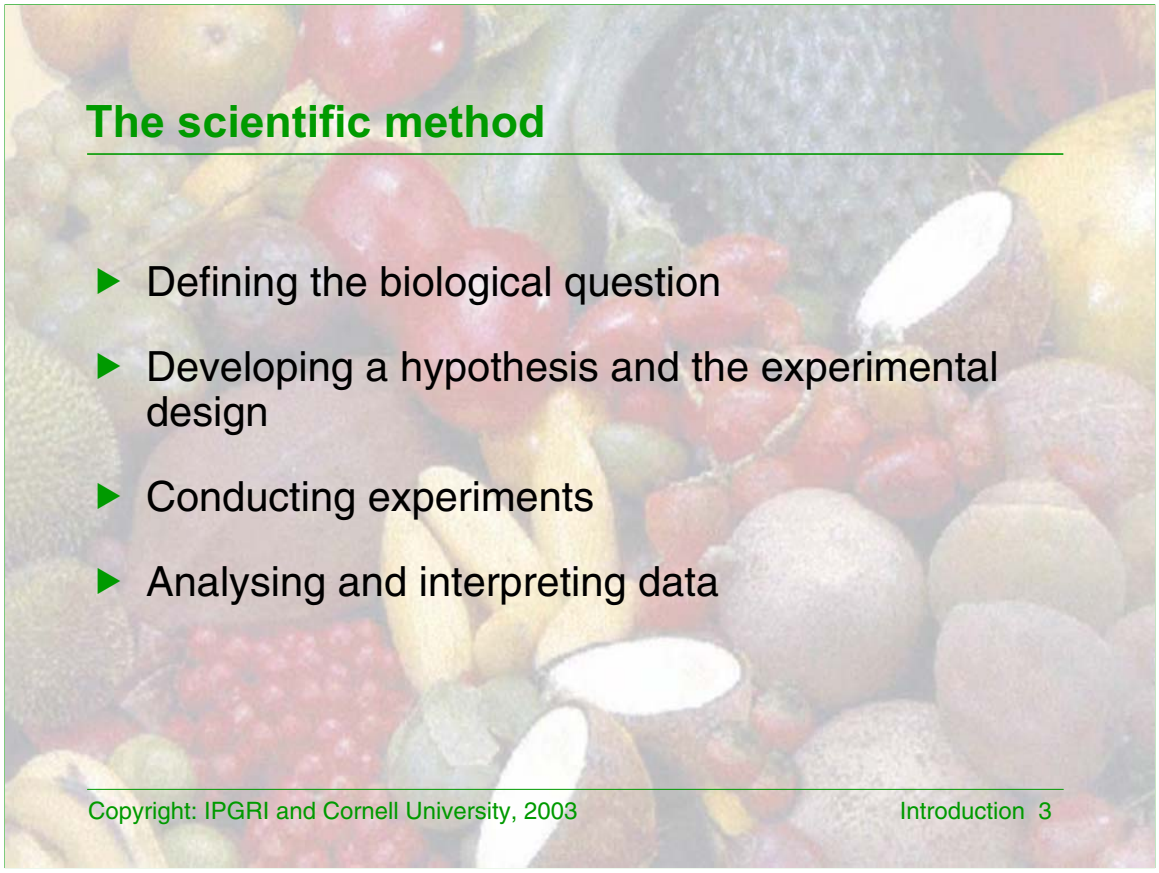
Genetic diversity analysis with molecular marker data: Learning module

Introduction

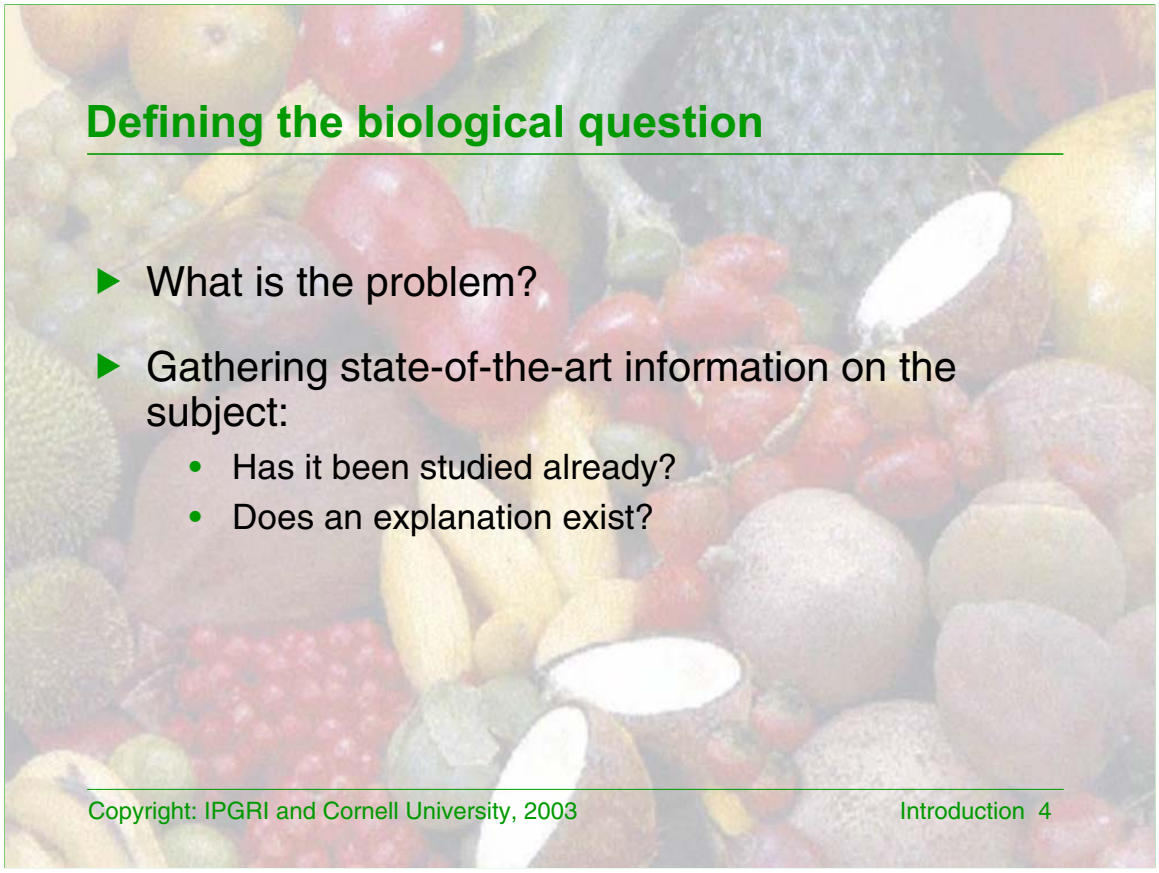
Contents



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.



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.

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?

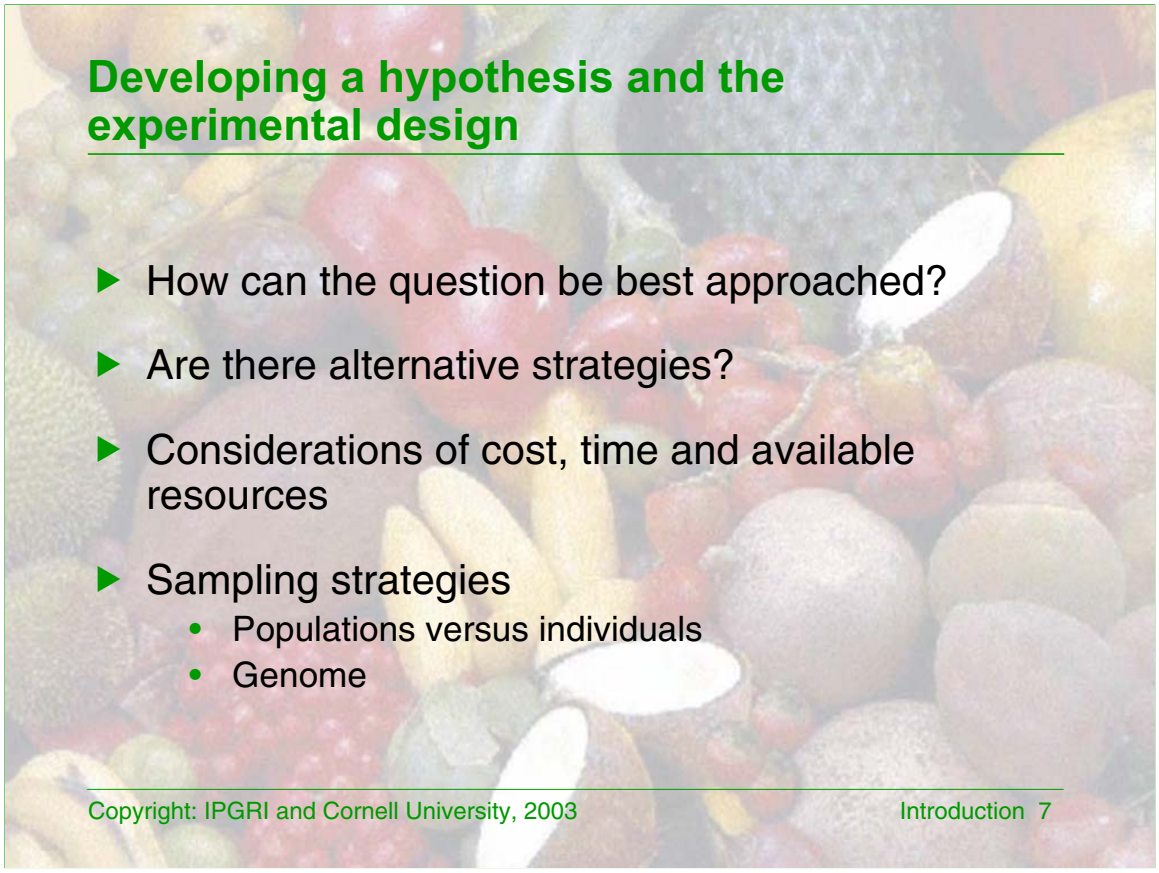
Loci	A	B	C	D	E	F	G	H	I	
Allele	A_1	B_1	C_1	D_1	E_1	F_1	G_1	H_1	I_1	
	A_2	B_2	C_2	—	E_2	F_2	G_2	—	I_2	
	—	—	C_3	—	E_3	—	G_3	—	I_3	
	—	—	—	—	E_4	—	—	—	—	
Genotype	A_1 A_1	B_1 B_1	C_1 C_1	D_1 D_1	E_1 E_1	F_1 F_1	G_1 G_1	H_1 H_1	I_1 I_1	
	A_1 A_2	B_1 B_2	C_1 C_2		E_1 E_2	F_1 F_2	G_1 G_2		I_1 I_2	
	A_2 A_2	B_2 B_2	C_1 C_3		E_1 E_3	F_2 F_2	G_1 G_3		I_1 I_3	
			⋮		⋮		⋮		⋮	
			etc.		etc.		etc.		etc.	
Total number of genotypes			$\Sigma K(K+1)/2$							
Loci	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
Genotypes (no.)	3	3	6	1	10	3	6	1	6	58320
One individual	A_1 A_2	B_2 B_2	C_2 C_3	D_1 D_1	E_2 E_4	F_1 F_1	G_1 G_3	H_1 H_1	I_2 I_2	
Another individual	A_1 A_1	B_1 B_2	C_1 C_1	D_1 D_1	E_1 E_3	F_2 F_2	G_2 G_3	H_1 H_1	I_3 I_3	

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: A_1 A_1, A_1 A_2, A_2 A_2, B_1 B_1 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.

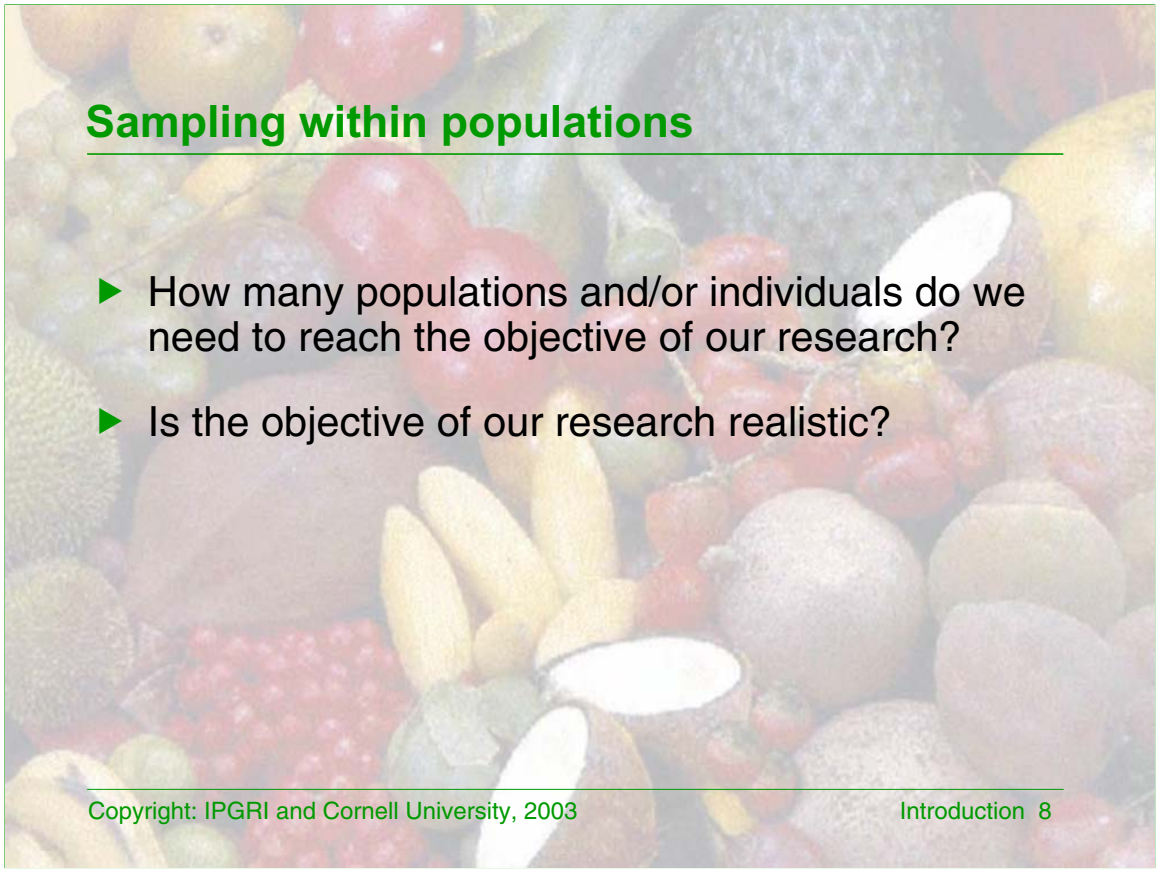


A good hypothesis is an assumption about an observation and should be amenable to experimental validation, which later serves to build broad conclusions. Defining a good hypothesis for studying genetic diversity requires a basic knowledge of, for example, reproduction systems, spatial and temporal distribution, and interspecific interactions. For instance, if we want to analyse 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).



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.

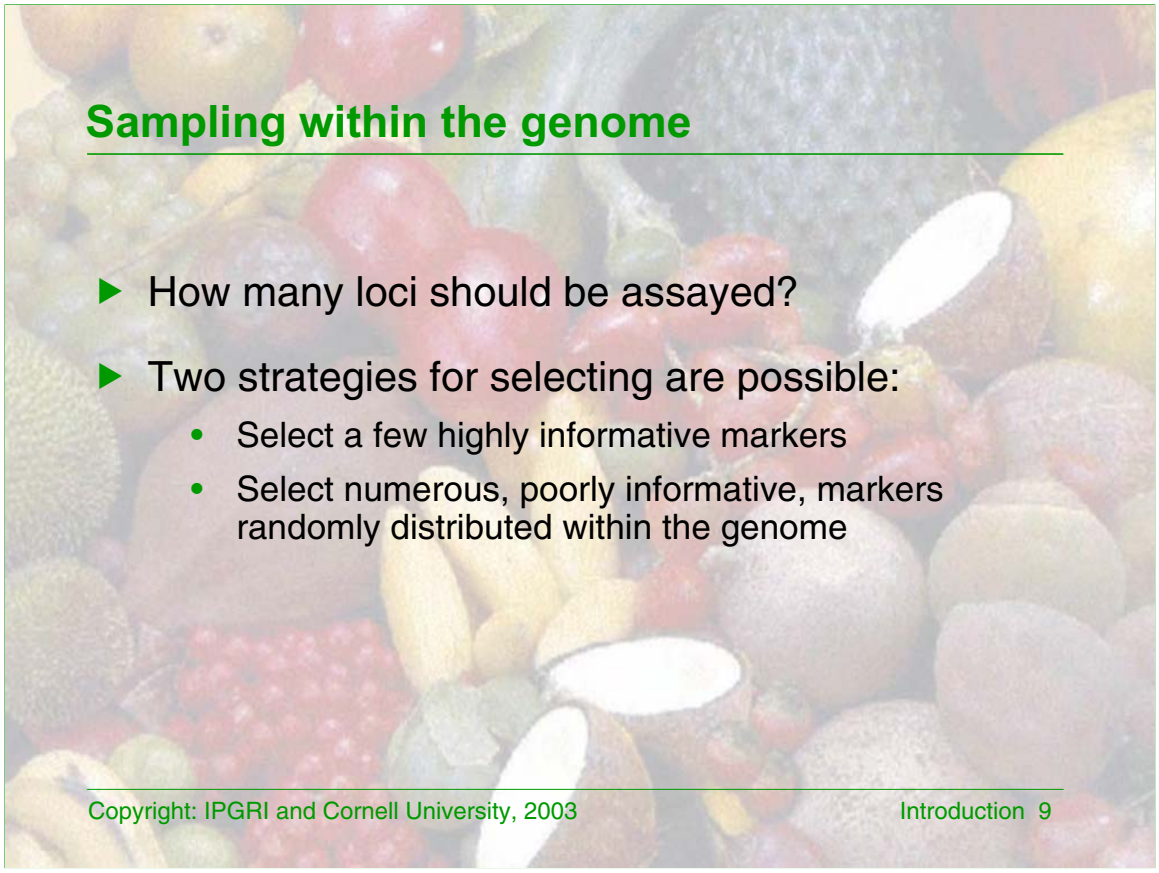
While no comprehensive publication dealing with questions for all situations exists, treatment of some situations can be found in the literature. For example, Brown and Weir (1983) on sampling to estimate allele frequency, number of alleles per locus and gene diversity. They, in turn, mention basic literature 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

Brown, A.H.D. and B.S. Weir. 1983. Measuring genetic variability in plant populations.
Pp. 219-239 *in* Isozymes in Plant Genetics and Breeding, part A (S.D. Tanksley and T.J. Orton, eds.). Elsevier Science Publishers, Amsterdam.

Gregorius, H.R. 1980. The probability of losing an allele when diploid genotypes are sampled. Biometrics 36:643-652.



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

Mariette, S., V. Lecorre and A. Kremer. 1999. Sampling within the genome for measuring within-population diversity: Trade-offs between markers *in* Sampling Strategies for Marker Analysis. A compendium of the research project *Development, Optimization and Validation of Molecular Tools for the Assessment of Biodiversity in Forest Trees* as part of the European Union DGXII Biotechnology FW IV Research Programme *Molecular Tools for Biodiversity*. (E.M. Gillet, ed.). <http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm>

Problems of bad sampling

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

	Genotypes			Allele freq.	
Indivs. (no.)	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	p	q
10	7	1	2	0.75	0.25
30	4	16	10	0.4	0.6
100	27	39	34	0.465	0.535
1000	299	395	306	0.4965	0.5035
1 000 000	300 000	400 000	300 000	0.5	0.5

(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:

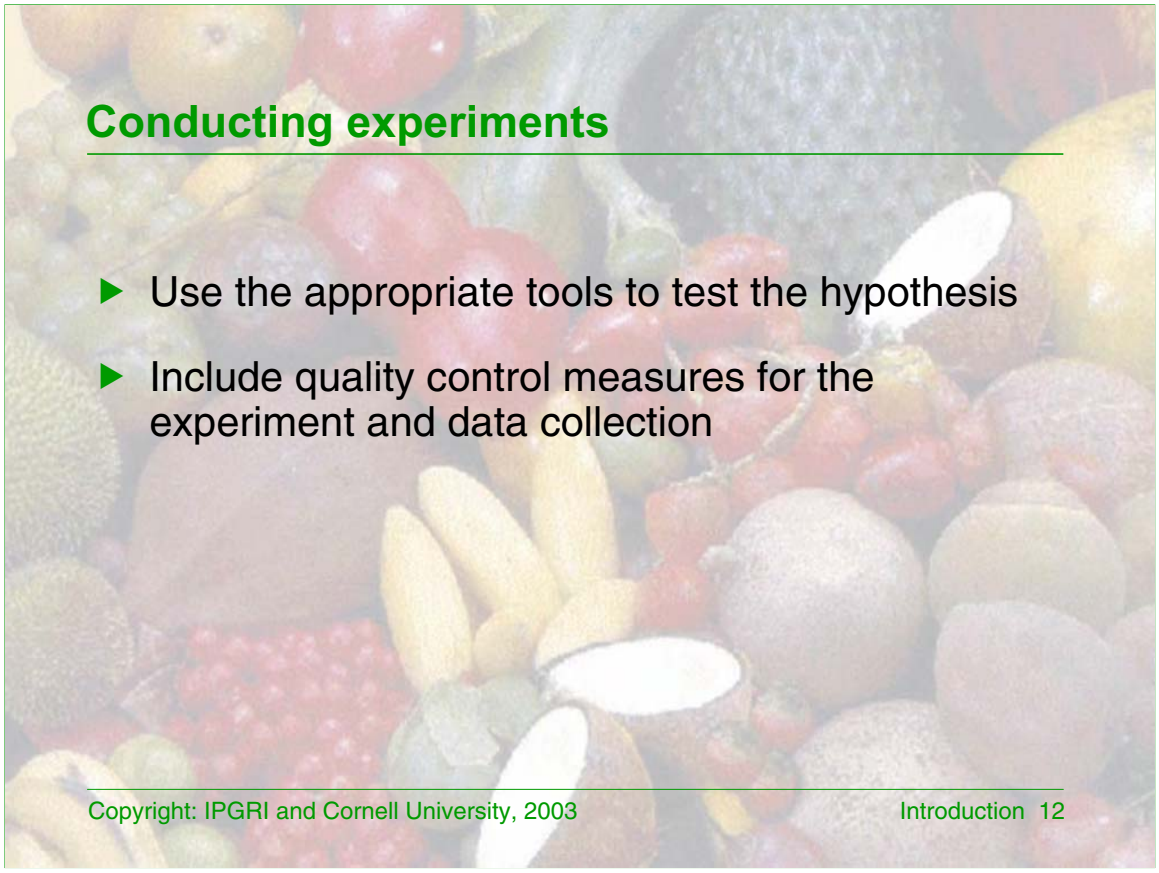
Initial population

Frequency of A_1 (p) = 25/96 = 0.260
Frequency of A_2 (q) = 68/96 = 0.708
Frequency of A_3 (s) = 3/96 = 0.031

Sample

Frequency of A_1 (p) = 8/36 = 0.222
Frequency of A_2 (q) = 28/36 = 0.778
Frequency of A_3 (s) = 0/36 = 0.000 (lost)

The example above shows a situation in which a subsample is taken from a population in such a way that allele A_3, 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.



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).

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.

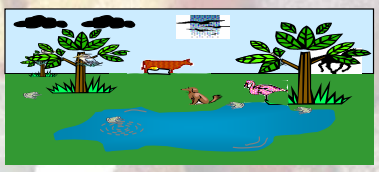
At this point, different analytical tools, as well as different software, may be available. These will be discussed in the last two sections of this module. Simultaneously, we need to be familiar with the basic scientific concepts behind genetic diversity and population genetics (which is that part of 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.

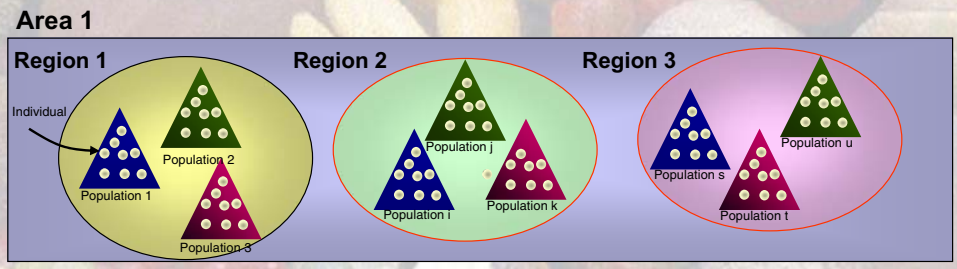
'Genetic diversity' refers to the variation of genes within species, that is to the mix of genes contained within individuals. This covers distinct populations of the same species or genetic variation within a population. Ultimately, genetic diversity resides in changes in the sequence of the four base-pairs 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)

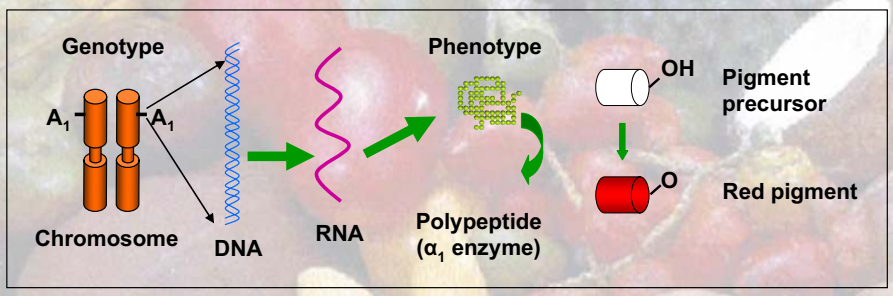


Genetic differences are found between individuals within a population, and differences are also found in allelic frequencies between populations. In all, the relative amount of variation depends on the species, history and environment.

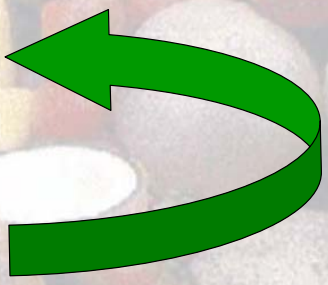
The presence of variants (polymorphism) in a sample may be assessed by its genotypes, alleles, haplotypes or nucleotides. Samples can be hierarchically divided at the level of the species, population or within the population. Sampling size should be sufficient when the goal is to study genetic variation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



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



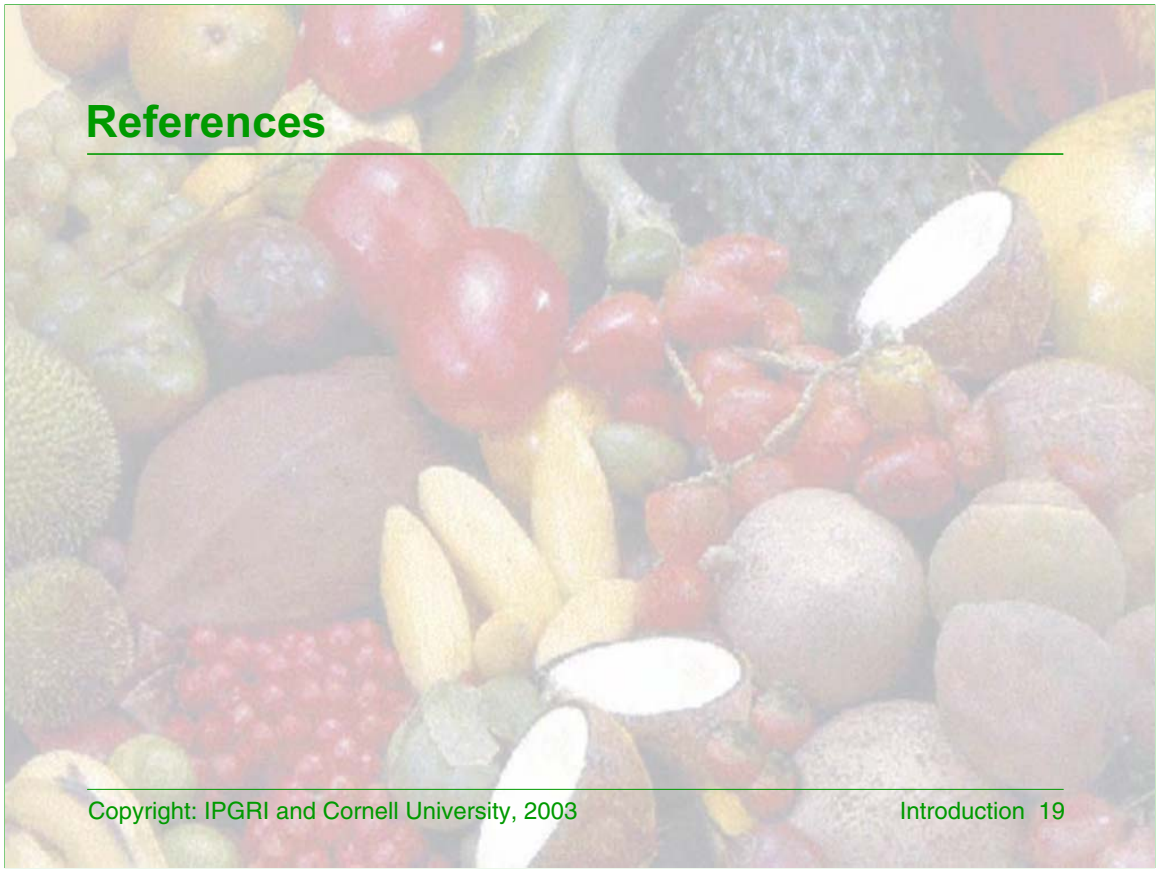
Analysing genetic variation with molecular technologies gives information at the DNA level. It can be neutral diversity, identified along the DNA sequence in regions whose function is unknown, such as when we use anonymous types of markers (e.g. AFLPs, RAPDs). Or, the diversity can be based on known 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

Brown, A.H.D. and B.S. Weir. 1983. Measuring genetic variability in plant populations. Pp. 219-239 *in* Isozymes in Plant Genetics and Breeding, part A (S.D. Tanksley and T.J. Orton, eds.). Elsevier Science Publishers, Amsterdam.

de Vicente, M.C. and T. Fulton. 2003. Using Molecular Marker Technology in Studies on Plant Genetic Diversity. <www.ipgri.cgiar.org/publications/pubfile.asp?ID_PUB=912>

Gregorius, H.R. 1980. The probability of losing an allele when diploid genotypes are sampled. Biometrics 36:643-652.

Mariette, S., V. Lecorre and A. Kremer. 1999. Sampling within the genome for measuring within-population diversity: Trade-offs between markers *in* Sampling Strategies for Marker Analysis. A compendium of the research project *Development, Optimization and Validation of Molecular Tools for the Assessment of Biodiversity in Forest Trees* as part of the European Union DGXII Biotechnology FW IV Research Programme *Molecular Tools for Biodiversity*. (E.M. Gillet, ed.). <http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm>

Next

- Basic concepts of population genetics
- Measures of genetic diversity
- Software programs for analysing genetic diversity
- Glossary