parents belonging to the same stock. Also called endogamy or self-breeding.

(cf. **Outbreeding**)

**in situ conservation:** A conservation method that attempts to preserve the genetic integrity of gene resources by conserving them within the evolutionary dynamic ecosystems of the original habitat or natural environment.

(cf. **ex situ conservation**)

**Isozyme:** Multiple forms of an enzyme whose synthesis is controlled by more than one gene.

**Locus (pl. loci):** The specific place on a chromosome where a gene or particular piece of DNA is located.

**Marker:** An identifiable physical location on a chromosome whose inheritance can be monitored (e.g. gene, restriction enzyme site or RFLP marker).

**Mating system:** The pattern of mating between individuals of a population, including such factors as extent of inbreeding, pair-bonding, and number of simultaneous mates. The mating system is of major importance in determining both the genetic structure and evolutionary potential of natural populations.

**Microsatellite DNA:** A type of repetitive DNA based on very short repeats such as dinucleotides, trinucleotides or tetranucleotides. Also called simple sequence repeats (SSRs).

**Migration:** Movement of individuals between otherwise reproductively isolated populations.
**Multiple alleles:** The existence of several known alleles of a gene.

**Mutation:** The term to describe an abrupt change of phenotype that is inherited. Any permanent and heritable change in DNA sequence. Types of mutations include point mutations, deletions, insertions, and changes in number and structure of chromosomes.

**Outbreeding:** An allogamous mating system in which mating is between individuals that are less closely related than are average pairs chosen from the population at random. Also called exogamy or cross-breeding. (cf. Inbreeding)

**Outcrossing:** see Allogamy.

**Phylogeny:** Evolutionary history of a species. A diagram illustrating the deduced evolutionary history of populations of related organisms.

**Polymorphism:** The appearance of different forms associated with various alleles of one gene or homologous of one chromosome.

**Population:** In genetics, a group of individuals who share a common genepool and have the potential to interbreed.

**Population genetics:** The quantitative study and measurement of populations in statistical terms, e.g. the study of genetic phenomena in terms of standard statistical parameters such as frequency tables and distributions, means, variance and standard deviations.

**Purine:** A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine.

**Pyrimidine:** A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine. The pyrimidines in RNA are cytosine and uracil.

**RAPD:** Random amplified polymorphic DNA. A technique for amplifying anonymous stretches of DNA, using PCR with arbitrary primers.

**Recombination:** The production of a DNA molecule with segments derived from more than one parental DNA molecule. In eukaryotes, this is achieved by the reciprocal exchange of DNA between non-sister chromatids within a homologous pair of chromosomes during prophase of the first meiotic division. Recombination allows the chromosomes to rearrange their genetic material, thereby increasing the potential of genetic diversity. Also known as crossing-over.

**Regeneration (of genetic resources collections):**
- The process of restoring a whole plant from individual cells by manipulating an *in vitro* culture.
- The growing of a sample of seeds from an accession to replenish the viability of the original accession. It is usually done when the viability of the original material drops to less than 85%.
**Restriction enzyme:** An endonuclease that will recognize a specific sequence and cut the DNA chain at that point.

**RFLP:** Restriction fragment length polymorphism. Variation between individuals as detected by differences in DNA fragment sizes after restriction digestion.

**Selfing:** To fertilize by means of pollen from the same plant.

**Self-pollination:** see Autogamy.

**Sexual reproduction:** The production of new individuals, following the mixing in a single cell of the genes of two different cells, usually gametes and usually from different parents.

**Species diversity:** A function of the distribution and abundance of species, similar in meaning to ‘species richness’. In more technical literature, includes considerations of the evenness of species abundances. An ecosystem is said to be more diverse, according to the more technical definition, if species present have equal population sizes and less diverse if many species are rare and some are very common.

**SSR:** See Microsatellite DNA.

**Sympatric:** Occurring in the same geographic area.

**Weed:**
- In agriculture, an individual plant or species growing where it is not wanted.
- In ecology, a plant that is adapted to grow in disturbed or open habitats, e.g. after fire or human disturbance.

**Wild relative:** A relative of a crop species that grows in the wild and is not used for agricultural purposes.
Feedback form

The subject of this module is continuously changing, making periodical updating desirable. For this reason we offer the module in electronic format. To help us with the task of updating, we would appreciate receiving relevant comments and requests from users to further improve the module. Your comments on this tool’s usefulness will also be very helpful.

Below we have asked a few guiding questions for you to answer. You may send your answers to us directly by clicking the ‘Send’ button at the end of this page, by fax (no.: 57-2 445 0096), by post to your contact in IPGRI or by e-mail to either cdevicente@cgiar.org; cflb@lamolina.edu.pe; or tf12@cornell.edu

Questions:

1. Tell us how you use this tool…

2. Do you think the module:
   a. Provides an overview of the basic scientific concepts of population genetic for assessing genetic diversity?
      Yes
      No
      Why?

   b. Shows how the use of molecular marker technologies can be a practical approach for anyone planning research on genetic diversity analysis.
      Yes
      No
      Why?

   c. Gives information on how to analyze and interpret molecular marker data.

   d. Makes available a list of current bibliographical resources for each of the items discussed?
      Yes
      No
      Why?
3. Tell us what you liked and what you would change about the following:
   a. Contents, in general
   b. Structure within the submodules
   c. Drawings and diagrams
   d. Appendices
   e. Other

4. What would you add to the module to make it more useful in your particular situation?
   a. Basic references
   b. Examples of applications
   c. Additional information on:
      I. Concepts of population genetics
      II. Measures of genetic diversity
      III. Software described
   d. Other

5. What other topics would you like to see in this module?

6. Was the module, or parts of it, difficult to download? Why?
IPGRI—The International Plant Genetic Resources Institute (IPGRI) is an autonomous international scientific organization, supported by the Consultative Group on International Agricultural Research (CGIAR). IPGRI’s mandate is to advance the conservation and use of genetic diversity for the well-being of present and future generations. IPGRI has its headquarters in Rome, Italy, and offices in another 22 countries worldwide. It operates through three programmes: (1) the Plant Genetic Resources Programme; (2) the CGIAR Genetic Resources Support Programme; and (3) the International Network for the Improvement of Banana and Plantain (INIBAP).

IGD—The Institute for Genomic Diversity is located at Cornell University in Ithaca, New York, USA. IGD’s mission is to develop, transfer, and provide technologies and educational resources to solve problems affecting the conservation of biodiversity and global food security. Specifically, our goals are: to utilize evolutionary and comparative genomics to solve applied problems; to develop and transfer enabling technologies in genomics and bioinformatics; to provide continuing support for national and international efforts in conservation, agriculture, and crop improvement; and to educate and train at all levels (including but not limited to undergraduate and graduate students, visiting scientists, faculty, teachers).

Copyright © 2004 International Plant Genetic Resources Institute (IPGRI) and Institute for Genomic Diversity (IGD), Cornell University.

Except as expressly stated otherwise, the copyright and all other rights related to this work are held by IPGRI and Cornell University. Individuals may freely copy and print materials for educational or other non-commercial purposes without prior permission from the copyright holders. Acknowledgement of the source of the materials is required.

Citation

M. Carmen de Vicente, PhD, is Plant Molecular Geneticist at the IPGRI-Americas Regional Office, Cali, Colombia; César López, MSc, is Lecturer of Genetics at the Universidad Nacional Agraria ‘La Molina’ in Lima, Peru; and Theresa Fulton, PhD, is Director of Outreach at the Institute for Genomic Diversity of Cornell University, Ithaca, NY.