IPGRI TECHNICAL BULLETIN NO. 10

IPGRI

Molecular markers

for genebank management

D. Spooner, R. van Treuren and M. C. de Vicente

Agricultural Research Service

USDA

WAGENINGEN UR

IPGRI Technical Bulletins are published by the International Plant Genetic Resources Institute with the intention of putting forward definitive recommendations for techniques in genetic resources. They are specifically aimed at National Programme and genebank personnel.

Previous titles in this series:

A protocol to determine seed storage behaviour
T.D. Hong and R.H. Ellis
IPGRI Technical Bulletin No. 1, 1996.

Molecular tools in plant genetic resources conservation: a guide to the technologies
A. Karp, S. Kresovich, K.V. Bhat, W.G. Ayad and T. Hodgkin
IPGRI Technical Bulletin No. 2, 1997.

Core collections of plant genetic resources
Th.J.L. van Hintum, A.H.D. Brown, C. Spillane and T. Hodgkin
IPGRI Technical Bulletin No. 3, 2000.

Design and analysis of evaluation trials of genetic resources collections
Statistical Services Centre and University of Reading
IPGRI Technical Bulletin No. 4, 2001.

Accession management: combining or splitting accessions as a tool to improve germplasm management efficiency
N.R. Sackville Hamilton, J. M.M. Engels, Th.J.L. van Hintum, B. Koo and M. Smale
IPGRI Technical Bulletin No. 5, 2002.

Forest tree seed health
J.R. Sutherland, M. Diekmann and P. Berjak
IPGRI Technical Bulletin No. 6, 2002.

***In vitro* collecting techniques for germplasm conservation**
V.C. Pence, J.A. Sandoval, V.M. Villalobos A. and F. Engelmann
IPGRI Technical Bulletin No. 7, 2002.

Análisis estadístico de datos de caracterización morfológica
T.L. Franco y R. Hidalgo
IPGRI Technical Bulletin No. 8, 2002.

A methodological model for ecogeographic surveys of crops
L. Guarino, N. Maxted and E.A. Chiwona, editors
IPGRI Technical Bulletin No. 9, 2005.

Copies can be obtained in PDF format from IPGRI's Website (www.ipgri.cgiar.org) or in printed format by sending a request to ipgri-publications@cgiar.org.

Molecular markers
for genebank management

D. Spooner[1], R. van Treuren[2] and M. C. de Vicente[3]

[1] USDA, Agricultural Research Service
Department of Horticulture
University of Wisconsin
1575 Linden Drive
Madison, Wisconsin
53706-1590
USA

[2] Centre for Genetic Resources,
The Netherlands (CGN)
Wageningen University and Research Centre
P.O. Box 16
6700 AA
Wageningen
The Netherlands

[3] International Plant Genetic Resources Institute
IPGRI Office for the Americas A.A.
6713 Cali
Colombia

Introduction to the Series

The Technical Bulletin series is targeted at scientists and technicians managing genetic resources collections. Each title will aim to provide guidance on choices while implementing conservation techniques and procedures and in the experimentation required to adapt these to local operating conditions and target species. Techniques are discussed and, where relevant, options presented and suggestions made for experiments. The Technical Bulletins are authored by scientists working in the genetic resources area. IPGRI welcomes suggestions of topics for future volumes. In addition, IPGRI would encourage, and is prepared to support, the exchange of research findings obtained at the various genebanks and laboratories.

The International Plant Genetic Resources Institute (IPGRI) is an independent international scientific organization that seeks to improve the well-being of present and future generations of people by enhancing conservation and the deployment of agricultural biodiversity on farms and in forests. It is one of 15 Future Harvest Centres supported by the Consultative Group on International Agricultural Research (CGIAR), an association of public and private members who support efforts to mobilize cutting-edge science to reduce hunger and poverty, improve human nutrition and health, and protect the environment. IPGRI has its headquarters in Maccarese, near Rome, Italy, with offices in more than 20 other countries worldwide. The Institute operates through four programmes: Diversity for Livelihoods, Understanding and Managing Biodiversity, Global Partnerships, and Improving Livelihoods in Commodity-based Systems.

The Agricultural Research Service (ARS) is the US Department of Agriculture's chief scientific research agency. Its mission is to find solutions to agricultural problems that affect consumers every day, from field to table. The agency conducts research to develop and transfer solutions to agricultural problems of high national priority and provide information access and dissemination.

The Centre for Genetic Resources, the Netherlands (CGN) is part of Wageningen University and Research Centre. Under a mandate of the Netherlands government, CGN is responsible for research tasks that relate to biodiversity and identity of species of importance to agriculture and forestry. CGN carries out the co-ordination of the governmental programme aimed at conservation and utilization of genetic resources. CGN's mission is to contribute to global conservation efforts.

Citation: Spooner D., R. van Treuren and M.C. de Vicente. 2005. Molecular markers for genebank management. IPGRI Technical Bulletin No. 10. International Plant Genetic Resources Institute, Rome, Italy.

Cover Image: Flowers of the wild tomato species *Solanum arcanum* Peralta (photo: D. Spooner) and part of a polyacrylamide sequencing gel S^{35} radio-labelling (photo: R. van Treuren). Design: P. Tazza.

ISBN-10: 92-9043-684-0
ISBN-13: 978-92-9043-684-3

IPGRI
Via dei Tre Denari 472/a
00057 Maccarese - Rome, Italy

Table of Contents

List of figures

List of tables

Introduction

In the last decade, the use of DNA markers for the study of crop genetic diversity has become routine, and has revolutionized biology. Increasingly, techniques are being developed to more precisely, quickly and cheaply assess genetic variation. These techniques have changed the standard equipment of many labs, and most germplasm scientists are expected to be trained in DNA data generation and interpretation. The rapid growth of new techniques has stimulated this update of IPGRI's Technical Bulletin No. 2, "Molecular tools in plant genetic resources conservation: a guide to the technologies" (Karp et al. 1997b). Our goal is to update DNA techniques from this publication, to show examples of their applications, and to guide genebank researchers towards ways to maximize their use. This bulletin reviews basic qualities of molecular markers, their characteristics, the advantages and disadvantages of their applications, and analytical techniques, and provides some examples of their use.

There is no single molecular approach for many of the problems facing genebank managers, and many techniques complement each other. However, some techniques are clearly more appropriate than others for some specific applications. In an ideal situation, the most appropriate marker(s) can be chosen irrespective of time or funding constraints, but in other cases the choice of marker(s) will depend on constraints of equipment or funds. The purpose of this publication is to explain the characteristics of different markers and guide to their use through a number of real examples that represent well-informed choices. What is most important is to choose a marker that can appropriately address well-defined questions through good experimental design, ideally leading to peer-reviewed scientific publications.

Experimental design has many definitions depending on the type of question being asked and on the field of science addressed. We use the term here in a very general way to cover all aspects of planning an experiment, including a clear definition of the question being addressed; knowledge of prior studies addressing the question; proper choice of molecular markers and of data used to address the question; knowledge of the characteristics, strengths and weaknesses of the data; sources of unexpected variation in the data; how much data are needed; proper methods to analyze the data; and limits to conclusions you can make from the results.

One of the most important considerations before beginning any experiment is to address proper experimental design. Improper

experimental design can make the work inconclusive, misleading, insignificant, and most likely unpublishable. Similarly, improvements in experimental design can change an uninspired study to a highly significant one with little to no increase in time and funds. Poor experimental design can waste significant resources and damage the reputation and impact of your genebank.

It is beyond the scope of any publication to outline all possible pitfalls that can lead to poorly designed experiments, analyses or conclusions, and different considerations of proper experimental design need to be made in particular fields. This technical bulletin outlines some basic considerations regarding molecular marker types and analyses to lead the reader. There is no substitute, however, for basic knowledge of the biological questions being addressed, knowledge of the taxonomic group under consideration and a thorough literature review to ensure that similar work has not been done before. If limitations of any type hinder genebank and germplasm managers with regards to these factors, collaboration or consultation with experts is well worth the effort.

Excellent reviews of methodology and data interpretation are presented in Weising et al. (1995), Hillis et al. (1996), Staub et al. (1996), Hillis (1997), Karp et al. (1997a,b) and Avise (2004). Hamrick and Godt (1997) present a review of isozyme data; Doebley (1992), Clegg (1993b) and Spooner and Lara-Cabrera (2001) present a review of molecular data for plant genetic resources and crop evolution; Bruford and Wayne (1993), Wang et al. (1994), Gupta et al. (1996), Powell et al. (1996a) and Weising et al. (1998) of microsatellite data; Wolfe and Liston (1998) on Polymerase Chain Reaction (PCR) related data. Schlötterer (2004) reviews the history and relative utility of different molecular marker types. Sytsma and Hahn (1997) present reviews of molecular studies in crop and non-crop plants. Some information from Spooner and Lara-Cabrera (2001) for crop diversity studies was used and updated; Spooner et al. (2003) was used for taxonomy studies.

An overview of the main marker techniques and their comparative qualities is presented in the section titled, "Overview of molecular technologies". Applications of molecular techniques in genebank management and crop breeding are the subject of the following sections. The section titled, "Future challenges" focuses on the current developments in molecular marker applications and future challenges that could result from these developments. Elements of experimental design are discussed throughout and some basic aspects of data analysis are discussed in "Genebank management".

Overview of molecular technologies

Due to the rapid developments in the field of molecular genetics, a variety of different techniques have emerged to analyze genetic variation during the last few decades (Whitkus et al. 1994; Karp et al. 1996, 1997a,b; Parker et al. 1998; Schlötterer 2004). These genetic markers may differ with respect to important features, such as genomic abundance, level of polymorphism detected, locus specificity, reproducibility, technical requirements and financial investment. No marker is superior to all others for a wide range of applications. The most appropriate genetic marker will depend on the specific application, the presumed level of polymorphism, the presence of sufficient technical facilities and know-how, time constraints and financial limitations. The main marker technologies that have been widely applied during the last decades are summarized in Table 1, and briefly outlined below, together with their strengths and weaknesses. Information about the technologies and their applications may also be accessed via the website of the Centre for Genetic Resources, The Netherlands (CGN) at http://www.cgn.wur.nl/pgr/.

Main marker technologies

Allozymes

Description: Allozymes are allelic variants of enzymes encoded by structural genes. Enzymes are proteins consisting of amino acids, some of which are electrically charged. As a result, enzymes have a net electric charge, depending on the stretch of amino acids comprising the protein. When a mutation in the DNA results in an amino acid being replaced, the net electric charge of the protein may be modified, and the overall shape (conformation) of the molecule can change. Because changes in electric charge and conformation can affect the migration rate of proteins in an electric field, allelic variation can be detected by gel electrophoresis and subsequent enzyme-specific stains that contain substrate for the enzyme, cofactors and an oxidized salt (e.g. nitro-blue tetrazolium). Usually two, or sometimes even more loci can be distinguished for an enzyme and these are termed isoloci. Therefore, allozyme variation is often also referred to as isozyme variation (Kephart 1990; May 1992).

Strengths: The strength of allozymes is simplicity. Because allozyme analysis does not require DNA extraction or the availability of sequence information, primers or probes, they are quick and easy to use. Some species, however, can require considerable optimization of techniques for certain enzymes. Simple analytical

procedures, allow some allozymes to be applied at relatively low costs, depending on the enzyme staining reagents used. Allozymes are codominant markers that have high reproducibility. Zymograms (the banding pattern of isozymes) can be readily interpreted in terms of loci and alleles, or they may require segregation analysis of progeny of known parental crosses for interpretation. Sometimes, however, zymograms present complex banding profiles arising from polyploidy or duplicated genes and the formation of intergenic heterodimers, which may complicate interpretation.

Weaknesses: The main weakness of allozymes is their relatively low abundance and low level of polymorphism. Moreover, proteins with identical electrophoretic mobility (co-migration) may not be homologous for distantly related germplasm. In addition, their selective neutrality may be in question (Berry and Kreitman 1993; Hudson et al. 1994; Krieger and Ross 2002). Lastly, often allozymes are considered molecular markers since they represent enzyme variants, and enzymes are molecules. However, allozymes are in fact phenotypic markers, and as such they may be affected by environmental conditions. For example, the banding profile obtained for a particular allozyme marker may change depending on the type of tissue used for the analysis (e.g. root vs. leaf). This is because a gene that is being expressed in one tissue might not be expressed in other tissues. On the contrary, molecular markers, because they are based on differences in the DNA sequence, are not environmentally influenced, which means that the same banding profiles can be expected at all times for the same genotype.

Applications: Allozymes have been applied in many population genetics studies, including measurements of outcrossing rates (Erskine and Muehlenbauer 1991), (sub)population structure and population divergence (Freville et al. 2001). Allozymes are particularly useful at the level of conspecific populations and closely related species, and are therefore useful to study diversity in crops and their relatives (Hamrick and Godt 1997). They have been used, often in concert with other markers, for fingerprinting purposes (Tao and Sugiura 1987; Maass and Ocampo 1995), and diversity studies (Lamboy et al. 1994; Ronning and Schnell 1994), to study interspecific relationships (Garvin and Weeden 1994), the mode of genetic inheritance (Warnke et al. 1998), and allelic frequencies in germplasm collections over serial increase cycles in germplasm banks (Reedy et al. 1995), and to identify parents in hybrids (Parani et al. 1997).

Table 1. Classification of marker techniques for relatively closely related germplasm

A. Biochemical markers

- Allozymes (Tanksley and Orton 1983; Kephart 1990; May 1992)

B. Molecular markers[1]

i) **Non-PCR**[2] **based techniques**

- Restriction Fragment Length Polymorphisms (**RFLP,** Botstein et al. 1980; Neale and Williams 1991)
- Minisatellites or Variable Number of Tandem Repeats (**VNTR,** Jeffreys et al. 1985a,b)

ii) **PCR-based techniques**

- DNA sequencing
  - Multi-copy DNA, Internal Transcribed Spacer regions of nuclear ribosomal genes (**ITS,** Takaiwa et al. 1985; Dillon et al. 2001)
  - Single-copy DNA, including both introns and exons (Sanger et al. 1977; Clegg 1993a)
- Sequence-Tagged Sites (**STS**)
  - Microsatellites, Simple Sequence Repeat (**SSR**), Short Tandem Repeat (**STR**), Sequence Tagged Microsatellite (**STMS**) or Simple Sequence Length Polymorphism (**SSLP**) (Hearne et al. 1992; Morgante and Olivieri 1993; Queller et al. 1993; Jarne and Lagoda 1996)
  - Amplified Sequence Length Polymorphism (**ASLP,** Maughan et al. 1995)
  - Sequence Characterized Amplified Region (**SCAR,** Paran and Michelmore 1993)
  - Cleaved Amplified Polymorphic Sequence (**CAPS,** Akopyanz et al. 1992; Konieczny and Ausubel 1993)
  - Single-Strand Conformation Polymorphism (**SSCP,** Hayashi 1992)
  - Denaturing Gradient Gel Electrophoresis (**DGGE**, Riedel et al. 1990)
  - Thermal Gradient Gel Electrophoresis (**TGGE,** Riesner et al. 1989)
  - Heteroduplex Analysis (**HDA,** Perez et al. 1999; Schneider et al. 1999)
  - Denaturing High Performance Liquid Chromatography (**DHPLC,** Hauser *et a*l. 1998; Steinmetz et al. 2000; Kota et al. 2001)
- Multiple Arbitrary Amplicon Profiling (**MAAP,** Caetano-Anolles 1996; Caetano Anolles et al. 1992)
  - Random Amplified Polymorphic DNA (**RAPD,** Williams et al. 1990; Hadrys et al. 1992)
  - DNA Amplification Fingerprinting (**DAF,** Caetano-Anolles et al. 1991)
  - Arbitrarily Primed Polymerase Chain Reaction (AP-PCR, Welsh and McClelland 1990; Williams et al. 1990)
  - Inter-Simple Sequence Repeat (ISSR, Zietkiewicz et al. 1994; Godwin et al. 1997)
  - Single Primer Amplification Reaction (SPAR, Staub et al. 1996)
  - Directed Amplification of Minisatellite DNA (DAMD, Heath et al. 1993; Somers and Demmon 2002)
  - Amplified Fragment Length Polymorphism (AFLP, Vos et al. 1995)
  - Selectively Amplified Microsatellite Polymorphic Loci (SAMPL, Witsenboer et al. 1997)

[1] Molecular markers can be based on cytoplasmic DNA (chloroplast, cpDNA, and/or mitochondrion, mtDNA) or nuclear DNA.

[2] PCR is an abbreviation of Polymerase Chain Reaction, a technology that amplifies DNA fragments.

Restriction Fragment Length Polymorphism (RFLP)

Description: RFLPs are bands that correspond to DNA fragments, usually within the range of 2–10 kb, that have resulted from the digestion of genomic DNA with restriction enzymes. DNA fragments are separated by agarose gel electrophoresis and are detected by subsequent Southern blot hybridization to a labelled DNA probe. Labelling of the probe may be performed with a radioactive isotope or with alternative non-radioactive stains, such as digoxigenin or fluorescein. The locus specific RFLP probes consist of a homologous sequence of a specific chromosomal region. Probes are generated through the construction of genomic or complementary DNA (cDNA) libraries and therefore may be composed of a specific sequence of unknown identity (genomic DNA) or part of the sequence of a functional gene (exons only, cDNA). RFLP probes are maintained as clones in suitable bacterial vectors that conveniently allow the isolation of the DNA fragments they hold. Probes from related species may be used (heterologous probes). DNA sequence variation affecting the absence or presence of recognition sites of restriction enzymes, and insertions and deletions within two adjacent restriction sites, form the basis of length polymorphisms.

Strengths: RFLPs are generally found to be moderately polymorphic. In addition to their high genomic abundance and their random distribution, RFLPs have the advantages of showing codominant alleles and having high reproducibility.

Weaknesses: The main drawbacks of RFLPs are the requirement of laborious and technically demanding methodological procedures, and high expense. In general, if research is conducted with poorly studied groups of wild species or poorly studied crops (orphan crops) suitable probes may not yet be available, so considerable investments are needed for development. Moreover, large quantities (1–10 µg) of purified, high molecular weight DNA are required for each DNA digestion. Larger quantities are needed for species with larger genomes, and for the greater number of times needed to probe each blot. RFLPs are not amenable to automation and collaboration among research teams requires distribution of probes.

Applications: RFLPs can be applied in diversity and phylogenetic studies ranging from individuals within populations or species, to closely related species. RFLPs have been widely used in gene mapping studies because of their high genomic abundance due to the ample availability of different restriction enzymes and random distribution throughout the genome (Neale and Williams 1991). They also have been used to investigate relationships of closely related taxa (Miller and Tanksley 1990; Lanner et al. 1997), as fingerprinting tools (Fang et al. 1997), for diversity studies (Debreuil et al. 1996),

and for studies of hybridization and introgression, including studies of gene flow between crops and weeds (Brubaker and Wendel 1994; Clausen and Spooner 1998; Desplanque et al. 1999).

Minisatellites

Description: Minisatellite analysis, like RFLPs, also involves digestion of genomic DNA with restriction endonucleases, but minisatellites are a conceptually very different class of marker. They consist of chromosomal regions containing tandem repeat units of a 10–50 base motif, flanked by conserved DNA restriction sites. A minisatellite profile consisting of many bands, usually within a 4–20 kb size range, is generated by using common multilocus probes that are able to hybridize to minisatellite sequences in different species. Locus specific probes can be developed by molecular cloning of DNA restriction fragments, subsequent screening with a multilocus minisatellite probe and isolation of specific fragments. Variation in the number of repeat units, due to unequal crossing over or gene conversion, is considered to be the main cause of length polymorphisms. Due to the high mutation rate of minisatellites, the level of polymorphism is substantial, generally resulting in unique multilocus profiles for different individuals within a population. Minisatellite loci are also often referred to as Variable Number of Tandem Repeats (VNTR) loci.

Strengths: The main advantages of minisatellites are their high level of polymorphism and high reproducibility.

Weaknesses: Disadvantages of minisatellites are similar to RFLPs due to the high similarity in methodological procedures. If multilocus probes are used, highly informative profiles are generally observed due to the generation of many informative bands per reaction. In that case, band profiles can not be interpreted in terms of loci and alleles and similar sized fragments may be non-homologous. In addition, the random distribution of minisatellites across the genome has been questioned (Schlötterer 2004).

Applications: The term DNA fingerprinting was introduced for minisatellites, though DNA fingerprinting is now used in a more general way to refer to a DNA-based assay to uniquely identify individuals. Minisatellites are particularly useful in studies involving genetic identity, parentage, clonal growth and structure, and identification of varieties and cultivars (Jeffreys et al. 1985a,b; Zhou et al. 1997), and for population-level studies (Wolff et al. 1994). Minisatellites are of reduced value for taxonomic studies because of hypervariability.

Polymerase Chain Reaction (PCR)-sequencing

Description: PCR was a major breakthrough for molecular markers in that for the first time, any genomic region could be amplified and analyzed in many individuals without the requirement for cloning and isolating large amounts of ultra-pure genomic DNA (Schlötterer 2004). PCR sequencing involves determination of the nucleotide sequence within a DNA fragment amplified by the PCR, using primers specific for a particular genomic site. The method that has been most commonly used to determine nucleotide sequences is based on the termination of in vitro DNA replication. The procedure is initiated by annealing a primer to the amplified DNA fragment, followed by dividing the mixture into four subsamples. Subsequently, DNA is replicated in vitro by adding the four deoxynucleotides (adenine, cytocine, guanine, thymidine; dA, dC, dG and dT), a single dideoxynucleotide (ddA, ddC, ddG or ddT) and the enzyme DNA polymerase to each reaction. Sequence extension occurs as long as deoxynucleotides are incorporated in the newly synthesized DNA strand.

However, when a dideoxynucleotide is incorporated, DNA replication is terminated. Because each reaction contains many DNA molecules and incorporation of dideoxynucleotides occurs at random, each of the four subsamples contains fragments of varying length terminated at any occurrence of the particular dideoxy base used in the subsample. Finally, the fragments in each of the four subsamples are separated by gel electrophoresis (Figure 1).

Strengths: Because all possible sequence differences within the amplified fragment can be resolved between individuals, PCR sequencing provides the ultimate measurement of genetic variation. Universal primer pairs to target specific sequences in a wide range of species are available for the chloroplast, mitochondrial and ribosomal genomes. Advantages of PCR sequencing include its high reproducibility and the fact that sequences of known identity are studied, increasing the chance of detecting truly homologous differences. Due to the amplification of fragments by PCR only low quantities of template DNA (the "target"° DNA used for the initial reaction) are required, e.g. 10–100 ng per reaction. Moreover, most of the technical procedures are amenable to automation. Much of this process is now automated (see "Developments in detection techniques".

Weaknesses: Disadvantages include low genome coverage and low levels of variation below the species level. In the event that primers for a genomic region of interest are unavailable, high development costs are involved. If sequences are visualized by polyacrylamide gel electrophoresis and autoradiography, analytical

procedures are laborious and technically demanding. Fluorescent detection systems and reliable analytical software to score base pairs using automated sequencers are now widely applied. This requires considerable investments for equipment or substantial costs in the case of outsourcing. Because sequencing is costly and time-consuming, most studies have focused on only one or a few loci (but see Rokas et al. 2003 in "Genebank management" below). This restricts genome coverage and together with the fact that different genes may evolve at different rates, the extent to which the estimated gene diversity reflects overall genetic diversity is yet to be determined.

Applications: In general, insufficient nucleotide variation is detected below the species level, and PCR sequencing is most useful to address questions of interspecific and intergeneric relationships (Sanger et al. 1977; Clegg 1993a). Until recently, chloroplast DNA and nuclear ribosomal DNA have provided the major datasets for phylogenetic inference because of the ease of obtaining data due to high copy number. Recently, single- to low-copy nuclear DNA markers (here referred to simply as low-copy) have been developed as powerful new tools for phylogenetic analyses (Mort and Crawford 2004; Small et al. 2004). Low-copy nuclear markers generally circumvent problems of uniparental inheritance frequently found in plastid markers (Corriveau and Coleman 1988), and concerted evolution found in nuclear ribosomal DNA (Arnheim 1983) that limits their utility and reliability in phylogenetic studies (Bailey et al. 2003). In addition to biparental inheritance, low-copy nuclear markers exhibit higher rates of evolution (particularly in intron regions) than cpDNA and nrDNA markers (Wolfe et al. 1987; Small et al. 2004) making them useful for closely related species. Yet another advantage is that low-copy sequences generally evolve independently of paralogous sequences and tend to be stable in position and copy number.

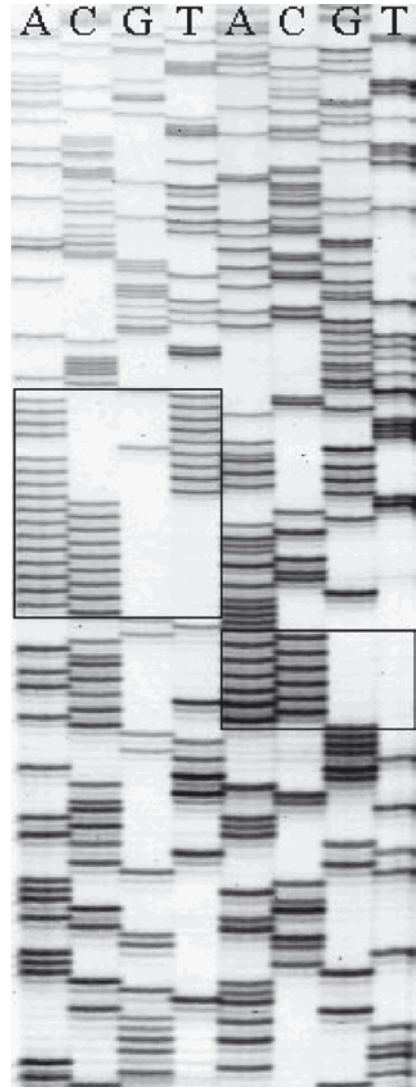


Figure 1. Section of a polyacrylamide sequencing gel using S^{35} radiolabelling. Part of the nucleotide sequence of two cloned DNA fragments containing microsatellite loci (indicated by boxes) is shown. This sequence analysis was part of a project to develop microsatellite markers for the oystercatcher (van Treuren et al. 1999).

Random Amplified Polymorphic DNA (RAPD)

Description: RAPDs are DNA fragments amplified by the PCR using short synthetic primers (generally 10 bp) of random sequence. These oligonucleotides serve as

both forward and reverse primer, and are usually able to amplify fragments from 1–10 genomic sites simultaneously. Amplified fragments, usually within the 0.5–5 kb size range, are separated by agarose gel electrophoresis, and polymorphisms are detected, after ethidium bromide staining, as the presence or absence of bands of particular sizes. These polymorphisms are considered to be primarily due to variation in the primer annealing sites, but they can also be generated by length differences in the amplified sequence between primer annealing sites.

Strengths: The main advantage of RAPDs is that they are quick and easy to assay. Because PCR is involved, only low quantities of template DNA are required, usually 5–50 ng per reaction. Since random primers are commercially available, no sequence data for primer construction are needed. Moreover, RAPDs have a very high genomic abundance and are randomly distributed throughout the genome.

Weaknesses: The main drawback of RAPDs is their low reproducibility (Schierwater and Ender 1993), and hence highly standardized experimental procedures are needed because of their sensitivity to the reaction conditions. RAPD analyses generally require purified, high molecular weight DNA, and precautions are needed to avoid contamination of DNA samples because short random primers are used that are able to amplify DNA fragments in a variety of organisms. Altogether, the inherent problems of reproducibility make RAPDs unsuitable markers for transference or comparison of results among research teams working in a similar species and subject. As for most other multilocus techniques, RAPD markers are not locus-specific, band profiles cannot be interpreted in terms of loci and alleles (dominance of markers), and similar sized fragments may not be homologous.

Applications: RAPDs have been used for many purposes, ranging from studies at the individual level (e.g. genetic identity) to studies involving closely related species. RAPDs have also been applied in gene mapping studies to fill gaps not covered by other markers (Williams et al. 1990; Hadrys et al. 1992). Variants of the RAPD technique include Arbitrarily Primed Polymerase Chain Reaction (AP-PCR) which uses longer arbitrary primers than RAPDs, and DNA Amplification Fingerprinting (DAF) that uses shorter, 5–8 bp primers to generate a larger number of fragments. Multiple Arbitrary Amplicon Profiling (MAAP) is the collective term for techniques using single arbitrary primers.

Microsatellites

Description: Microsatellites, like minisatellites, represent tandem repeats, but their repeat motifs are shorter (1–6 base pairs). If nucleotide sequences in the flanking regions of the microsatellite are known, specific primers (generally 20–25 bp) can be designed to amplify the microsatellite by PCR. Microsatellites and their flanking sequences can be identified by constructing a small-insert genomic library, screening the library with a synthetically labelled oligonucleotide repeat and sequencing the positive clones (Figure 1). Alternatively, microsatellites may be identified by screening sequence databases for microsatellite sequence motifs from which adjacent primers may then be designed. In addition, primers may be used that have already been designed for closely related species. Polymerase slippage during DNA replication, or slipped strand mispairing, is considered to be the main cause of variation in the number of repeat units of a microsatellite, resulting in length polymorphisms that can be detected by gel electrophoresis. Other causes have also been reported (Matsuoka et al. 2002).

Strengths: The strengths of microsatellites include the codominance of alleles, their high genomic abundance in eukaryotes and their random distribution throughout the genome, with preferential association in low-copy regions (Morgante et al. 2002). Because the technique is PCR-based, only low quantities of template DNA (10–100 ng per reaction) are required. Due to the use of long PCR primers, the reproducibility of microsatellites is high and analyses do not require high quality DNA. Although microsatellite analysis is, in principle, a single-locus technique, multiple microsatellites may be multiplexed during PCR or gel electrophoresis if the size ranges of the alleles of different loci do not overlap (Ghislain et al. 2004). This decreases significantly the analytical costs. Furthermore, the screening of microsatellite variation can be automated, if the use of automatic sequencers is an option (Figure 2).

Weaknesses: One of the main drawbacks of microsatellites is that high development costs are involved if adequate primer sequences for the species of interest are unavailable, making them difficult to apply to unstudied groups. Although microsatellites are in principle codominant markers, mutations in the primer annealing sites may result in the occurrence of null alleles (no amplification of the intended PCR product), which may lead to errors in genotype scoring. The potential presence of null alleles increases with the use of microsatellite primers generated from germplasm unrelated to the species used to generate the microsatellite primers (poor "cross-species amplification"). Null alleles may result in a biased estimate of the allelic and genotypic frequencies and an underestimation of heterozygosity. Furthermore, the underlying mutation model of

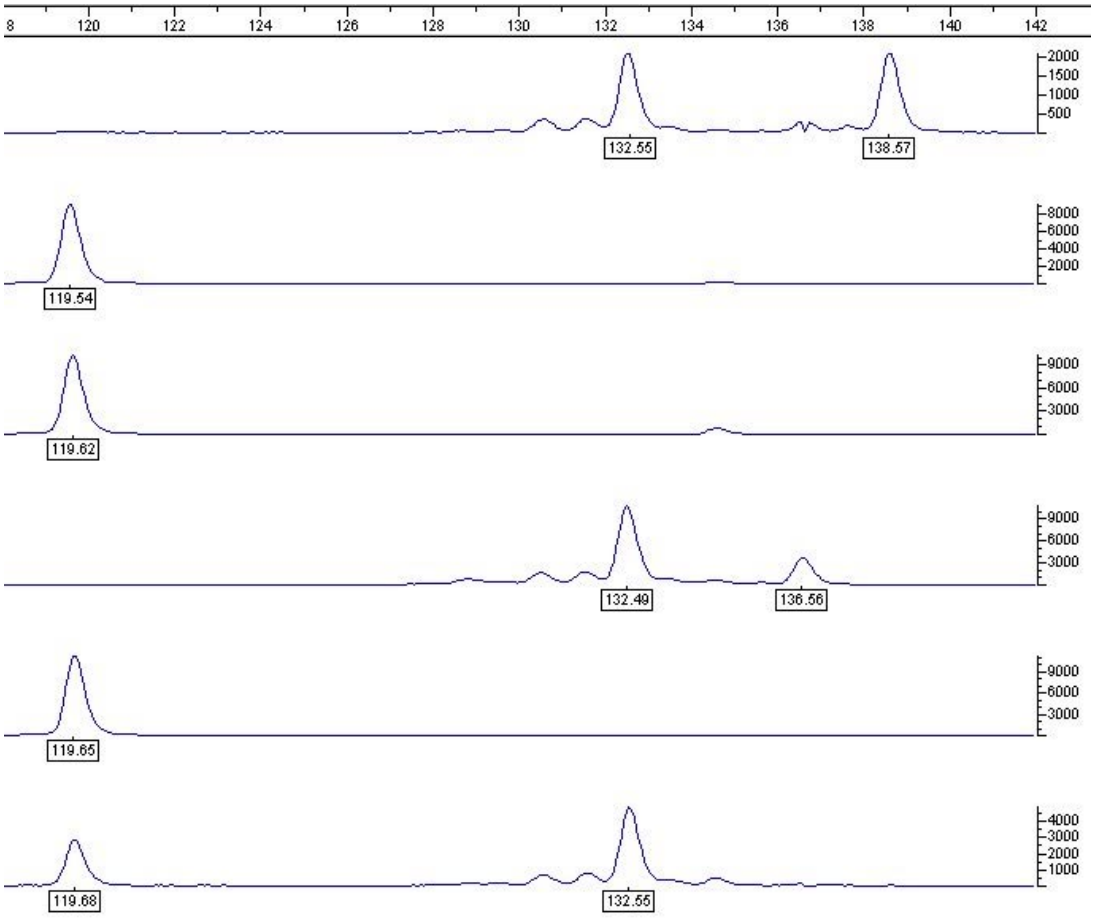


Figure 2. Peak patterns of six perennial ryegrass samples screened for a microsatellite locus using fluorescent labelling on an ABI Prism 3700 DNA analyzer. At the top a size reference in number of base pairs is shown, and at the right the strength of the fluorescent signal. Alleles can be distinguished at respectively 120, 133, 137 and 139 bp. This microsatellite analysis was carried out within the framework of a project to study mating patterns in a regeneration population of perennial ryegrass (van Treuren, unpublished).

microsatellites (infinite allele model or stepwise mutation model) is still under debate. Homoplasy may occur at microsatellite loci due to different forward and backward mutations, which may cause underestimation of genetic divergence. A very common observation in microsatellite analysis is the appearance of stutter bands that are artifacts in the technique that occur by DNA slippage during PCR amplification. These can complicate the interpretation of the band profiles because size determination of the fragments is more difficult

and heterozygotes may be confused with homozygotes. However, the interpretation may be clarified by including appropriate reference genotypes of known band sizes in the experiment.

Applications: In general, microsatellites show a high level of polymorphism. As a consequence, they are very informative markers that can be used for many population genetics studies, ranging from the individual level (e.g. clone and strain identification) to that of closely related species. Conversely, their high mutation rate makes them unsuitable for studies involving higher taxonomic levels. Microsatellites are also considered ideal markers in gene mapping studies (Hearne et al. 1992; Morgante and Olivieri 1993; Queller et al. 1993; Jarne and Lagoda 1996).

Inter Simple Sequence Repeats (ISSR)

Description: ISSRs are DNA fragments of about 100–3000 bp located between adjacent, oppositely oriented microsatellite regions. ISSRs are amplified by PCR using microsatellite core sequences as primers with a few selective nucleotides as anchors into the non-repeat adjacent regions (16–18 bp). About 10–60 fragments from multiple loci are generated simultaneously, separated by gel electrophoresis and scored as the presence or absence of fragments of particular size. Techniques related to ISSR analysis are Single Primer Amplification Reaction (SPAR) that uses a single primer containing only the core motif of a microsatellite, and Directed Amplification of Minisatellite-region DNA (DAMD) that uses a single primer containing only the core motif of a minisatellite.

Strengths: The main advantage of ISSRs is that no sequence data for primer construction are needed. Because the analytical procedures include PCR, only low quantities of template DNA are required (5–50 ng per reaction). Furthermore, ISSRs are randomly distributed throughout the genome.

Weaknesses: Because ISSR is a multilocus technique, disadvantages include the possible non-homology of similar sized fragments. Moreover, ISSRs, like RAPDs, can have reproducibility problems.

Applications: Because of the multilocus fingerprinting profiles obtained, ISSR analysis can be applied in studies involving genetic identity, parentage, clone and strain identification, and taxonomic studies of closely related species. In addition, ISSRs are considered useful in gene mapping studies (Godwin et al. 1997; Zietkiewicz et al. 1994; Gupta et al. 1994).

Single-Strand Conformation Polymorphism (SSCP)

Description: SSCPs are DNA fragments of about 200–800 bp amplified by PCR using specific primers of 20–25 bp. Gel electrophoresis of single-strand DNA is used to detect nucleotide sequence variation among the amplified fragments. The method is based on the fact that the electrophoretic mobility of single-strand DNA depends on the secondary structure (conformation) of the molecule, which is changed significantly with mutation. Thus, SSCP provides a method to detect nucleotide variation among DNA samples without having to perform sequence reactions. In SSCP the amplified DNA is first denatured, and then subject to non-denaturing gel electrophoresis. Related techniques to SSCP are Denaturing Gradient Gel Electrophoresis (DGGE) that uses double stranded DNA which is converted to single stranded DNA in an increasingly denaturing physical environment during gel electrophoresis, and Thermal Gradient Gel Electrophoresis (TGGE) which uses temperature gradients to denature double stranded DNA during electrophoresis.

Strengths: Advantages of SSCP are the codominance of alleles and the low quantities of template DNA required (10–100 ng per reaction) due to the fact that the technique is PCR-based.

Weaknesses: Drawbacks include the need for sequence data to design PCR primers and the necessity of highly standardized electrophoretic conditions in order to obtain reproducible results. Furthermore, some mutations may remain undetected, and hence absence of mutation cannot be proven.

Applications: SSCPs have been used to detect mutations in genes using gene sequence information for primer construction (Hayashi 1992).

Cleaved Amplified Polymorphic Sequence (CAPS)

Description: CAPS are DNA fragments amplified by PCR using specific 20–25 bp primers, followed by digestion of the PCR products with a restriction enzyme. Subsequently, length polymorphisms resulting from variation in the occurrence of restriction sites are identified by gel electrophoresis of the digested products. CAPS have also been referred to as PCR-Restriction Fragment Length Polymorphism (PCR-RFLP).

Strengths: Advantages of CAPS include the involvement of PCR requiring only low quantities of template DNA (50–100 ng per reaction), the codominance of alleles and the high reproducibility. Compared to RFLPs, CAPS analysis does not include the laborious and technically demanding steps of Southern blot hybridization and radioactive detection procedures.

Weaknesses: In comparison with RFLP analysis, CAPS polymorphisms are more difficult to find because of the limited size of the amplified fragments (300–1800 bp). Furthermore, sequence data are needed to design the PCR primers.

Applications: CAPS markers have been applied predominantly in gene mapping studies (Akopyanz et al. 1992; Konieczny and Ausubel 1993).

Sequence Characterized Amplified Region (SCAR)

Description: SCARs are DNA fragments amplified by the PCR using specific 15–30 bp primers, designed from nucleotide sequences established from cloned RAPD fragments linked to a trait of interest. By using longer PCR primers, SCARs do not face the problem of low reproducibility generally encountered with RAPDs. Obtaining a codominant marker may be an additional advantage of converting RAPDs into SCARs, although SCARs may exhibit dominance when one or both primers partially overlap the site of sequence variation. Length polymorphisms are detected by gel electrophoresis.

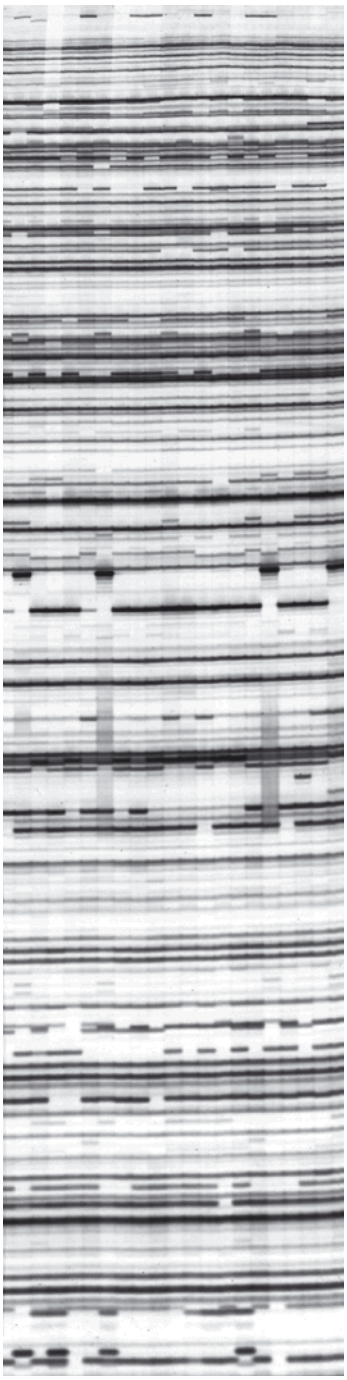
Strengths: The main advantage of SCARs is that they are quick and easy to use. In addition, SCARs have a high reproducibility and are locus-specific. Due to the use of PCR, only low quantities of template DNA are required (10–100 ng per reaction).

Weaknesses: Disadvantages include the need for sequence data to design the PCR primers.

Applications: SCARs are locus specific and have been applied in gene mapping studies and marker assisted selection (Paran and Michelmore 1993).

Amplified Fragment Length Polymorphism (AFLP)

Description: AFLP is a trademark of KeyGene (Wageningen, The Netherlands). AFLPs are DNA fragments (80–500 bp) obtained from digestion with restriction enzymes, followed by ligation of oligonucleotide adapters to the digestion products and selective amplification by the PCR. AFLPs therefore involve both RFLP and PCR. The PCR primers consist of a core sequence (part of the adapter), and a restriction enzyme specific sequence and 1–5 selective nucleotides (the higher the number of selective nucleotides, the lower the number of bands obtained per profile). The AFLP banding profiles are the result of variations in the restriction sites or in the intervening region. The AFLP technique simultaneously generates fragments from many genomic sites (usually 50–100 fragments per reaction) that are separated by polyacrylamide gel electrophoresis and that are generally scored as dominant markers (Figure 3). Selective Fragment Length Amplification (SFLA) and Selective Restriction Fragment Amplification (SRFA) are synonyms sometimes



used to refer to AFLPs. A variation of the AFLP technique is known as Selectively Amplified Microsatellite Polymorphic Locus (SAMPL). This technology amplifies microsatellite loci by using a single AFLP primer in combination with a primer complementary to compound microsatellite sequences, which do not require prior cloning and characterization.

Strengths: The strengths of AFLPs lie in their high genomic abundance, considerable reproducibility, the generation of many informative bands per reaction, their wide range of applications, and the fact that no sequence data for primer construction are required. AFLPs may not be totally randomly distributed around the genome as clustering in certain genomic regions, such as centromers, has been reported for some crops (Alonso-Blanco et al. 1998; Young et al. 1999; Saal and Wricke 2002). AFLPs can be analyzed on automatic sequencers, but software problems concerning the scoring of AFLPs are encountered on some systems.

Weaknesses: Disadvantages include the need for purified, high molecular weight DNA, the dominance of alleles, and the possible non-homology of comigrating fragments belonging to different loci. In addition, due to the high number and different intensity of bands per primer combination, there is the need to adopt certain strict but subjectively determined criteria for acceptance of bands in the analysis. Special attention should be paid to the fact that AFLP bands are not always independent. For example, in case of an insertion between two restriction sites the amplified DNA fragment results in increased band size. This will be interpreted as the loss of a small band and at the same time as the gain of a larger band. This is important for the analysis of genetic relatedness, because it would enhance the weight of non-independent bands compared to the other bands.

Applications: Because of the highly informative fingerprinting profiles generally obtained, AFLPs can be applied in studies involving genetic identity, parentage and identification of clones and cultivars, and phylogenetic studies of closely related species. Their high genomic abundance and generally random distribution throughout the genome make AFLPs a widely valued technology for gene mapping studies (Vos et al. 1995). SAMPL is considered more applicable to intraspecific than to interspecific studies due to frequent null alleles.

Figure 3. Variation among flax samples in part of an AFLP autoradiogam using P^{33} radiolabelling. This AFLP analysis formed part of a marker-assisted rationalization study in a flax collection (van Treuren et al. 2001).

Comparative qualities of marker techniques

DNA provides many advantages that make it especially attractive in studies of diversity and relationships. These advantages have been reviewed elsewhere (e.g. Crawford 1990). They include: 1) Freedom from environmental and pleiotropic effects. Molecular markers do not exhibit phenotypic plasticity, while morphological and biochemical markers can vary in different environments. DNA characters have a much better chance of providing homologous traits. Most morphological or biochemical markers, in contrast, are under polygenic control, and subject to epistatic control and environmental modification (plasticity); 2) A potentially unlimited number of independent markers are available, unlike morphological or biochemical data; 3) DNA characters can be more easily scored as discrete states of alleles or DNA base pairs, while some morphological, biochemical and field evaluation data must be scored as continuously variable characters that are less amenable to robust analytical methods; 4) Many molecular markers are selectively neutral. These advantages do not imply that other more traditional data used to characterize biodiversity are not valuable. On the contrary, morphological, ecological and other "traditional" data will continue to provide practical and often critical information needed to characterize genetic resources.

Molecular markers differ in many qualities and must therefore be carefully chosen and analyzed differently with their differences in mind. To assist in choosing the appropriate marker technique, an overview of the main properties of the 11 main marker technologies described in "Overview of molecular technologies" will follow. These properties are summarized in Table 2.

Genomic abundance

The number of markers that can be generated is determined mainly by the frequency at which the sites of interest occur within the genome. RFLPs and AFLPs generate abundant markers due to the large number of restriction enzymes available and the frequent occurrence of their recognition sites within genomes. Within eukaryotic genomes, microsatellites have also been found to occur frequently. RAPD markers are even more abundant because numerous random sequences can be used for primer construction. In contrast, the number of allozyme markers is restricted due to the limited number (about 30) of enzyme detection systems available for analysis. To investigate specific genomic regions by PCR sequencing, SSCP, CAPS or SCAR, sequence data of the sites of interest (structural genes mainly) are required for primer construction. Although, in principle, many sites of interest may occur within genomes, the

Table 2. Overview of the relevant characteristics of 11 main marker technologies

	Allozymes	RFLP	Mini-satellites	PCR sequencing	RAPD	Micro-satellites	ISSR	SSCP	CAPS	SCAR	AFLP
Genomic abundance	Low	High	Medium	Low	High	High	Medium-High	Low	Low	Low	High
Level of polymorphism	Low	Medium	High	Low	Medium	High	Medium	Low	Low-Medium	Medium	Medium
Locus-specificity	Yes	Yes	No/Yes	Yes	No	Yes	No	Yes	Yes	Yes	No
Codominance of alleles	Yes	Yes	No/Yes	Yes	No	Yes	No	Yes	Yes	No/Yes	No/Yes
Reproducibility	High	High	High	High	Low	High	Medium-High	Medium	High	High	Medium-High
Labour-intensity	Low	High	High	Low/High	Low	Low	Low	Low-Medium	Low-Medium	Low	Medium
Technical demands	Low	High	High	High	Low	Low-Medium	Low-Medium	Medium	Low	Low	Medium
Operational costs	Low	High	High	High	Low	Low	Low-Medium	Low-Medium	Low	Low	Medium
Development costs	Low	Medium-High	Medium-High	High	Low-Medium	High	Low	High	Medium	Medium	Low
Quantity of DNA required	-	High	High	Low	Low	Low	Low	Low	Low	Low	Medium
Amenability to automation	No	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

proportion of the genome covered by PCR sequencing, SSCP, CAPS and SCAR in studies reported to date is limited. However, this is expected to change due to the wealth of sequence information that is becoming increasingly available for different crops. Genomic abundance is essential to studies where a large fraction of the genome needs to be covered, e.g. for the development of high-density linkage maps in gene mapping studies.

If, in addition to genomic abundance, genome coverage is also sought, caution should be taken in marker selection. While some markers are known to be scattered quite evenly across the genomes, others, such as some AFLP markers, sometimes cluster in certain genomic regions. For example, clustering of AFLP markers has been reported in centromeric regions of *Arabidopsis thaliana* (Alonso-Blanco et al. 1998), soybean (Young et al. 1999) and rye (Saal and Wricke 2002).

Level of polymorphism

The resolving power of genetic markers is determined by the level of polymorphism detected, which is determined by the mutation rate at the genomic sites involved. Variation at allozyme loci is caused by point mutations, which occur at low frequency ($<10^{-6}$ per meiosis). Moreover, only mutations modifying the net electric charge and conformation of proteins can be detected, reducing the resolving power of allozymes. In contrast, mutations at minisatellite and microsatellite loci, mainly due to changes in the number of repeat units of the core sequence, have been estimated to occur at the relatively high frequency of 10^{-3}–10^{-2} and 10^{-5}–10^{-2} per meiosis, respectively (Jarne and Lagoda 1996). The other markers presented in Table 2 generally show intermediate levels of polymorphism, resulting from base substitutions, insertions or deletions which may alter primer annealing sites and recognition sites of restriction enzymes, or change the size of restriction fragments and amplified products. In choosing the appropriate technique, the level of polymorphism detected by the marker needs to be considered in relation to the presumed degree of genetic relatedness within the material to be studied. Higher resolving power is required when samples are more closely related. For example, analyses within species or among closely related species may call for fast-evolving markers such as microsatellites. However if the objective is to study genetic relatedness at higher taxonomic levels (such as congeneric species), AFLPs or RFLPs may be a better choice because co-migrating fast-evolving markers will have less chance of being homologous. A primary guiding principle in marker selection is that more conservative markers (those having slower evolutionary rates) are needed with increasing evolutionary distance and vice-versa.

Locus-specificity

Genetic markers using multilocus probes or primers benefit from the fact that multiple polymorphisms, representing various genomic regions, are generated simultaneously. However, a major drawback is that in general the band profiles cannot be interpreted in terms of loci and alleles, but are scored as the presence or absence of bands of a particular size. As a consequence, similar sized fragments may represent alleles from different loci and not be homologous. Therefore, locus-specific markers should be considered for questions of phylogeny or genetic relatedness. Alternatively, markers for fingerprinting studies rely on differences only, and homology is not a concern. In general, locus-specific markers generate polymorphisms of known identity, however in most cases sequencing data are needed for their development.

Codominance of alleles

Codominant markers are markers for which both alleles are expressed when co-occurring in an individual. Therefore, with codominant markers, heterozygotes can be distinguished from homozygotes, allowing the determination of genotypes and allele frequencies at loci. In contrast, band profiles of dominant markers are scored as the presence or absence of fragments of a particular size, and heterozygosity cannot be determined directly. As a consequence, only an approximation of allele frequency can be obtained by assuming Hardy-Weinberg equilibrium in a population and estimating allele frequency from the proportion of individuals with the absent phenotype (homozygous recessive). For predominantly self-fertilizing species, heterozygosity could be disregarded and allele frequencies be considered equal to observed band frequencies. Codominant markers are preferred for most applications. The majority of codominant markers are single locus markers and hence the degree of information per assay is usually lower compared to the multilocus techniques.

Reproducibility

Reproducibility is always an important property of markers, but even more important with collaborative projects, involving the generation of data by different labs whose results need to be assembled. To obtain reproducible results, the extraction of purified, high quality DNA is a prerequisite for the majority of the marker techniques. For example, degraded and/or unpurified DNA may affect the amplification or restriction of DNA, resulting in unspecific polymorphisms. Even when purified and high molecular weight DNA is used, RAPDs often fail to show reproducible results. This

is because RAPD primers are very short (10 bp), which can result in alterations in their annealing behaviour to the template DNA and the resulting band profiles as a result of small deviations in experimental conditions. Therefore, highly standardized experimental procedures are required when RAPD markers are being used. This implies the need for including repeated samples and also the inclusion of reference genotypes which represent bands of known size. Problems with reproducibility in RAPD analysis could be overcome by focusing on mapped markers for which their inheritance has already been verified.

Labour-intensity

RFLPs and minisatellites are labour-intensive markers because their analysis includes the time-consuming steps of Southern blotting, labelling of probes and hybridization. Therefore, PCR-based techniques are currently preferred, some of which can even be automated to decrease the labour-intensity. PCR sequencing may still be quite labour-intensive if performed by the old time-consuming method of performing four separate sequence reactions per sample. However, automated procedures have greatly reduced labour-intensity of PCR-sequencing. The labour-intensity of the other PCR-based techniques presented varies from low to medium, depending on the methodological procedures required in addition to PCR (Table 2).

Technical demands

RFLPs, minisatellites and manual PCR sequencing require higher technical skills and facilities for analysis. RFLP and minisatellite analyses require Southern blot hybridizations and may include radioactive labelling. This calls for expertise and exclusive facilities needed to comply with special legal and safety requirements. These technologies are therefore among the most technically demanding markers. Another type of technical demand arises from the use of polyacrylamide gels and automated equipment. Allozymes and PCR-based markers analyzed on agarose gels (e.g. RAPD, SCAR and microsatellites) are the least technically demanding.

Operational costs

Wages, laboratory facilities, technical equipment and consumables all contribute to the operational costs of the technologies. Relatively expensive consumables include Taq-polymerase needed for all PCR-based marker types, restriction enzymes (for RFLPs, minisatellites and CAPS, and particularly the restriction enzyme MseI often used in AFLPs) and isotopes where polymorphisms are visualized by means

of radioactive labelling. Polyacrylamide gels are more expensive to run than agarose gels and require visualization of polymorphisms by autoradiography or silver staining procedures, which are more costly compared to ethidium-bromide staining. Laborious and technically demanding markers, such as RFLPs, minisatellites, PCR sequencing, and those techniques being performed by automated equipment, are quite expensive. Costs of performing RAPD analyses are usually considered low. However, if measures to ensure reproducibility and low numbers of markers per primer are taken into account, costs may increase to the level of the more complex technologies.

In general, operational costs of markers will vary depending on the methodology. Regarding automated procedures and technologies, while purchasing the equipment is usually very expensive and the technical expertise required is high, a significant increase in throughput may be obtained through multiplexing. An additional consideration is the emergence of cost effective "outsourcing" companies to generate marker-based and DNA sequencing data, as service laboratories keep up with efficient equipment developments. Outsourcing allows researchers to concentrate on defining questions, experimental design, data analysis and interpretation. The relative costs/benefits of outsourcing will vary in different labs according to local labour and supply costs, availability of equipment, the benefit of generating your own data for quality control or educational purposes, and the legal requirements to ship crop germplasm DNA out of a country.

Development costs

Marker development may be very time-consuming and costly when suitable probes or sequence data for primer construction are unavailable. Development of suitable probes for Southern blot hybridizations (e.g. for RFLP analysis) requires the construction of either genomic or cDNA libraries and the examination of various probe/restriction enzyme combinations for their ability to detect polymorphisms. The development of site-specific PCR primers (e.g. for microsatellite analysis) also requires the construction of libraries, which then need to be screened to identify the fragments of interest. Subsequently, the identified fragments need to be sequenced to verify their suitability and to design primers. Therefore, the investment required for marker development should be evaluated in relation to the intended range of application of the technique. Alternatively, new genomic tools are allowing probes, primers and sequence data to be obtained from genome databases of other species, with the understanding, as in all DNA tools, that their

usefulness may decrease with increasing evolutionary distance between the species.

Quantity of DNA required

Because only small quantities of template DNA (5–100 ng per reaction) are required, techniques which are based on the PCR are currently preferred. Although RFLPs and minisatellites require the largest amount of DNA (5–10 µg per reaction), Southern blot membranes may be probed several times. Intermediate quantities of DNA are needed for AFLP-analysis (0.3–1 µg per reaction) because restriction of the DNA precedes the PCR reaction. In general, consideration should be given to the use of PCR-based markers if only small amounts of DNA can be obtained.

Amenability to automation

Currently, if adequate equipment and resources are available, techniques that can be automated are highly preferred because of the potential for high sample throughput. Although considerable financial investment is still required, automation may be cost-effective when techniques are applied on a routine basis. As pointed out above, outsourcing of data generation may also be an alternative strategy. Nearly all techniques that are based on the PCR are amenable to a certain degree of automation.

Genebank management

The realization that the world was rapidly losing much of its agrobiodiversity led to a global effort to collect and conserve germplasm. An increasing awareness of the narrow genetic base of crops in advanced agriculture and potential susceptibility to crop failures (National Research Council 1972) further stimulated the efforts to collect, and a system of national and international genebanks eventually amassed holdings of 6.1 million collections in 1300 genebanks worldwide (FAO 1996).

The great success of this collecting phase has presented new challenges to genebank managers to determine needs for new collections, maintain existing collections, determine optimum regeneration methods, characterize collections for useful agronomic traits, classify the collections, and reduce the size of the working collection to a manageable size (the core collection concept; Frankel 1984; Hamon et al. 1995). A major question facing genebank managers concerns which material needs to be included in a collection to conserve a representative sample of the total genetic diversity range of a crop. Schoen and Brown (1993) document how molecular marker techniques provide the optimal strategy to sample materials from populations of wild relatives to maximize the number of alleles. Hamilton (1994) argues that simple marker diversity is an insufficient measure of maintenance of diversity, and suggests that more detailed studies of genetic correlations of quantitative genetic variation and gene and environment interaction are needed.

Another important question is how the genetic diversity present in the original collections is being affected and retained after serial germplasm increase cycles. Germplasm is stored as seeds for the majority of species, and under proper conditions of low temperature and humidity, seeds remain viable for 20 years or more, but other species have seed that rapidly lose viability and, consequently must be regenerated vegetatively. Each time a collection is regenerated from seeds, it has the potential to lose genetic diversity, especially if it is increased under greenhouse or experimental field conditions where it is removed from natural evolutionary forces (Bretting and Duvick 1997). Others question the adequacy of population sizes needed for increase cycles (Brown et al. 1997), proper pollination and seed increase strategies (Crossa and Vencovsky 1994), and avoidance of the accumulation of deleterious mutations within accessions under long-term storage (Schoen et al. 1998).

Acquisition of collection material

Biogeography is the science that attempts to document and understand spatial patterns of biodiversity, and includes the study of historical and climatic effects on plant distributions (Brown and Lomolino 1998). The literature sometimes shows association of DNA-based relationships to geography and sometimes fails to show such associations. One study showing association is that of Whitkus et al. (1998), who investigated the origins of different cultivars of cacao. Wild populations of cacao are widely distributed from the upper Amazonian basin (South American populations) to southern Mexico (Mesoamerican populations). Cultivars from Mesoamerica are termed "criollo", while South American cultivars are termed "forastero". One hypothesis suggested a single origin of cultivars in South America, with human dispersion and later differentiation to Mesoamerica, while the alternative hypothesis was for independent origins in both areas from indigenous wild populations. Whitkus et al. (1998) examined these alternative hypotheses using RAPD markers from a wide variety of geographically diverse wild and cultivated populations. Their results supported a single origin of modern cultivars in South America. However, they also discovered relationships of distinct populations in ancient Mayan groves in southern Mexico to wild Mesoamerican populations. If these ancient Mayan populations are truly remnants of ancient cultivated sites, and if RFLP data are providing a good recapitulation of relationships, a separate origin of cacao in both regions is supported.

Cronn et al. (1997) evaluated isozyme variability of 146 accessions of wild and weedy sunflower and two outgroup species, representing the geographic range of the species throughout much of North America. The great genetic variation in this species has led to a multitude of names. The goals of the study were to quantify the range of genetic diversity, divergence and redundancy in the collection, and to elucidate possible patterns of ecogeographic variation that would clarify interrelationships and place of origin of the domesticated forms. Prior hypotheses suggested domestication in the central United States with later dissemination from this region, or domestication in the southwestern United States. The results showed greater diversity in wild compared to cultivated accessions. There was a geographic association with isozyme diversity, with the greatest diversity from the Great Plains. The domesticated accessions are most similar to those from the Great Plains, suggesting this as the site of domestication. The alternative hypothesis of origin in the southwestern United States cannot be discounted however if early cultivars were introgressed with germplasm from the southwest.

Other studies showing a relationship of molecular variation to geography or ecology are Espejo-Ibañez et al. (1994), Garvin and

Weeden (1994), Yang et al. (1996), Lanner et al. (1997), Nevo et al. (1997, 1998) Fahima et al. (1998), and Zerega (2004). However, there is often not a congruence of genetic distance and geographic distance as shown in Maass and Ocampo (1995), Lanner et al. (1996), Ursla et al. (1997), Varghese et al. (1997), Comes and Abbott (1998), Freville et al. (2001) and del Rio and Bamberg (2002).

Marker studies addressing the distribution of genetic variation within and between populations have been used to guide the acquisition of new material for germplasm conservation. For example, the substantially higher level of RFLP variation observed in self-incompatible, as compared to self-compatible species of tomatoes was used to recommend predominantly sampling the self-incompatible species for germplasm acquisition (Miller and Tanksley 1990). Priority regions for further sampling of sorghum were identified by high diversity levels in some populations estimated from allozymes (Aldrich et al. 1992). Sampling of marginal populations was recommended in order to capture most of the rare and local alleles responsible for differentiation of pawpaw [*Asimina triloba* (L.) Dunal] in the US, based on the distribution of variation for allozme and RAPD markers (Huang et al. 1998, 2000). Steiner et al. (1998) used RAPDs to measure genetic diversity of germplasm collections of *Trifolium incarnatum* L. In combination with data on pedigree, they documented low diversity in the existing cultivars and identified populations that need additional collecting. Lamboy et al. (1996) examined, with six enzyme systems, 291 seedlings of 31 sib families of *Malus sieversii* (Ledeb.) M. Roem., the primary progenitor of the cultivated apple. The families were produced from 14 populations collected in 4 regions in eastern Kazakhstan. They found that there was only a very weak correspondence of allele frequencies to geographic region, that most of the sib families were more closely related to sib families from other regions, and that there were no alleles that were both fixed within and unique to a region. They concluded that populations of this species formed a large panmictic population and that thorough sampling of a few large populations would efficiently capture most of its genetic diversity. Other studies addressing the application of marker data for germplasm acquisition include Brown and Munday (1982), Murphy and Philips (1993), Lamboy et al. (1994), Tsegaye et al. (1996), Nebauer et al. (1999), Maquet et al. (1996, 1997) and Zoro Bi et al. (1998).

Marker data have also been used to identify unique germplasm that may be underrepresented in a genebank. Doebley (1989) examined, by RFLP analysis, the chloroplast DNA of 39 accessions representing the range of maize and its wild relatives in the genus *Zea* L. and the related *Tripsacum dactyloides* (L.) L. and *T. pilosum* Schibn. & Merr. The finding of an atypical cytoplasmic genome, incorporated into the

nuclear background of some *Z. perennis* (Hitchc.) Reeves & Mangelsd. accessions led to the conclusion that introgression had occurred from an unknown *Zea* taxon. Kesseli et al. (1991) compared variation at 143 RFLP loci between 67 accessions of cultivated lettuce and five related species. *Lactuca serriola* L. is closely related to cultivated lettuce, but the RFLP data indicated that unknown populations of *L. serriola* or other unknown entities have been involved in the evolution of cultivated lettuce. In summary, molecular data are useful to guide genebank curators in decisions concerning acquisition in cases where molecular variation is associated with geographic variation.

Taxonomic issues

Genebank researchers often address questions of species delimitations and species interrelationships (taxonomy or systematics). These data can be analyzed in different ways, depending on the taxonomic level of the question and the marker type used. The discussion that follows is intended only as a very basic introduction to fundamental differences of various classes of methods (phenetic and cladistic) to analyze molecular data; the agreement (congruence) or disagreement of results from different molecular markers of the same germplasm sources; the predictive value of taxonomic results; and different ways in which these and other data are used to define species. We guide the reader to Judd et al. (2002) for detailed treatments of molecular systematics and analytical techniques.

Taxonomy, as all branches of science, continues to advance with new sources of data, new ways to analyze these data, new ways to interpret the results of data analyses, and different philosophies on how to interpret the results of a data analysis. This manual will touch on some of these sources of ambiguity or controversy. These include different methods to use in phenetic and cladistic analyses, different ways to interpret the results, how to interpret the different results from the same group of organisms but with different sources of data (congruence of results), and different ways to define species. A discussion of these unresolved ambiguities is not meant to confuse the reader but rather to impart an understanding of different interpretations from different investigators and to help you more intelligently design and interpret your own results.

Phenetic vs. cladistic analyses

Cladistics and phenetics are systematic approaches that share some points in common but have many fundamental differences. Both focus on the explicit examination of many characters, in contrast to earlier intuitive systematic techniques that sometimes used limited characters,

or placed greater importance (weight) on only some characters. Nevertheless, many early and intuitive taxonomic interpretations concur with more recent molecular studies and have stood the test of time. They continue to form the basis for newer approaches and have provided invaluable points of reference for comparison of datasets.

Cladistic and phenetic procedures begin in the same way, by gathering an organism-by-character rectangular data matrix. The organism can be an individual, from which actual data are measured, or a "virtual taxon" (a species or higher rank as genus or family) from which taxon-specific (or near taxon-specific) characters are inferred. The term Operational Taxonomic Unit, (OTU), is frequently used to refer to the individual or taxon being scored for characters. It is from this point of explicit data gathering that the techniques differ. Some basic references for phenetic procedures are Sneath and Sokal (1962), Sokal and Crovello (1970), Rohlf (1992); and for cladistic procedures are Wiley et al. (1991) and Hall (2001).

Phenetics

The advent of computers allowed the application of multivariate phenetic techniques to taxonomic data (grouping based on overall similarity (Sneath and Sokal 1962)). A phenetic philosophy typically analyzes all data types (as morphological, anatomical, chemical, molecular, or any character type) as having equal value, even reproductive characters that in the taxonomic morphological species concept (see below) were typically given more value. An added claim was that these explicit, computer-based methods opened up the new classifications to scrutiny, as all data and analytical techniques were clearly stated and open to re-evaluation by others.

Phenetic philosophy holds that changes in character states are so common that the evolutionary history of organisms can never be determined. As such, organisms are best classified by overall similarity. A phenetic organism-by-character matrix can be: 1) actual quantitative measurements, as lengths, widths, or other quantitative measures of many plant parts; or 2) simple coded qualitative characters as 0/1; or 3) ranges of qualitative measurements, as 0, 1, 2, n; or 4) a mixture of each type of coded character. Generally, it is best to minimize the combination of qualitative and quantitative data because there are different algorithms to analyze each type of data. These characters can be either morphological or molecular characters, but we concentrate in this manual on molecular characters.

As described in "Overview of molecular technologies", molecular marker data are useful in that they can be scored as discrete characters (present=1, absent=0), or for DNA sequence data can be scored as one of the four nucleotides. As detailed in Table 2, some molecular markers

are codominant (showing both alleles) and others dominant (showing only one allele). Dominant markers, such as AFLPs or RAPDs have a higher information content, therefore, in shared present/present (1/1) or present/absent comparisons (1/0) than in absent/absent comparisons (0/0) comparisons; that is, they have more chance of representing homologous states. Because of this, they are best analyzed with proximity algorithms that place no weight on 0/0 matches, as a Jaccard's algorithm (Jaccard 1908), or reduced weight on 0/0 matches, as a Dice algorithm (Dice 1945) also called Nei-Li (Nei and Li 1979). Codominant markers, on the other hand, can be analyzed with similarity algorithms that provide equal weight to all pair-wise combinations, including 0/0 matches, such as the simple matching coefficient.

Phenograms are then generated from these proximity matrices. As with proximity algorithms, choices must be made from a wide range of phenogram methods to use, such as single linkage, complete linkage, unweighted pair group mean with arithmetic averaging (UPGMA), and others. More than one tree can be generated from the same proximity matrix. If more than one tree is produced, "consensus" programs summarize the common branching points of these alternative trees.

Another type of tree building method is the neighbour-joining method. It was developed by Saitou and Nei (1987) as a method for estimating phylogenetic trees. While the method is based on the idea of parsimony (a cladistic technique) as described later, the neighbour-joining method does not attempt to obtain the shortest possible tree for a set of data. Rather, it attempts to find a tree that is close to the shortest phylogenetic tree (Rohlf 1992). This method does not require the use of outgroups but they can be used. It can be viewed as an intermediate type of analysis to phenetics and cladistics.

Different results are produced, therefore, from using different combinations of proximity coefficients and tree building methods. Despite the goal of objectivity of phenetic approaches, there are different choices (many more than presented here) for proximity algorithms and tree building methods. They provide yet different results. In addition to the branching trees (dendrograms) discussed above, there are various ordination analyses, as principal components analysis (PCA) and canonical discriminant analysis (CDA), that provide plots of OTUs in two or three-dimensional space (dots placed in a square or in a box, or cube, with the closer-spaced dots inferring closer similarity among OTUs than farther-spaced dots).

Because the above phenetic approaches use very different algorithms and operate under different assumptions about data sets, multiple phenetic analyses using different proximity algorithms and tree building methods are often used to explore similarities. Since different procedures give various results, conclusions must be drawn through either choice

of only one result, or discussion of the commonality of different results, and both choices have a certain degree of subjectivity. One objective method sometimes applied to choose the "best" tree is to compare the distortion in the different trees relative to the proximity matrix with cophenetic correlation coefficients (a procedure present in NTSYS-pc). These values vary from 1 (no distortion) to 0 (total distortion) and the tree with the least distortion is presented.

After these various analyses are presented, they must be interpreted in light of the question being addressed. Here more subjectivity is encountered, because there is no universal or statistical method to interpret these results. Some investigators interpret phenograms by drawing in a vertical line (phenon line) across branches to make taxonomic decisions by group membership relative to this line (e.g. Figure 4), but it is the investigator's decision where to draw this line, ideally based on inferences from other data. For example, in Figure 4, if the phenon line is drawn as shown, the conclusion may be that there are four taxa (as species or subspecies): 1) OTU 1 + OTU 5, 2) OTU 3, 3) OTU 2 and 4) OTU 4. A different line would reach a different conclusion.

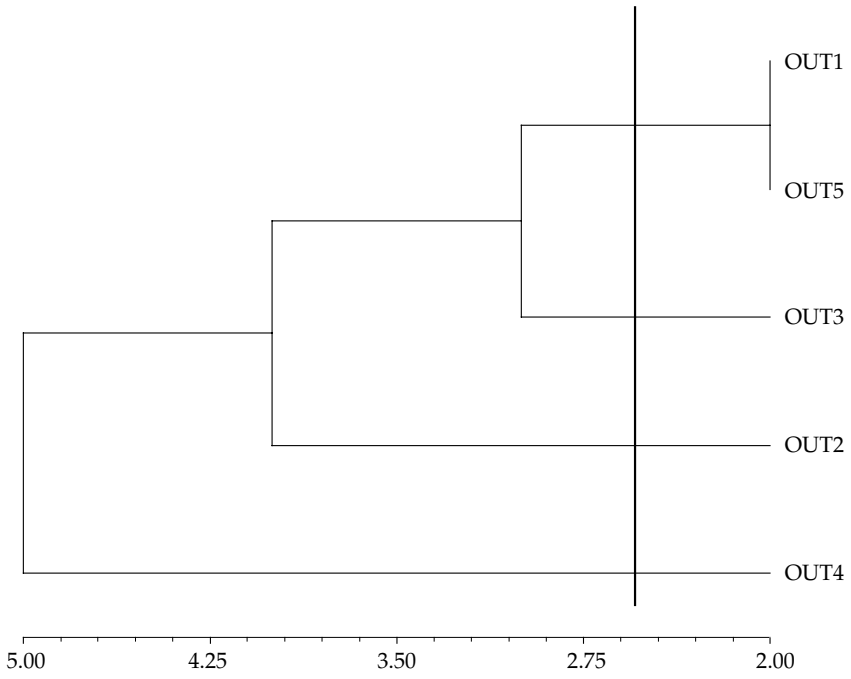


Figure 4. A phenogram with a vertical phenon line drawn at distance coefficient about 2.5.

Cladistics

Cladistics, like phenetics, begins with an explicit analysis of data, and scores these data as an organism-by-character data matrix. However, the assumptions and methods diverge at this point. While pheneticists state that phylogeny is so difficult to reconstruct that overall similarity is the only objective criterion for grouping organisms, cladists state that phylogenetic relationship is the only valid criterion for classifying organisms, no matter how hard it may be to infer.

Cladistics has its own terminology that conveys the basic assumptions and methods that distinguish cladistics from phenetics and some basic cladistic terms are here briefly explained; the reader is directed to Wiley et al. (1991) for more detailed explanations. A monophyletic group encompasses an ancestor and all of its descendants. Put another way, it refers to all of the taxa that trace down to a common branching point in a phylogenetic tree or cladogram. Monophyletic groups are determined by constructing cladograms and the results of the cladograms are used to make decisions of what is a monophyletic group.

To construct a cladogram you begin with what you think may be a monophyletic group, referred to as an ingroup, such as "species A"or "tuber-bearing solanums," or "the sunflower family". Evolutionary relationships within the ingroup are determined by the use of one or more outgroups, that are thought to be closely related to the ingroup. As in phenetics you make a character-by-organism data matrix of the putative ingroup and outgroup(s). A sister group is the most closely related monophyletic outgroup to the ingroup. The proper choice of outgroups is critical to the results, and sometimes the sister group relationships are unclear. To alleviate this problem, further outgroups can be analyzed for a multiple outgroup analysis (Maddison et al. 1984) (Figure 5.).

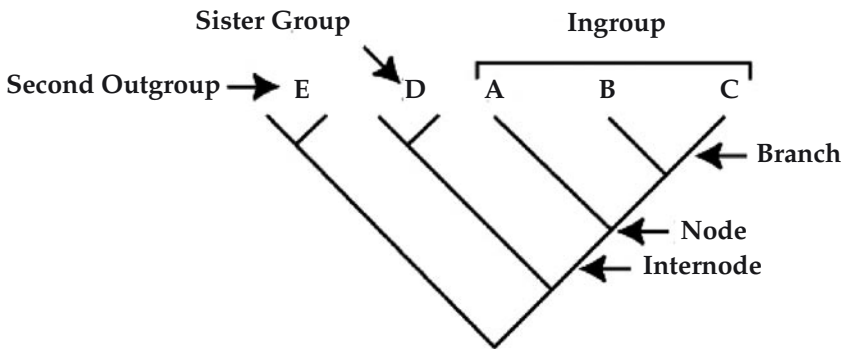


Figure 5. Terms relative to cladograms.

Any character type can potentially be used for a cladistic analysis, including morphological or molecular characters (as DNA nucleotides or restriction endonuclease sites). Most analyses score these characters qualitatively, as presence or absence (1-0), as DNA nucleotides, or as a range of discrete character states (0-1-2-n). Care is taken to score only homologous characters arising from common ancestry, avoiding characters that may look similar but actually arise in parallel from different ancestors, when homology is known. This has important implications for the choice of a molecular marker. As mentioned above, homology is a function of taxonomic distance, and the more rapidly evolving markers would not be expected to be homologous (and therefore inappropriate) for increasingly distant taxa. Orthologous characters are homologous by a speciation event, meaning that they trace their ancestry to a common progenitor, and are taken as the only useful type of homologous character. Molecular taxonomists are searching for single-copy nuclear genes for phylogeny construction while doing everything possible to avoid paralogous characters that have arisen from gene duplication. Such duplicated genes can evolve separately in the same lineage, may falsely appear to be homologous, and can provide misleading phylogenetic information if each copy has independently diverged from the other.

In addition to parsimony, other techniques are used to produce cladograms, depending on the data type. For example, while parsimony is frequently used for DNA sequence data, other techniques, such as maximum likelihood (Felsenstein 1981) and Bayesian analysis (Mau et al. 1999) search for trees that may be longer but that represent character changes based on certain evolutionary models. All of these methods "root" trees based on characters of the outgroup(s), and monophyletic groups are supported relative to the branching patterns of the ingroups. Cladograms may superficially look like phenetic trees (also called dendrograms), but branches of a parsimony-based cladogram are supported by specific characters and phenograms by an average of all characters. As a result, parsimony cladograms have characters supporting each branch shown on the tree, while phenograms only have average similarity values placed under the entire tree.

Pheneticists infer only overall similarity of organisms from their phenograms, not phylogeny, while most cladists interpret cladograms phylogenetically. That is, cladists try to recognize only monophyletic groups and think that phenetic groups (groups resulting from a phenetic analysis) have no reality and should not be used in classification. Cladists refer to different classes of non-monophyletic groups as is diagrammed in Figure 6. These non-

monophyletic terms are used in cladistic discussions all the time so it is important to understand them. They include 1) paraphyletic groups (groups containing some, but not all, descendants of the most recent common ancestor), and 2) polyphyletic groups (the common ancestor is placed in another taxon).

There is a wide diversity of opinion on application and interpretation of these concepts. For example, as described in "Comparison of molecular marker data" below, phylogenetic results of the same organisms obtained from different data sources are frequently in conflict (Wendel and Doyle 1998). Some researchers advocate analyzing different data sources separately to discover datasets providing misleading results, while others advocate combining all data into a single matrix for a total evidence analysis (e.g. Eernisse and Kluge 1993). Cladistic results can be affected by poor choice of outgroups, by analysis of unrecognized non-orthologous characters, by different choice of cladistic algorithms to construct trees, by insufficient ingroup or outgroup sampling, and by different methods to handle missing data. There also is debate among cladists whether cladograms truly reflect recently evolved groups sharing common ancestry (process cladists), or whether they need to be theory neutral and only show patterns decoupled from assumptions of ancestry (pattern cladists or transformed cladists)

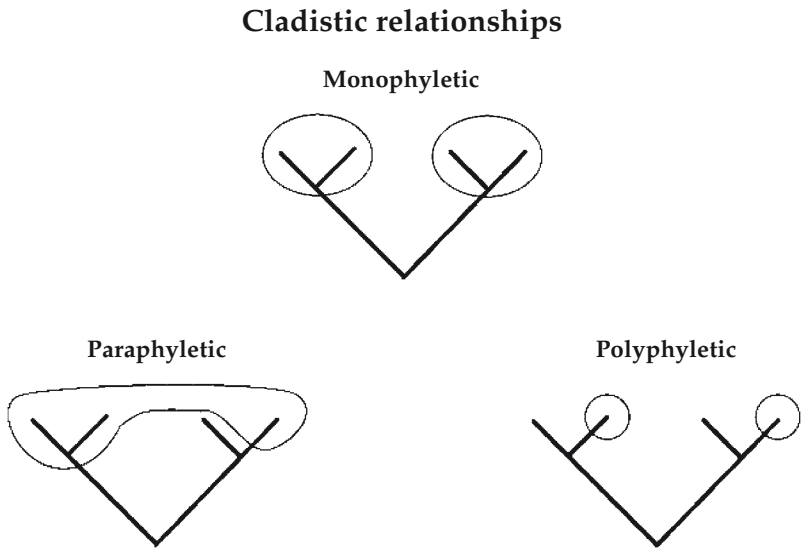


Figure 6. Cladistic relationships relative to cladograms.

(Ereshefsky 2001). Perhaps the greatest source of debate is the use of cladistics at the species level. This is because cladistic procedures assume divergent taxa, yet individuals within species generally hybridize, leading some to consider cladistics to be an inappropriate method to define species (Templeton 1989).

So is it best to use a phenetic or cladistic analysis of your molecular data? A case can be made that low taxonomic level molecular marker datasets (within species or among closely-related species) or dominant characters, should be analyzed with phenetic, as well as with cladistic methods (Koopman et al. 2001). The use of characters that have a greater chance of being homologous (as DNA sequence data), in questions above the species level, should use cladistic approaches, but such cladistic analyses should always choose outgroups to properly root the tree. If outgroups are unknown or problematic, multiple possible outgroups should be chosen as mentioned above.

Congruence of molecular marker data

Congruence of cladistic data

Systematics has progressed through a variety of conceptual and procedural developments including experimental crossing data, secondary chemical and isozyme data, and molecular data. Each development was preceded by an excitement of the primacy of the new technique, and the advantages of molecular data led to similar excitement. However, the use of separate molecular markers for the same germplasm sources frequently show at least somewhat different results, questioning the utility of any individual marker. An excellent review of phylogenetic discordance by Wendel and Doyle (1998) summarized its causes and biological interpretation. Although discordant results can cause problems in the strict translation of molecular results into hypotheses of evolution and into taxonomic treatments, they have the potential to provide insights into potential problems of individual markers, and insights into unexpected evolutionary events (Wendel and Doyle 1998).

Wendel and Doyle (1998) classified discordance into three classes to include: 1) technical causes such as poor gene choice, sequencing error, or insufficient taxonomic sampling; 2) organism level processes such as convergent or rapid morphological evolution, rapid diversification of organisms, hybridization and introgression, lineage sorting, and horizontal gene transfer; and 3) gene and genome-level processes such as intra-genic recombination, use of paralogous genes (genes arising from a single-copy gene that has duplicated and can diverge separately in the same organism) inter-locus interactions and concerted evolution, and non-independence of sites examined in the analyses.

Examples of each possible cause of congruence are given in their paper. We recommend it as standard reading for anyone wishing to choose markers for phylogenetic analysis, to gain appreciation of the need to wisely choose multiple genes for analysis, and possible explanations of why different genes give different phylogenetic results. We discuss some of these reasons below.

Every study is limited by time and funds, and decisions need to be made regarding number of taxa and characters to use for well-supported phylogenetic results. The importance of outgroup(s) is discussed in cladistic studies (see "Genebank management"). Sufficient types and numbers of taxa and numbers of characters also are important. Often, inclusion of taxa is determined by simple availability. For example, studies of small taxonomic groups or of rare taxa must choose accessions from the limited material available. Lack of readily available germplasm, however, is not an excuse for poor experimental design; we have reviewed many unpublishable papers that analyzed only the available accessions in their genebank without realizing that these were insufficient to address significant questions.

Hillis (1998) outlined five possible strategies for adding taxa to a cladistic analysis of different species: 1) add taxa randomly from all living organisms; 2) choose taxa randomly within the monophyletic group being investigated; 3) select taxa within the monophyletic group that represent the overall diversity of the group; 4) select taxa within the monophyletic group that are expected to alleviate problems of "long branch attraction" (this is a phenomenon where strongly unequal rates of evolutionary change in different members of a group under study causes cladistics to produce incorrect trees, see Felsenstein 1978 for details); 5) add or delete taxa until the a-priori biases of the investigator are supported. Most taxonomists would avoid choice 1 because it adds distant outgroups; that is, if you are interested in investigating maize, it adds little to an analysis to use pines as an outgroup, the gene used for ingroup analysis may not analyze distant outgroups well, and the outgroups form long branches. The second choice fails to best analyze diversity within clades, and the fifth choice is unscientific. Most taxonomists choose taxa based on a combination of options 3 or 4.

The largest study to date to analyze discordance of different genes in a phylogenetic analysis was Rokas et al. (2003). They investigated the incongruence of phylogenetic trees produced from an analysis of 106 orthologous genes, from eight species of yeast and an outgroup. These genes, distributed on all 16 yeast chromosomes, comprised a total length of 127 026 nucleotides, encoding 42 324 amino acids, containing roughly 1% of the genomic sequence and 2% of the predicted genes in yeast. This dataset is an order of

magnitude larger than any other one for a study of phylogeny and incongruence. It is made possible because of the complete genome sequences of these eight yeast species and outgroup. The analysis consisted of maximum likelihood (ML) of the nucleotide data, maximum parsimony (MP) of the nucleotide data, and MP of the inferred amino acid sequences, with branch support estimated by bootstrap analysis. Comparisons were made of single genes and all genes together (a concatenated dataset). The results of individual genes produced more than 20 alternative ML or MP trees, but all 3 analytical methods produced a single tree with 100% bootstrap support on each branch when applied to the concatenated dataset. This level of bootstrap support is unprecedented and suggests that large datasets can overcome incongruence in any single-gene analysis. In this study they concluded that: 1) eight genes were required to obtain a mean bootstrap value of greater than 70% with a 95% confidence interval; 2) 20 genes were required for a mean bootstrap value of at least 90% bootstrap support; 3) only 3000 randomly selected nucleotides from the entire concatenated dataset of 127 026 nucleotides were needed to obtain the 70% bootstrap 95% confidence interval tree, corresponding to an average length of only 2.5 genes. The implications, of 8–20 separate genes needed for well-supported phylogenies, need to be tested with other organisms varying by different levels of phylogenetic divergence. The study highlights, however, the need for the use of many genes to obtain well-supported phylogenies.

Congruence of phenetic data

We summarize here some studies examining congruence of phenetic results statistically. Powell et al. (1996b) examined ten accessions of cultivated soyabean, *Glycine max* (L.) Merr. and its progenitor species *G. soja* Siebold & Zucc., chosen to represent the maximum diversity present in the United States collection. They examined expected heterozygosity, multiplex ratios, and effectiveness in assessing relationships between accessions for AFLPs, RAPDs, RFLPs and SSRs. The heterozygosity measures were calculated from actual allele frequencies. The multiplex ratios were calculated from the number of bands simultaneously analyzed per experiment, for example the number of bands resolved on a particular gel. The marker index was the product of heterozygosity and multiplex ratio and is a measure to evaluate the overall information content of a marker system. They found that the average heterozygosity increased from AFLP and RAPD (similar in magnitude) to RFLP to SSR (greatest heterozygosity). The effective multiplex ratio increased from RFLP to SSR to RAPD to AFLP. The marker index increased from RFLP to

RAPD to SSR to AFLP (highest). Despite the low heterozygosity of AFLPs, its high marker index is caused by its very high multiplex ratio (many more bands are detected per each experiment). The hypervariable SSR alleles proved best at detecting individual genetic differences. SSRs were the most effective in showing the differences among individual accessions (average genetic similarity between genotypes 0.341), while AFLPs, RFLPs and RAPDs were less able to distinguish genotypes (0.64–0.66). All marker types were very effective at the interspecific level and distinguished *G. max* from *G. soja*, but differed markedly in congruence within species. At the intraspecific level (when only *G. max* was considered), only AFLP and RAPD similarities were significantly correlated.

Milbourne et al. (1997) examined 16 cultivars of potato. The marker index in this study increased from SSR to RAPD to AFLP. Proximity matrices showed low correlation between the marker types, but the best was between AFLP and RAPD. Russell et al. (1997a) examined RFLPs, AFLPs, RAPDs and SSRs using 18 cultivated barley accessions, chosen to represent the majority of the ancestors of European cultivated barley. SSRs appeared to be the most polymorphic, while AFLPs displayed the highest marker index (0.937). The lowest marker index was observed for RFLPs (0.322). Spearman rank correlations of genetic similarity values ranked over 70% of the pairwise comparisons between AFLPs and RFLPs in the same order. SSRs showed the lowest correlation to the other marker systems. Spooner et al. (1996) examined congruence as a function of the number of markers per marker type, and showed that at the interspecific level there was a gradation of resolution from isozymes to RAPD to RFLP (highest), while at the intraspecific level it was RFLP to RAPD (there were insufficient isozyme markers for intraspecific questions). Lu et al. (1996) compared RFLPs, RAPDs, AFLPs, SSRs and ISSRs for informativeness (level of polymorphism detected and ability to discriminate between germplasm lines) and genetic diversity assessment among 10 pea genotypes. The PCR-based techniques were found to be more informative than RFLPs. A Mantel test revealed significant correlations among trees derived from the different marker systems, except for ISSRs. Pejic et al. (1998) evaluated informativeness and applicability for genetic diversity studies of RFLPs, RAPDs, SSRs and AFLPs using 33 maize inbred lines. SSRs displayed the highest expected heterozygosity and average number of alleles, while AFLPs showed the lowest polymorphism. However, the marker index of AFLPs was found to be more than 10-fold higher compared to the other marker systems. Genetic similarity trees for the different marker types appeared highly correlated, with the exception of the RAPD-based tree.

Other comparative studies of closely related taxa have been conducted by Thorman et al. (1994), Dos Santos et al. (1994) and Hallden et al. (1994) in mustards (RFLP and RAPD); Sharma et al. (1996) in lentil (AFLP and RAPD); Lin et al. (1996) in soyabean (RFLP, RAPD and AFLP); Yang et al. (1996) in Chinese sorghum (RFLP, RAPD and ISSR); Olufowote et al. (1997) in rice (RFLP and SSR); Nagaoka and Ogihara (1997) in wheat (RFLP, RAPD and ISSR); Parsons et al. (1997) in rice (isozymes, RAPD and ISSR); Fang et al. (1997) in trifoliate orange (isozymes, RFLP and ISSR); Virk et al. (2000b) in rice (isozymes, RAPD, ISSR and AFLP); Anthony et al. (2002) in coffee (SSR and AFLP); Palombi and Damiano (2002) in kiwifruit (RAPD and SSR); and Lopez-Sese et al. (2002) in Spanish melon (RAPD and SSR). In general, the highest level of polymorphism per marker was found for SSRs, while AFLPs displayed the highest marker index. Compared to closely related accessions, better congruence was found between marker systems for more distantly related accessions.

An additional comparison of congruence is based on pedigree data. There are potential problems in these comparisons because of incomplete or incorrect pedigree records, assumptions of equal parental contribution to the genetic makeup of the cultivar, and different methods to generate pedigree estimators and compare them to genetic identity statistics. Schut et al. (1997) used 31 barley lines to investigate the congruence between similarity values for 681 AFLP markers, pedigree-based coefficients of co-ancestry, and morphological distance based on 25 characters. Using a core collection of 25 European two-row spring barleys, AFLP and pedigree data showed poor to moderate correlation, while the morphological data were not significantly correlated with either the AFLP or pedigree data. However, inclusion of more distantly related barleys (winter types, North American origin, six-row barleys) improved the correlation between the AFLP and pedigree data considerably. In general, congruence between molecular and pedigree data varies greatly among studies. For example, significant but low correlations between pedigree and genetic similarity were shown by Graner et al. (1994) in barley (RFLP), O'Donoughue et al. (1994) in oat (RFLP), Hallden et al. (1994) in mustards (RFLP and RAPD), Ahnert et al. (1996) in sorghum (RFLP). "Reasonable to good" correlations exist between pedigree and genetic similarity as shown by Mumm and Dudley (1994) in maize (RFLP); Hill et al. (1996) in lettuce (AFLP); Doldi et al. (1997) in soyabean (RAPD and SSR); Huff (1997) in ryegrass (RAPD); Paz and Villeux (1997) in potato (RAPD); and Prabhu et al. (1997) in soyabean (DAF and RFLP).

The choice of markers for phylogenetic or diversity studies will depend on many factors. These include the anticipated degree of relatedness between accessions. Because of their extensive polymorphism and frequently observed PCR problems at higher taxonomic levels, the utility of SSRs lies in investigating very closely related taxa. The high marker index and general concordant results make AFLPs very useful for close-to medium-divergent (within genera) studies. The high conservation of RFLP markers makes them more useful for comparative genome studies. It is likely that at low taxonomic levels (closely related accessions), perfect congruence of any markers will be an elusive goal.

One long-held assumption in the choice of molecular markers was that the most reliable ones for diversity assessment would be those that, through prior mapping studies, were shown to be evenly spread throughout the genome (e.g. Bonierbale et al. 1995; Karp et al. 1997a). However, Virk et al. (2000a) found no advantage of using mapped AFLP markers compared to unmapped markers in assessing genetic relationships among rice accessions. They even concluded that the use of mapped markers may lead to misleading patterns of diversity because results are biased towards differences between the parents used to obtain the mapping population.

Predictive value of taxonomy

A major justification for investment in taxonomic research is that it produces something useful. One claim of its use is its "predictivity". Taxonomy has many components, but two of the most important are: to determine what is a species; and to look for the interrelationships among these species. Interrelated species are grouped into higher-level taxonomic ranks such as genera. The predictive goal is to use these species, genera, and other taxonomic ranks to make inferences about the entire group when you have data for only some of its members. For example, we may know from actual cases of poisoning that, "species A has is a deadly toxin to humans", or that, "species B is a useful wild species to be used in a potato breeding program for late blight resistance". Similarly, we may infer that the species most closely related to a species with a known trait more likely share that trait than do unrelated species. These predictive statements help us avoid or choose these species to fit our needs.

The idea that taxonomy serves to make such predictive statements has long been accepted (Michener 1963; Rollins 1965; Warburton 1967; Sokal 1985; Stuessy 1990; Miller and Rossman 1995; Daly et al. 2001). Clearly, better taxonomic classifications would be expected to make better predictions. Modern taxonomy has undergone a renaissance during the last 20 years due to new molecular data

and improved taxonomic theory and methodology that allows the recognition of monophyletic ("natural") groups. These developments have upgraded the position of taxonomy in the biological sciences and clearly have improved the link to prediction, addressing a range of questions regarding biosystematic and developmental pathways, sources of natural products, origins and migrations of evolutionary lineages, and conservation.

There are dramatic examples of this increased prediction role through cladistic studies, as outlined in Daly et al. (2001). For example, 15 families of angiosperms were known to produce glucosinolates, and these traits were thought to have evolved separately many times. Cladistic analyses of these families showed them to be part of a single clade (monophyletic branch of a phylogenetic tree), called the Brassicales, with the exception of the genus *Drypetes* Vahl that was in another clade called the Malpigiales. This cladistic result suggested that there were two evolutionary origins for mustard glucosinolates (Rodman et al. 1993), and indeed glucosinolates are derived by two different biosystematic pathways in these two orders, showing they had different historical origins. Daly et al. (2001) point out two other dramatic examples of prediction. The important bioactive compound taxol was known only from members of the plant family Taxaceae, but searches for other sources in the most closely related family Podocarpaceae identified new sources there (Stahlhut et al. 1999). In another example, new data on the interrelationships of the angiosperms have shown that those families exhibiting nitrogen fixation are members of one clade, suggesting a single origin of this syndrome of traits (Soltis et al. 1995). Such striking concordances have increased our confidence of the power of the cladistic method to address prediction. We know of no similar studies to demonstrate the predictive role of taxonomy at lower taxonomic levels, but it is widely assumed to occur.

In an applied sense, prediction means that germplasm can be chosen or avoided by breeders based on past positive or negative evaluations of related species. Germplasm evaluations of resistance or agronomic traits are common in the literature. For example, species-specific statements about the breeding value of wild potato germplasm are found in Ross (1986), Hawkes (1990) and Ruiz de Galerreta et al. (1998). Clearly, not all accessions of a species will share all traits, but lacking prior evaluation data, taxonomy should provide a useful guide to make inferences about unevaluated germplasm based on present knowledge.

However, claims of the predictive component of taxonomy have always been by post-hoc discoveries of associations that match expected prediction models. The dramatic examples of the association of glucosinolates, taxol and nitrogen fixation to new knowledge of

cladistic relationships are intriguing, but they ignore non-associations in other groups that need to be considered and tested statistically. Without such tests, there is no way to convincingly demonstrate that associations of traits to taxonomy are not simply due to chance. A-priori experiments are needed to specifically test the prediction assumption empirically before convincing statements can be made of the overall strength of taxonomy as a predictive tool. Most of the above predictions are for traits under polygenic control, but prediction also can address simply inherited traits.

Taxonomy is not the only possible predictor. Biogeographical variables have also been used to predict the presence or absence of traits in wild plants. Biogeography-based hypotheses of association are fundamental in guidelines for collecting plant genetic resources. For example, it is often suggested that germplasm collectors should sample from as many ecologically different environments as possible (Brown and Marshall 1995). It is also thought to be important to include geographical extremes of the range of a species (Allard 1970). Although such populations may not present great genotypic variation, they may harbor unique traits or taxa (von Bothmer and Seberg 1995).

The presence of biogeographical associations might reflect adaptation of plants to prevailing ecological conditions where they grow. Rick (1973) and Nevo et al. (1982) found similar convergences in resistance to drought in populations growing in dry areas. One would also expect to find similar traits in areas with a comparable bioclimatic environment, even when these areas are far apart and there is no genetic exchange among the populations. In the case of disease resistance, adaptation could arise in an area as a result of coevolution of a pathogen with a limited range. Thus, resistance to a certain disease may be present in areas where the pathogen is endemic, whereas it may be absent in areas that are similar from an ecological and/or taxonomic perspective but where the pathogen is absent. For example, nearly all R genes for resistance to potato late blight have been found in accessions from central Mexico. This is the centre of diversity of late blight and its likely centre of origin (Fry and Goodwin 1997). Patterson and Givinish (2002) showed convergence of several non-disease related traits in different clades in the Liliales, and termed this phenomenon concerted convergence.

In convergence situations one expects to find the presence of traits to be restricted to certain areas and independent of species. In the case of widespread species, one would expect the trait to appear (or increase in frequency) as the selection pressure (whether ecological or coevolutionary) increases. Thus, this strongly contrasts with the taxonomic prediction paradigm. In reality, both taxonomic and biogeographical factors may be associated with distribution of traits

(Hijmans et al. 2003). The question then becomes how to best use these predictive capacities. Which of the two is a stronger predictor? How may they complement each other?

Wild crop relatives provide an outstanding resource to test prediction. Many large germplasm collections are well studied regarding taxonomy, distribution, and have large databases of screening data for useful traits such as disease resistance. The greatest impacts of disease resistance from wild germplasm in food crops have been in wheat, potato and tomato (Lenné and Wood 1991). Because of the economic importance of disease resistances and other crop improvement traits, and for many other purposes, large germplasm collections have been assembled, and they represent a ready resource for prediction studies.

Species concepts

The idea of what constitutes a species has changed from original concepts based on special creation by God, to modern ideas incorporating biological and DNA data. Originally, species were defined by simple impressions of differences as determined by morphological observations. Our concepts of species changed with new discoveries of genetics, reproductive biology, data analyses, and theory of what forms and maintains species. Many different species concepts arose from these developments, and there is no consensus on what constitutes a species. The purpose of this section is to introduce the reader to basic species concepts to allow you to interpret your own data, and that of others, in the basic framework of these historical developments and conceptual differences of species.

Taxonomy is the theory and practice of describing, naming and classifying organisms (Lincoln et al. 1998). Systematics is a related term, sometimes used synonymously, but involves a broader discipline of discovering phylogenetic relationships through modern experimental methods using comparative anatomy, cytogenetics, ecology, morphology, molecular data, or other data (Stuessy 1990). We use the term taxonomy to describe all these aspects here for simplicity. Taxonomy has many components but primarily involves: 1) determining what is a species (or their subdivisions, as subspecies); 2) distinguishing these species from others through taxonomic keys and descriptions and examining the geographic ranges of species; 3) investigating their interrelationships; and 4) determining proper names of species and higher order ranks (as genera or families) using international rules of nomenclature. In addition, some taxonomists investigate processes of evolution that lead to the existing pattern of species and their interrelationships (Spooner et al. 2003).

Standard ranks in the taxonomic hierarchy from lowest to highest are form, variety, subspecies, species, series, section, genus, tribe, family, order, class, division and kingdom. More ranks can be added if desired by adding qualifier terms such as "sub" or "super", for example, to create subgenus or supergenus. Ranks only have meaning in a relative (not absolute) sense in that a genus is less inclusive than a family, and a family less inclusive than an order (Stevens 1998). There are no objective criteria or set of characters to indicate what taxonomic level is a genus, family, or order. As such, families or any rank are not comparable regarding the relative diversity they contain or how diverged phylogenetically they are to other ranks. Put in a phylogenetic context, traditional ranks are not necessarily equivalent in that they do not designate sister clades. Many taxonomists today are attempting to have taxonomy reflect the branching patterns of phylogenetic trees. In some cases traditional ranks are not monophyletic, and in others, one clade could represent a family and its sister clade could represent a genus.

There are many reasons why taxonomy is important. Taxonomists use standardized rules of nomenclature to make names as stable as possible. For wild and cultivated plants, taxonomists use the International Code of Botanical Nomenclature (ICBN) (Greuter et al. 2000), and for cultivated plants use the International Code of Nomenclature for Cultivated Plants (ICNCP) (Brickell et al. 2004); the use of either code is allowable for cultivated plants. Such stable names help maintain continuity of the scientific literature and allow all disciplines to communicate among each other with a common language. Taxonomy also produces descriptions and distribution maps and identification aids (as taxonomic keys) that aid identification of unknown specimens. The predictive role of taxonomy is described above.

Taxonomic treatments are especially useful for genebank managers. For example, a stable nomenclature allows efficient literature searches for traits of breeding interest, suggests localities where species are known to occur in the wild and may be in need of collecting, defines species diversity (as determined by numbers of species per area) for biodiversity conservation, guides genebank managers to rationally organize and document collections, and guides breeders to possible sexual crossability. Despite the above admirable taxonomic goals, a reliable taxonomy can be difficult to obtain because of different classification philosophies or practices based on morphological, crossability, phylogenetic, ecological, molecular, or other data. For example, Mayden (1997) lists a total of 22 different species concepts. We here list only the major variants of these concepts and summarize them into six major classes

(modified from Spooner et al. 2003). There are various perspectives on the proper criteria for recognizing species that lead to different classifications as mentioned below.

Morphological species concepts

Morphological species concepts define species entirely on morphological or anatomical characters. Because of their utility, they have been frequently applied, especially historically when taxonomy was primarily based on morphological data from herbarium specimens. Taxonomists can apply these concepts very effectively by gaining initial impressions of species limits from examination of herbarium specimens, sometimes followed by microscopic examination to gain additional data to modify species delimitation. Cronquist (1978) defined this very practical application of the morphological species concept as the taxonomic morphological species concept: "Species are the smallest groups that are consistently and persistently distinct and distinguishable by ordinary means".

The characters leading to this subjective judgement are often unclear, sometimes even to the taxonomist applying them. The advent of computers and phenetic philosophy led to a more objective evaluation of characters, and the phenetic species concept arose to allow taxonomic decisions based on clustering of individuals. Sokal and Crovello (1970) defined the phenetic morphological species concept as: "dense regions of hyperdimentional space" (referring to clustering of individuals in ordination analyses), but the definition refers to dendrograms as well.

Interbreeding species concepts

The interbreeding species concepts focus almost entirely on the ability of species to exchange genes, either naturally or artificially, as assessed by artificial crossing programmes, studies of mechanisms to facilitate gene flow, and biological isolating mechanisms. Mayr (1942) advanced the biological species concept (an interbreeding concept) as: "Species are groups of interbreeding natural populations that are reproductively isolated from other such groups". This concept matches that held in the minds of the general public and is intuitively appealing, but there are many practical and theoretical problems in applying this concept. Procedurally, it is almost impossible to apply to a group of any size because replicated pair-wise crosses are needed in most interspecific combinations to be confidently interpreted (Sokal and Crovello 1970). As well, data from greenhouse situations do not always match crossability data in natural situations, and organisms frequently display varying

degrees of crossing success that make interpretation of the data difficult. Also, the concept is inapplicable to species reproducing asexually. The lifetime of crossing studies by Rick (1963, 1979) in tomato is a notable application of this concept, but this depth of study is exceptional and rarely been applied as thoroughly in other groups.

As pointed out by Mallet (2004) Mayr's biological species concept had roots in a paper by Poulton (1904), perhaps the first paper devoted entirely to the discussion of species concepts. This early paper outlined key ideas important to species theory today, including reproductive and geographic isolation, the classification of isolating mechanisms, and the term sympatry, a key concept referring to species growing in the same geographical area.

Ecological species concept

Van Valen (1976) noted the perplexing array of variation in oaks that often have broadly sympatric (growing in the same area) sets of very similar species, often hybridizing among each other. He noted that despite many hybrids, oak species are often maintained in their distinct habitats. For example, swamp white oak is broadly sympatric with burr oak in the Great Lakes and Ohio River basins, and they frequently hybridize. The former, however, grows in wet bottomlands, stream sides and swamps, and the latter in moist habitats of rich woods and fertile slopes. Van Valen was so influenced by the ecological partitioning of distinct species in specific habitats that he contended that "The control of evolution is largely controlled by ecology and the constraints of individual development." He defined the ecological species concept as: "A species is a lineage (or closely defined related set of lineages) which occupies an adaptive zone minimally different from that of any other lineage in its range and which evolves separately from other sets of lineages outside its range." He contended that ecological factors are more closely related to genetic differences than reproductive isolation.

Cladistic species concepts

The most recent and conceptually difficult species concepts are those based on cladistic criteria. Cladistics, as discussed in "Genebank management", also has its own unique set of terms that can initially make cladistics difficult to understand. Cladistic species concepts arose out of the ideas of Hennig (1950, 1966) who grouped taxa entirely on historical relationships as determined by cladistic analyses. Hennig never used cladistics to define species, but only to group species, but some taxonomists have later applied cladistics at the species level to try to define species.

Cladists investigate progenitor-derivative relationships, based on shared derived character states, as determined from states in related taxa (outgroups) to form monophyletic groups. The ability to interbreed is viewed by cladists as a potentially misleading character for assessing interrelationships, however, because it is not a shared derived character. Rather, two species often diverge by a breeding barrier that separates them, and the inability to interbreed is often derived. For instance, two species sharing a common branching point on a phylogenetic tree (sister species) may have diverged after an isolating mechanism that prevented their interbreeding, while more distantly related species of the same group may have retained the ability to interbreed (a shared primitive character). In such cases, biological and cladistic species concepts provide very different delimitations of species (Cracraft 1989).

The attempt to apply cladistic criteria to define species poses problems, with the main one being that cladistic approaches search for patterns of divergence, yet species are composed of potentially interbreeding populations. Theory shows that those taxonomic markers useful for defining species and subspecific taxa can show different cladistic results depending on potential levels of crossing over and linkage relationships (Maddison 1995). As a result, studies using multiple genes useful for studies at the species level have been advocated to search for points of agreement in different gene trees that may indicate a species divergence (Baum and Donoghue 1995).

Rieseberg and Brouillet (1994) and Olmstead (1995) have argued that geographically localized models of speciation typically produce many monophyletic daughter species and an extant paraphyletic progenitor species, and argue that a strict concept for monophyly fails for many species. Olmstead (1995) termed the former apospecies and the latter plesiospecies. Castillo and Spooner (1997) applied this concept to locally endemic species of wild potatoes arising from more widespread progenitor species to recognize plesiospecies and apospecies.

Eclectic species concepts

The former species concepts highlight single processes to define species. Eclectic species concepts, in contrast, assume that species are formed and maintained by a variety of forces. Doyden and Slobobchikoff (1974) constructed a flow chart detailing a variety of morphological, geographical, biological and ecological criteria to define species. Ereshefsky (2001) outlined several classes of species concepts, and advanced a pluralist species view that no single correct definition of species exists and that a number of alternative concepts may be legitimate. Mallet (2004) highlights a key paper by Poulton

(1904) as perhaps the first devoted entirely to the discussion of species concepts. Poulton argued that species formed reproductive communities, the individual members of which were united by common descent, combining aspects of the biological and cladistic species concepts.

Nominalistic concepts of species

Some question the very existence of species, and believe that individuals or interbreeding populations are the only entities that have any objective reality. Such ideas arose out of the philosophy of nominalism, arguing that only individuals are real and that classes of any kind (as species, genera, or families) are artificial constructs. For example, Burma (1954) stated, "…species are highly abstract fictions". Levin (2000) likewise argued that the local population is the only unit of evolution, and species are artificial. Ehrlich and Raven (1969) documented many cases of reduced gene flow in both plants and animals that would preclude any cohesive force to maintain species. They contended, "Selection alone is both the primary cohesive and disruptive force in evolution…for sexual organisms it is the local interbreeding population and not the species that is clearly the evolutionary unit of importance".

So just what is a species?

Our discussion presents only a small sample of a wide diversity of opinions on what constitutes a species, and the criteria important for their recognition. Our division of criteria to define species is really a "taxonomy" of species concepts, and there are other ways to view these criteria. For instance, Mallet (2001) argues that a clear distinction needs to be made between the data used to define species (e.g. morphology, isolating mechanisms, molecular data), and the methods used to analyze these data (e.g. phenetics, cladistics, population biological methods). Under this perspective, the term "morphological species concept" is a misnomer in that it describes only one data type, which with others, such as DNA characters, should be used together to define species. He also points out that the "reality" of many species decreases over wide geographic ranges. For instance, in small areas species may appear to be discrete, but over wider ranges show more variability that makes them harder to distinguish from similar species. We fully agree with these ideas and present our species "taxonomy" as we do to highlight the different criteria that have developed to define species.

Despite some valid arguments for "species are not real", there is utility in their recognition as we can best define them because they have practical value. These include communication among plant

breeders and other biologists, biodiversity conservation, ecological studies, and legislation of biodiversity. Genebank managers are particularly dependent on species names to collect, manage, and legally distribute species across borders. Our perspective is that there is a wide continuum between well defined to almost impossible to define species, but that we must do our best to define them for these practical needs. It is crucial for us to understand that these different concepts exist, and to interpret the literature with these differences in mind. Ideally, genebank managers will cooperate to gather a common set of evaluation data to classify germplasm, in order that common concepts can be applied and compared across collections. We believe that a variety of criteria are useful to help define species, including morphology, molecular markers useful at the species level (as discussed in "Overview of molecular technologies"), crossing studies, and other types of data. There has never been a single set definition to define species and likely there never will be one. What is most important for genebank managers is to gather their data in well designed experiments relative to choice of taxa and markers, to analyze their data appropriately, to make their data publicly available for possible reanalysis and evaluation by others, and to interpret and discuss their results with full knowledge of the above concepts.

Characterization of germplasm

Systematics

Cultivated plants evolve through human selection, unlike natural selection of wild species, which can have a profound effect on the morphology of cultivars and thus on their classification. They are often distinguished from their wild relatives by artificially selected, novel and extreme morphological and physiological differences relating to seed dispersal, inflorescence architecture, seed size, and gigantism that may make the relationship to their progenitor(s) unclear. Traits that are typically selected for cultivated forms can reduce fitness in wild habitats (e.g. lack of seed dormancy or seed shattering) but can confer a selective advantage in the drastically different artificially selective environment of cultivated habitats.

Different taxonomists of the same crop often construct different taxonomic treatments and it is often hard for users to choose one that is "better" (Harlan and de Wet 1971). Also, crop species often are more "over described" (too many species recognized) than is typically found in taxonomic treatments of wild plants. Many crops have only recently diverged from their wild relatives, and often form hybrid swarms with them. The same crop may have arisen independently many times. For example, multiple crop origins have

been demonstrated for common bean (Sonnate et al. 1994), cotton (Wendel 1995), millet (Yabuno 1962), rice (Second 1982), and squash (Decker 1988). They often form intergrading, reticulating gene pools with their wild relatives, and species distinctions are often obscure. Because of this, genebank curators should be skeptical of taxonomic treatments of their cultivated collections and should critically review the criteria that were used for the taxonomy, and compare possible alternative taxonomic treatments. Genebanks with global collections provide a wonderful resource of well-represented collections to address these taxonomic questions.

Taxonomic questions are addressed by almost every molecular marker class from microsatellites useful at the species level to DNA sequences useful at the generic and family levels. For example, Roa et al. (1997) used phenetic analyses of AFLPs to investigate the origin of cassava (*Manihot esculenta* Crantz) relative to four other wild species and two non-cultivated subspecies of *M. esculenta*. These subspecies were thought to be progenitors of the cultivars or escapes from cultivation. As is typical in crops, the cultivars contained less variation than their putative wild progenitors. The two non-cultivated subspecies showed distinct AFLP differentiation from the cultivars suggesting they were progenitors, not escapes from cultivation. Species-specific markers characterized the species but not the two subspecies. A morphological analysis of the same accessions likewise failed to separate subspecies, suggesting that they were unworthy of separate taxonomic status. *Manihot tristis* Müll. Arg. was supported to be most similar to *M. esculenta*.

Kardolus et al. (1998) used cladistic and phenetic analyses of AFLPs to investigate interspecific relationships of wild potatoes, with tomatoes as outgroups. The results were in broad congruence with other modern molecular results, showing the applicability of AFLPs at this taxonomic level. Many other interspecific studies, including appropriate outgroups, have examined relationships with RAPDs and isozymes (Maass and Klaas 1995, onion; Chan and Sun 1997, amaranths), and RFLPs and RAPDs (Miller and Spooner 1999, potato).

In general, genebank collections consist of multiple accessions of the same or similar species, and their identification can be difficult. Taxonomic misidentifications of genebank accessions can be common, leading to confusion by users of this germplasm and perpetuation of errors in resulting publications. Molecular markers provide tools to test the taxonomic validity of species and can provide species-specific diagnostic markers. For example, Martin et al. (1997) used RAPDs to re-identify a collection of oat accessions. Lee et al. (1996) used RAPDs to discriminate *Brassica* L. varieties. Sharma et al. (1995) effectively used RAPDs to identify species-

specific markers for lentil. Species-specific markers also have been found with minisatellites (Baurens et al. 1997, banana), and Zhou et al. (1997) used minisatellite sequences to discover genome-specific fragments in oat, but not cultivar-specific fragments.

McGregor et al. (2002) used AFLPs to characterize 314 accessions of the wild potato subgroup *Solanum* L. series *Acaulia* Juz. Series *Acaulia* consists of the species *S. albicans* (Ochoa) Ochoa and *S. acaule* Bitter, with the latter subdivided in four subspecies. An UPGMA cluster analysis of the AFLP data grouped the majority of accessions into species and subspecies. The AFLP data uncovered 16 taxonomic misidentifications to species or subspecies including four cases that were later identified as different species outside the series. The AFLP data also allowed determination to subspecies rank of 97 accessions that previously were described as *S. acaule* only. Two accessions appeared to consist of a mixture of species. AFLP analysis, therefore, proved to be a very powerful tool to verify the taxonomic status of accessions within the series *Acaulia*, and hence contributed considerably to more reliable identification of this collection.

Santacruz-Varela et al. (2004) used morphology, isozymes, and microsatellites to investigate the origin of North American maize relative to populations from Mexico and South America. They discovered three distinctive groups from North America, and proposed their recognition as distinct races: 1) North American Yellow Pearl Popcorns that represent the common popcorn for US production and originated from Chilean races in the 19th century; 2) North American Pointed Rice Popcorns that were commercially important in the first-half of the 20th century and originated in central Mexico; and 3) North American Early Popcorns, sharing a diversity of traits with Northern Flint Maize, other Mexican races and European popcorn varieties introduced in the late 19th century.

Fingerprinting studies

Advanced cultivars of many crops frequently are morphologically very similar, and may have arisen from their progenitors only by somatic mutations. "Fingerprinting" is an attempt to discover cultivar-specific molecular markers that aid their identification. Certain cultivars are economically very important and form the backbone of regionally important industries. The grape cultivar 'Sangiovese' is a case in point. It is one of the economically most important grape cultivars in Italy and is the major cultivar of Tuscan wines. Many phenotypic variants of 'Sangiovese' exist, and the identification and maintenance of these lineages is important to the industry. Some members of the 'Sangiovese' group are local variants and are given different names. The identification of these variants is

important for identification of clones to be used in breeding and for germplasm conservation. Grapes show great SSR allelic diversity and priming sites have been characterized (Vignani et al. 1996; Lamboy and Alpha 1998). SSRs are logical choices for investigating differences at the taxonomic level of these closely related clones. Vignani et al. (1996) used 7 SSRs to investigate 12 'Sangiovese' clones. Eleven of these 12 clones were identical at all 7 loci, but 1 clone differed by 1 allele at each of 4loci. The results support the "clonal" nature of these 11 accessions. SSRs do not provide a clear answer regarding the relationship of the twelfth clone to 'Sangiovese', but a close relationship is inferred from much allele sharing with the other clones.

Potato is another clonal crop with many morphologically similar cultivars. Schneider and Douches (1997) were able to distinguish 24 of 40 potato cultivars with 7 SSR primer pairs. When the SSR data were combined with the tuber morphology, only five pairs of cultivars could not be distinguished. Similarly, Mandolino et al. (1996) distinguished eight potato cultivars with two probe-enzyme RFLP combinations and three RAPD primers. Crops propagated by seed have been distinguished as well. For example, Charters et al. (1996) demonstrated the ability to distinguish all 20 *Brassica* cultivars examined with just 2 SSR probes. Other fingerprinting studies involve DNA Amplification Fingerprinting (Weaver et al. 1995, Eremochola Buse; Scott et al. 1996, *Chrysanthemum* L.), RAPDs (Lee et al. 1996; mustards; Golembiewski et al. 1997, *Agrostis* L.; Degani et al. 1998, strawberry), and SSR (Guilford et al. 1997, apple; Russell et al. 1997b, barley; Lamboy and Alpha 1998, grape). These many studies clearly document the use of SSRs for fingerprinting genebank accessions.

Putative natural hybrids

Hybridization is thought to be a major evolutionary force at both the diploid and polyploid levels (Rieseberg 1995). The major data used to infer whether accessions are hybrids have been additive morphological traits. Nonetheless, Rieseberg and Ellstrand (1993) showed that hybrids are no more likely to display intermediate character states than parental ones, and can additionally express an array of transgressive or novel traits. Other long-held beliefs that hybrids are less fit than their parents, and that they exhibit character coherence (parental characteristics remain associated) have also been questioned (Rieseberg 1995).

The utility of molecular markers to investigate hybrids decreases with time of divergence, as species-specific markers from both parents can be disrupted through recombination, and both can

mutate to new markers. This can happen very rapidly as was shown by Song et al. (1995). They showed extensive loss of markers and the appearance of novel new markers in just the F_5 generation of an artificial interspecific *Brassica* hybrid. This was constructed as a homozygous allopolyploid, demonstrating that at least on the polyploid level species can generate extensive genetic diversity in a very short period of time.

Molecular markers present powerful new tools to reinvestigate hypotheses of hybridization. One method is to search for markers in the hybrid that are found in both putative parents. The utility of this method depends on discovering species-specific markers, yet this may be difficult because many hybrids are between closely related taxa that have not diverged enough to have formed specific markers. Additive molecular markers were used to investigate three hypotheses of hybrid origins in wild potatoes. In all three hypotheses discussed below, hybrid origins were strongly inferred from intermediate morphology and distribution of the hybrids at contact zones of the parents, yet molecular data supported only one of these hybrid origins. Clausen and Spooner (1998) supported a prior hypothesis of hybridization in the wild potato species *Solanum* × *rechei* Hawkes & Hjert. with RFLP data. Miller and Spooner (1996) on the other hand, failed to support a hypothesis of introgression of the wild potato species *S. microdontum* Bitter into *S. chacoense* Bitter with RAPD and RFLP data, and Giannattasio and Spooner (1994) and Spooner et al. (1991) failed to support a hypothesis of hybrid origin of *S. raphanifolium* Cárdenas & Hawkes with nuclear RFLP and chloroplast DNA data. These results show that morphology can provide misleading clues to the hybrid nature of accessions and show the use of molecular markers to reinvestigate hybrids.

A classical experimental method to reinvestigate hybrid origins is to re-synthesize an artificial hybrid and to compare it to the natural hybrid relative to morphology, reproductive success, or a compliment of parental molecular, biochemical, or ecological traits. Rieseberg et al. (1996) investigated the diploid hybrid origin of the wild sunflower species, *Helianthus anomalus* S. F. Blake, which is derived from two chromosomally divergent parental species, *H. annuus* L. and *H. petiolaris* Nutt. Earlier studies had already confirmed the hybrid origin of *H. anomalous* (Rieseberg 1991). Rieseberg et al. (1996) additionally investigated its genome composition via the construction of molecular linkage maps with RAPDs. They produced three separate hybrid lineages to the F_3 generation by different designs of backcrossing, and compared the genomes of these three lineages to the natural hybrid. The genome compositions of all three artificial hybrids and natural hybrid were

remarkably similar. In addition to providing yet more evidence for hybridity, this study showed that blocks of chromosomes were rearranged as conserved units in both natural and artificial hybrids. It also showed the resistance of certain genes for recombination, and suggested that positive gene interactions drive the selection for these rearrangements in the hybrid. Because many crops are thought to be of hybrid origin, or undergoing introgression with related weeds, such studies can be applied to many other crops.

Genomic differences in wild relatives of crops

The genome refers to the array of genes carried by different individuals. Major genomic differences are usually represented by separate capitol letters (as genomes AA, BB, AB for diploid species, or AAAA or AABB or BBBB for a tetraploid), with one letter per haploid genome set. Minor genome differences are typically expressed as subscripted or superscripted variants of individual genomes (as A^aA^a, or A^1A^1).

Because of the practical importance of interbreeding crops for crop improvement, the discovery and manipulation of species-specific genomic differences is of use to agronomists (Smartt and Simmonds 1995). Although genomes are primarily defined by crossability data, they have been inferred in the absence of these data by phylogenetic investigations. For examples, Kollipara et al. (1997), Singh et al. (1998) and Hymowitz et al. (1998) collectively investigated cladistic relationships, via DNA sequencing from the internal transcribed spacer region of nuclear ribosomal DNA (ITS), of all 18 wild and cultivated soyabean species, and assigned genome symbols to unknown species by phylogenetic results. Biochemical, cytogenetic and molecular data are congruent with genome designations in these species.

Singh and Smartt (1998), however, point out incongruence between cytogenetic and molecular evidence to genome hypotheses in the cultivated peanut (*Arachis hypogaea* L., a tetraploid of AABB genome constitution). Cytogenetic evidence supports *A. batizocoi* Krapov. & W.C. Greg. as the closest B-genome diploid species relative of *Arachis hypogaea* but RFLP data suggests that *A. batizocoi* is more distantly related to *A. hypogaea* than other species of the section *Arachis*. In this case, therefore, RFLP data may not be a good indicator of genomes in *Arachis*. They suggest that additional crossing evidence is needed to infer genome progenitors in this crop. Because genomic differences are so important for predicting the ability to hybridize or to introgress genes among different species, further studies of genome differences in crops are needed.

Rationalization of germplasm collections

Redundancies (identical or near identical accessions) may occur in collections for various reasons, e.g. documentation errors, exchange of identical accessions between genebanks, and the sampling of multiple populations from genetically homogeneous collection sites. Clearly, redundant accessions are a nuisance to genebank curators because they do not contribute to the genetic diversity of a collection, but require resources to maintain them. Identification and elimination of redundancies are therefore important aspects in plant genetic resources management, both from a genetic and economic view. Comparison of passport data may only suggest potential redundancies within collections, which subsequently may be validated with morphological data.

Currently, molecular markers are being widely applied to identify or validate redundancies. However, these analyses are by no means straightforward. In a strict sense, absolute certainty that two samples are identical can only be inferred from DNA sequence comparison of their entire genomes. For obvious practical reasons, analyses are restricted to a limited number of markers, which can only prove that samples are different, but not identical. Sexually reproducing species consist of heterogenous populations and, even for self-fertilizing species, intra-accession variation is often observed (e.g. van Treuren and van Hintum 2001). Therefore, accessions will rarely be found identical, and "redundancy" analyses must make decisions on the degree of difference needed to justify discarding very similar accessions. For this reason the term "redundancy" is more appropriate than "duplicate". Think for instance of an outcrossing population that is independently sampled twice and analyzed with molecular markers. Unless the entire population is sampled twice and no scoring errors are made, the probability that the two samples are completely identical will be close to zero. However, the two samples may be very similar and one sample may be considered redundant. The question then is how similar two samples should be to consider them redundant, or how different two samples should be to consider them not redundant. The use of statistical theory may help to decide whether accessions are sufficiently different and to obtain probability estimates of making incorrect decisions in rationalization (e.g. Excoffier et al. 1992).

Another consideration is that neutral marker data do not necessarily reflect variation in functional characteristics. Despite identical marker profiles, accessions may still differ in important phenotypes, such as a disease resistance. Therefore, in redundancy studies, marker data should preferably be used in conjunction with passport, morphological and evaluation data. If redundancy studies are carried out mainly for economic reasons, the costs to perform the molecular analyses should

be evaluated against the expected benefits of a smaller collection. For example, for crops that can be regenerated relatively easy and that can be stored for relatively long times at low cost (e.g. maize), the benefits may not outweigh the costs for the molecular analyses (Engels and Visser 2003).

With these caveats in mind, redundancy studies will benefit from marker data. To identify redundancies among rice accessions, potential duplicates selected by examination of passport data were characterized by morphological and/or RAPD analysis (Virk et al. 1995). To reduce the size of a *Brassica oleracea* L. collection, van Hintum et al. (1996) used allozymes to validate bulking of accessions into groups that were formed by crop experts based on historical and morphological data. Phippen et al. (1997) used RAPDs to analyze 14 phenotypically uniform *Brassica oleracea* accessions. Statistical analysis that partitions variation among accessions showed that the accessions could be reduced to four groups with only minimal loss of variation, thereby saving about 70% of the costs per cycle of regeneration. In a similar approach using microsatellites, Dean et al. (1999) achieved a 50% reduction among accessions within a sorghum collection without jeopardizing the overall genetic diversity. Using microsatellites, isozymes and AFLPs, potential duplicates were identified in a cassava core collection (Chavarriaga-Aquirre et al. 1999). Prior to the development of a core collection of the cultivated potato Group Andigenum, morphological characters and electrophoresis of total proteins and esterases were used to reduce a potato collection from 10 722 to 2 379 (Huamán et al. 2000). AFLPs were used to characterize 29 flax accessions belonging to 3 groups based on similar accession names. A procedure called "stepwise bulking" was subsequently used to form significantly different groups containing genetically similar material. To form these groups, an analysis of molecular variance (AMOVA) was used to evaluate within and between accession variation. Using this bulking approach, the 29 accessions could be reduced to 14 with only 2.6% loss of the among-population component of variance (van Treuren et al. 2001). Accessions of the wild potato *Solanum* L. series *Acaulia* Juz. were analyzed with AFLPs, which revealed a redundancy of about 5%. It was estimated that the costs to identify these redundancies were about 2.5 times as high as the savings that could be expected per generation by rationalization of the collection (McGregor et al. 2002; van Treuren et al. 2004). Other studies addressing the use of molecular markers in the identification or validation of redundancies include Tao and Sugiura (1987), Waycott and Fort (1994), Lamboy et al. (1994), van Hintum and Visser (1995), Cao et al. (1998), Cervera et al. (1998), Zeven et al. (1998), Fregene et al. (2000), Negash et al. (2002) and Engels and Visser (2003).

Assembly of core collections

Genebanks are reservoirs of genetic diversity. However, the maintenance of maximum genetic diversity is done at a significant cost. One of the most important purposes for genetic resources conservation is their utilization by plant breeders for germplasm enhancement through variety development. However, the actual clientele of genebank collections is much broader and may include taxonomists, entomologists, molecular geneticists and scientists from many other disciplines (Engels and Visser 2003).

Whatever the purpose for conserving a collection, its size can create difficulties because of costs and inability to evaluate all of the accessions. Without evaluation data, breeders may have difficulty in selecting germplasm for their needs, which may in turn reduce the utilization of collections. To facilitate utilization, core collections have been developed by genebanks, following the concept developed by Frankel (1984). Core collections would represent, with a minimum of repetitiveness, the genetic diversity of a crop and its relatives. The remaining accessions would be maintained as a reserve collection. The basic idea is to subject the smaller core collection (e.g. 10% of the entire collection) to intensive characterization and to maintain the reserve collection (the remaining 90%), as a back-up of potential, but largely unevaluated variability for future use if new or extraordinary needs arise, such as emerging diseases.

The core collection could be constructed in different ways to fulfill varying objectives, from subsets of large genetic resources collections, including diverse arrays of taxa, to smaller collections reflecting particular priorities within a given genepool mostly addressing user requests. Core collections should in any case integrate a representative sample of the variation within the crop of interest, including its wild relatives, with a minimum of redundancy (Hodgkin et al. 1995; van Hintum et al. 2000).

The concept of core collections was further developed by Frankel and Brown (1984) and Brown (1989a,b). Traditionally, the core has been constructed by a variety of taxonomic, morphological, agronomic and ecogeographical criteria. Often, after a core collection has been completed by any of the traditional criteria, biochemical and/or molecular data have been later applied either to validate the strategy adopted or to verify the amount of genetic diversity gathered in comparison with the original collection. For example, variation at 3 morphological trait loci, 17 isozyme loci, and an amylase gene in a collection of lentil was determined to compare the usefulness of 2 different sampling methods (Erskine and Muehlbauer 1991). A set of 19 isozyme profiles, C-banding variants

in all chromosomes, 3 morphological descriptors, and phenotypic patterns for hordein and DDT-susceptibility were used to validate the selection of a core collection based on pedigree data against a random strategy (van Hintum 1994). Some core collections have been established exclusively on molecular marker data. For instance, Ghislain et al. (1999) set up a cultivated potato core collection based on RAPD data. Marita et al. (2000) also used RAPDs to develop a core collection of cacao and pepper.

The success of alternative methods for composing a core collection has been evaluated by quantifying isozyme variation, with the assumption that the larger the number of alleles in the core collection, the more successful the method (van Hintum et al. 1995). Simulations have also addressed the usefulness of utilizing marker data as a selection criterion. For example, methods based on variation at isozyme and RFLP loci were compared with marker-independent methods and showed that marker-assisted methods were able to gather higher allelic richness (Schoen and Brown 1995). AFLPs were used to evaluate the genetic structure between and within gene pools of a wild bean core collection and the enhanced power of molecular markers was contrasted with the limitations of previous methods of analysis (Tohme et al. 1996).

Molecular markers have also been used to compare the genetic diversity of the core and reserve collections of common bean (Skroch et al. 1998, using RAPDs); potato (Huamán et al. 2000, using isozymes); sorghum (Grenier et al. 2000, using SSRs); sandalwood (Shashidhara et al. 2003, using RAPDs). In all cases, no significant genetic differences were found between core and reserve samples, indicating that the core collection formed a representative sample of the entire collection. In the bean and potato situations, data were interpreted to suggest that the traditional core collection selection criteria captured the most representative sample of genetic variation. In the sandalwood study, the authors concluded that molecular markers were efficient markers for estimating diversity and in the case of sorghum, the results showed that genetic diversity was captured in various sampling procedures. Another class of markers of possible use for core collections are functional diversity markers that may be able to screen for traits, as disease resistance markers (see "Future challenges").

In some instances, the assembly of core collections has been based on a combination of data, including biochemical and molecular markers. Eight polymorphic loci in combination with geographical origin and morphological characters were considered an optimal method to build a core collection of cacao (Ronning and Schnell 1994). The combination of quantitative and qualitative data, including molecular markers,

was shown to be adequate to assemble core collections of a number of tropical crop plants (coffee, rubber tree, rice and sorghum; Hamon et al. 1998) and *Cichorium* L. (Kiers et al. 2000).

The most important issue facing gene banks is that of improving the use of their collections. For plant breeders, a logical question to ask is: are molecular marker data useful to build the core collection if its purpose is the maximization of agronomically important characters? Stated another way, is neutral marker diversity correlated with functional diversity? However, the power of molecular markers to form a core collection to maximize useful phenotypic diversity has not been thoroughly tested. Molecular markers share a number of traits that make them differentially useful depending on their applications. As genotypic markers, they are likely more suited to evaluate genetic distance than to construct core collections that maximize a wide range of functional diversity (as disease resistances and agronomic characters). Most studies suggest that neutral markers have reduced utility for being linked to functional diversity, and therefore of reduced utility for the construction of most types of core collections. For example, an analysis of 71 publications showed only a weak mean correlation between neutral molecular markers and quantitative measures of variation. Furthermore, there was no significant correlation between genetic variation and life history traits or heritability (Reed and Frankham 2001). However, there was a moderate correlation between genetic diversity as measured by neutral molecular markers and fitness, despite theoretical considerations and empirical observations that have suggested otherwise (Reed and Frankham 2003).

Maintenance of the genetic integrity of accessions

Genetic resources conserved as seed populations need to be regenerated at certain intervals to regenerate stocks for distribution and because of loss of viability. Each time a variable accession is regenerated there is a risk that the genetic integrity of the accession is compromised by genetic drift, selection, or gene flow (Sackville Hamilton and Chorlton 1997).

Genetic drift is a stochastic phenomenon that describes fluctuations in allele frequencies in the offspring deviating from the parental population. This may cause random fluctuations in allele frequencies from generation to generation and may eventually result in the loss of alleles from accessions. For diploid organisms the change in allele frequency in one generation equals $q(1-q)/2Ne$, in which q represents the frequency of allele q and Ne the effective population size (Falconer 1981). Thus, the extent of genetic drift increases with reduced effective population size. The effective population

size represents the size of an idealized population in which all individuals have an equal probability to contribute gametes to the next generation.

Disparity between actual and effective population size occurs when only a fraction of the population participates in reproduction; for example, when not all plants are flowering. Effective population sizes are generally smaller than actual population sizes. Factors that lower Ne include unequal numbers of females and males, overlapping generations, non-random mating, differential fertility and fluctuations in population size (Falconer 1981; Barrett and Kohn 1991). In the regeneration of accessions the question is not so much whether genetic drift occurs, but rather to which extent. Potential measures to control genetic drift include adjustment of the size of the regenerating population or the development of improved regeneration methods (Engels and Visser 2003). Genetic drift influences all loci simultaneously and to the same extent. The neutrality of markers is an advantage in studies on genetic drift because of the random nature of the process. Co-dominant markers are to be preferred in assessments of genetic drift because they allow accurate estimation of allele frequencies.

During regeneration, the genetic composition of populations may also change due to selection. Selection towards particular genotypes may occur when variation in flowering time occurs and seed harvesting is performed only once. Accessions may become adapted to the circumstances under which they are regenerated and deviate from allele frequencies adapted to their natural habitats. The difference between selection and genetic drift is that selection is not random and does not affect all loci simultaneously. Therefore, selection during regeneration can be inferred from strong shifts in marker frequencies for certain loci between parental and offspring populations. Obviously, the absence of such observations cannot be taken as evidence that selection has not occurred because selection may act on a single gene or on a few genes that are not targeted by the markers.

Multiple accessions of a crop are usually regenerated simultaneously. Unless accessions are regenerated in complete isolation, gene flow between accessions may occur. Gene flow may occur by means of pollen or through seed contamination after harvesting. The probability of pollen flow between populations strongly depends on the mating system of the species. To prevent gene flow between outcrossing accessions, regeneration may be carried out in isolation, e.g. in isolation chambers or by using different crops to separate populations in the field. Molecular markers may be used for monitoring contamination of accessions or for evaluating the effectiveness of measures to prevent contamination. Gene flow

between populations can readily be studied with neutral, preferably co dominant, markers that are diagnostic for different populations. The probability of detecting gene flow between populations very much depends on the extent of genetic differentiation between the populations.

The objective of genebanks is to maintain the genetic integrity of their accessions as much as possible. The loss of genetic integrity from variable accessions will always occur, and the best genebanks can do is to reduce it as much as possible. Molecular markers are useful tools to evaluate and optimize the efficiency of the regeneration protocols. Reedy et al. (1995) used isozymes to investigate allele frequencies in maize accessions following several cycles of regeneration. Significant changes in allele frequencies over generations were observed that were attributed to genetic drift because of the absence of any linear trend. Del Rio et al. (1997a) used RAPDs to measure the loss of diversity in genebank samples of wild potatoes after one to four cycles of seed increase. The majority of the populations showed no significant or only very little loss, and they suggested that the seed increase methodology they used (using 20 individuals and pollinating from bulked pollen) was an appropriate seed increase strategy. Del Rio et al. (1997b) used RAPDs to measure genetic differences between genebank samples subjected to one cycle of seed increase and recollections from the original site of collections in the wild. These collections showed significant differences that could be due to genetic drift, gene flow from adjacent populations, or differences in sampling in the wild. Parzies et al. (2000) used morphological and isozyme markers to analyze diversity within barley landraces stored in genebanks for 10, 40 and 72 years. Data were compared with recent collections of the same landrace. Severe declines in genetic diversity of the landraces were observed with length of time in storage, as well as a strong increase of the level of genetic differentiation among accessions over time. Thus, not only was genetic variation lost from accessions, but populations also differentiated in genetic composition over time. If the accessions are regenerated about once every five years it was estimated from the genetic data that the effective population size over the period of seed storage was only 4.7. By means of allozyme analysis Wagner and Allard (1991) presented evidence of long-distance pollen migration in predominantly selfing barley, affecting the genetic integrity of pedigree stocks and experimental populations. Börner et al. (2000) investigated wheat accessions regenerated 24 times under ex situ conditions using microsatellite markers. No pollen or seed contamination was detected for any of the accessions, whereas genetic drift was observed in one case. Other studies using genetic markers to address questions about the genetic integrity of accessions include

Spagnoletti-Zeuli et al. (1995), Penteado et al. (1996), Schittenhelm et al. (1997) and Steiner et al. (1997). In summary, changes in the genetic constitution of accessions as a result of regeneration can readily be studied with molecular, preferably co-dominant, markers. Resulting data are useful to genebank curators in evaluating and improving their regeneration protocols in order to minimise loss of genetic integrity as much as possible.

Utilization of genetic resources

Germplasm base broadening

The replacement of traditional agricultural systems by modern industrial methods and the introduction of modern high yielding varieties have been considered to decrease crop diversity relative to landrace or wild species progenitors. Although some studies suggest that this reduction in genetic diversity may be less severe than previously assumed (Petersen et al. 1994; Struss and Plieske 1998; Backes et al. 2003; Khlestkina et al. 2004), broadening of the genetic base of crops may be essential in the development of new varieties. The striking success of the "Green Revolution" was based largely on introgression of genes for reduced plant height and increased resistance to diseases in wheat and rice. This success led to a widespread endeavour to assemble genetic resources in genebank collections for the present and future use of mankind. Much progress of plant breeding was possible thanks to developments such as advances in cell and tissue culture, embryo rescue, identification of somaclonal variation, protoplast fusion, double-haploids and genetic transformation (Koornneef and Stam 2001). A new era started in the 1970's with the development of genetic markers, initially biochemical and then molecular, and the possibilities they provide to assess genetic variation. DNA markers simplified the construction of genetic linkage maps, the identification of single-gene traits, quantitative trait loci, and application of molecular breeding strategies through "marker-assisted" introgression and selection.

Markers have been used to infer the origins of crop species and identify ancestors that could be sources of interesting traits. Maize and its wild relatives, i.e. teosinte, were analyzed at 21 allozyme loci to identify teosinte as the progenitor of maize and offered new perspectives for the improvement of maize with introgressed teosinte germplasm (Doebley et al. 1984). RFLPs have been used to confirm gene introgression of somatic hybrids of potato and one of its non-tuber-bearing wild relatives (Watanabe et al. 1995). RAPDs were successfully used for a genetic diversity study of faba bean germplasm aiming at the identification of heterotic groups for broadening the genetic base of the crop (Link et al. 1995). Allozymes and RFLPs were used to investigate

the success of hybrids recreating rapeseed from its parents as a source for broadening the genetic variation in *Brassica napus* L. (Becker et al. 1995). A PCR-based DNA marker able to detect a wide-compatibility allele in rice was found to be useful to facilitate the exploitation of crosses between different subspecies of *Oryza sativa* L. germplasm sources (Williams et al. 1997).

Introgression of desirable traits from exotic germplasm in traditional breeding approaches requires several back-crossing generations. Beckmann and Soller (1986) calculated the frequency of a favourable allele after several backcross generations with and without the help of a linked marker. They found that the use of a marker could improve efficiency of introgression ten-fold. RFLPs were used to determine whether unadapted landrace material persisted during selection for early flowering following introgression into maize (Koester et al. 1993). A codominant RAPD marker linked to a resistance gene for the bean golden mosaic virus was identified and proved useful to expedite the rapid introgression of partial resistance into susceptible germplasm of common bean (Urrea et al. 1996).

Both a CAPS and a microsatellite marker were developed for marker-assisted introgression of a locus encoding resistance to different strains of the Barley Yellow Mosaic virus (Graner et al. 1999). The availability of an RFLP genetic linkage map made possible the evaluation of several quantitative trait loci in an artificial cross of tomato, involving the wild relative *Solanum pennellii* Correll, and led to the understanding of the basis of transgressive segregation in a wide cross. The same study revealed the power of molecular markers to identify useful QTLs in wild species for traits in which they were considered phenotypically inferior in comparison with the cultivated species (de Vicente and Tanksley 1993). This research emphasized the importance of genetic rather than phenotypic traits in crop improvement. This evidence was then successfully pursued using a range of wild tomato species, while improving methodologies and similar approaches in other crops (Eshed and Zamir 1995; Eshed et al. 1996; Tanksley and Nelson 1996; Xiao et al. 1996; Bernacchi et al. 1998). These studies demonstrated the ability to identify accessions with DNA markers that possessed useful alleles (Tanksley and McCouch 1997). Determining the genotypes of accessions in a genebank collection may improve the choice of specific alleles to greatly improve the likelihood of uncovering novel and useful alleles.

A strategy was proposed to develop exotic libraries that are collections of elite crop lines containing defined genomic regions from wild species, to provide pre-breeding material for modern varieties (Stuber et al. 1999; Zamir 2001). The procedures involve

the construction of near isogenic lines (NILs), covering the donor genome, in a process monitored with molecular markers, in which sequential segments of the genome are replaced by the segments originated in the donor genome. These collections could be very useful for gene discovery and characterization, in particular for quantitative trait loci, and may represent an excellent strategy to exploit germplasm of genetic resources.

Molecular breeding has many challenges (Young 1999). Many markers may not be sufficiently close to the genes of interest to aid effective introgression; depending on the question being addressed, marker coverage of the genome might be insufficient; poor parental sources may be chosen; and when pursuing the location of QTLs, poor phenotyping and lack of sufficient replication trials may hamper the results. Tanksley and Nelson (1996) identified two reasons for the reduced success in the development of new varieties with important QTLs: 1) QTL mapping and variety development are usually two separate activities; and 2) most QTL studies have focused on elite materials. Stuber et al. (1999) indicated that the use of QTLs will depend very much on the genetic background where they were found and the need for their proper screening.

The development and applications of modern techniques will play a significant role in uncovering and using genetic variation. Markers have been identified and successfully used to introgress simple and complex traits. More research is needed to understand the genetic basis underlying complex variation. Ultimately, plant scientists will only be able to use genetic variation if it is properly maintained and accessible. With sufficient knowledge on the experimentation needed, genebanks may play a key role in variety development through pre-breeding.

Identification of relevant diversity for specific traits

The role of markers for identifying and transferring useful genes from germplasm to crops has already been mentioned. However, genebanks are often not linked to breeding programmes, or they may not have the capacity to conduct pre-breeding. Here we describe the use of molecular markers linked to agronomically important traits to screen germplasm for these traits. The value of these types of markers will depend on how well they uncover sufficient polymorphism in different genetic backgrounds. Rick and Fobes (1974) found a specific acid phosphatase allele (Aps1) in nematode resistance stocks, after testing a wide array of tomato cultivars for isozyme patterns. The particular allozyme was missing from other tested tomato germplasm and the test of segregating populations showed that only +/+ individuals were nematode susceptible. This

meant that an exceptional codominant marker had been detected, which could be efficiently used to screen for nematode resistance in germplasm without the need to expose materials to the parasite. However, given that the source of the resistance was the wild species *Solanum peruvianum* L., the usefulness of this marker would be mostly limited to breeding material that used *S. peruvianum* in its pedigree. A similar case is that of a shikimate dehydrogenase allozyme, which was found to be an effective marker in screening rice germplasm for high seed protein content (Shenoy et al. 1990). Tang and Zhang (1992) found a peroxidase enzyme marker linked to dwarfness that could select for this trait when trees were only two months old. More recently, a preliminary RAPD screening was used to identify markers linked to Fusarium wilt resistance in chickpea. Allele-specific associated primers were then developed for susceptibility of race 1 of Fusarium wilt (Mayer et al. 1997).

These linked markers may be useful to structure a collection according to a certain genotypic composition. Another alternative would be to characterize germplasm for diversity at specific gene loci. In both cases, genebanks would offer potential clients, whether plant breeders or the entire plant scientific community, value-added germplasm with will-characterized traits.

Manipulation of the phenotype is improved by identification of the gene controlling the character and also knowledge of its gene action and interactions. The study of variants in genes will eventually lead to the deciphering of gene function and might facilitate the management of genes in approaches tailored to solve specific agricultural challenges. Modern technologies and research currently focus on the diversity at its the DNA sequence level. Genome sequencing data will assist in the search for putative genes of agronomic importance. Since the genomes of all species will not be sequenced because of cost, comparative mapping and comparative genomics provide other options. Comparative mapping makes it possible to map the presence of a putative gene of interest in a related species. Markers derived from the known sequence of a gene in a related species may be used to identify the gene in the lesser-known species through the use of comparative genomics (see also "Future challenges").

Such genomic approaches have already been made. Primers were constructed from conserved regions of the leghemoglobin gene, a potentially important gene for enhanced nitrogen fixation. These primers had complete homology with the leghemoglobin-encoding genes of common bean, two leghemoglobin genes of soyabean and 90% homology with a third gene of soyabean. Grouping of eleven species of *Phaseolus* L. was possible based on the product of the PCR

amplification, while intraspecies variation could only be observed for *P. acutifolius* A. Gray and *P. angustissimus* A. Gray. (Skroch et al. 1993). Plant disease resistance genes have already been isolated, and many show common structural features in nucleotide binding sites and leucine-rich repeats. A study of molecular diversity was conducted with lettuce germplasm, both wild and cultivated, using markers derived of the NBS-LRR resistance gene type (Sicard et al. 1999). One microsatellite marker and two SCAR markers from the LRR-encoding regions were developed and were shown to be highly correlated with resistance phenotypes in *L. sativa* L. In addition, several haplotypes indicated the presence of numerous resistance genes in wild species. There are many copies of this type of resistance gene homologue in plant genomes, and they can be used as candidates at QTLs for resistance to plant diseases (Pflieger et al. 1999; Geffroy et al. 2000). Based on the sequence similarities of several plant resistance genes available, Chen et al. (1998) used denaturing polyacrylamide-gel electrophoresis to detect PCR products of genomic DNA amplified with primers designed from conserved regions of those genes. Because segregation analysis of the PCR products in breeding populations showed linkages with known resistance genes, the authors concluded that the markers developed could be used to assess genetic diversity based on candidate genes for resistance.

Different approaches aim at tackling the discovery of genetic variation within coding sequences in order to identify allelic differences responsible for phenotype, such as allele mining, comparison of EST sequence data, identification of single-nucleotide polymorphisms, correlation of SNP haplotypes with phenotypes, use of conserved ortholog gene primers in different species, etc.; see also "Future challenges". However, recent investigations show that changes in the phenotype of a character are not necessarily due to changes in the DNA sequence, or the structure of the protein they encode, but rather in the regulatory elements that affect the expression of the gene (Frary et al. 2000). Pursuing research in many of these areas will be beyond the role of most genebanks, but this may change as methodologies become more routine and cost effective.

Streamlining procedures and goals among cooperating genebanks

Experimental procedures to determine marker variation may vary between different laboratories and comparison of data can be problematic. However, international cooperation between genebanks is indispensable in achieving maximum efficiency in the management of genetic resources. Such cooperation includes the exchange of methodologies and technologies to research, document, manage and

utilize genetic resources. Since the diversity of many crops is typically distributed among many genebanks, the ability to compare different collections is important for investigating the extent of diversity of a crop and of the overlap existing between genebanks. Due to environmental effects, comparing accessions by morphological analysis at different locations has obvious limitations. In contrast, molecular markers can overcome such environmental effects and are particularly useful for these investigations. However, even though the majority of molecular markers are quite robust, several methodological issues need to be carefully considered in order to obtain results that can be compared between different laboratories.

In a small-scale study, Jones et al. (1997) compared the reproducibility of RAPDs, AFLPs and SSRs between different European laboratories through an examination of small numbers of poplar, sugar beet and tomato samples involving nine different laboratories. Molecular analyses were carried out on DNA samples extracted by a single laboratory and optimal protocols were used among participating laboratories. None of the laboratories was able to reproduce the RAPD profile exactly, but AFLP and SSR profiles were highly reproducible.

Bredemeijer et al. (2002) examined 521 tomato varieties using 20 microsatellites, carried out by a consortium of 5 laboratories using different detection methods. Each laboratory used the same set of reference alleles, predefined in a previously standardized experiment with 22 tomato varieties. No discrepancies between laboratories were observed for 361 varieties (70%), whereas 136 varieties (25%) showed discrepancies at 1 or 2 loci, and 24 varieties (5%) showed discrepancies at more than 2 loci. The majority of the discrepancies could be attributed to heterogeneity of the seed sample and only few discrepancies were ascribed to methodological differences, e.g. the use of different selection thresholds for allelic peaks. It was concluded that microsatellites could be used successfully to construct databases containing molecular data generated by different laboratories, provided that attention is paid to methodological considerations, including the careful selection of markers, the duplication of analyses in at least two laboratories, and knowledge of possible heterogeneity of seed samples.

In a similar study, Röder et al. (2002) characterized 502 European wheat varieties with 19 microsatellites. Data were collected in duplicate in at least two laboratories using different experimental procedures. Out of the 11 080 data points generated, 34 discrepancies remained unsolved, revealing an accuracy of more than 99.5%. The use of reference alleles was indispensable in achieving correct allele identification. Although studies on the reproducibility of marker data in network situations are still scarce, microsatellites appear to be one of the most repeatable marker technologies in collaborative projects.

Crop Breeding

Parental contributions of artificial hybrids

DNA markers are very useful for confirming hybridity of artificial sexual hybrids or somatic fusion hybrids. Molecular markers are especially useful when hybridity is questioned by morphological reasons or for early screening of large putative hybrid populations. Thieme et al. (1997) used isozyme and RAPD data to screen hundreds of first generation somatic fusion hybrids and later sexual backcross generations between cultivated and wild species of potato, to distinguish hybrid from non-hybrid progeny. Parani et al. (1997) also used isozymes and RAPDs to confirm hybridity of F_2 generations of *Sesamum* L. where the identity of hybrids was questioned on morphological grounds. Lee et al. (1998) used RAPDs to disprove hybridity in putative *Saccharum* L. and *Erianthus* Michx. hybrids. Durham and Korban (1994) used RAPDs to identify apple clones with introgressed genes from a wild apple species that was used to transfer apple scab resistance through many cycles of introgressive hybridization. Oberwalder et al. (1997) used RFLP and SSR probes to investigate the genome contribution to asymmetric somatic fusion hybrids. These differ from symmetric somatic fusion hybrids by the elimination of part of the genome of one of the parents, through irradiation or chemical treatments, before fusion, in order to try to eliminate undesirable traits. They found that RFLPs were better able to distinguish symmetric from asymmetric fusion hybrids. Yamada et al. (1997) used cpDNA and mtDNA to investigate symmetric somatic fusion hybrids of cultivated potato and one of its wild relatives. The results suggested that chloroplast genomes of the respective parents segregated randomly while the mitochondrial genomes favoured the cultivated species parent. Provan et al. (1996) used SSRs to show how somatic fusion hybrids could be rapidly screened at the callus level to quickly and unambiguously distinguish fused from non-fused products.

Chromosome substitution lines are genetic stocks differing by deleted chromosomes, and are used in genetic studies of the inheritance of quantitative traits. Korzun et al. (1997) used mapped SSRs to authenticate different substitution lines. The high polymorphism present in SSRs, in contrast to isozymes and RFLPs, made them ideal markers for this purpose. Crouch et al. (1998) used SSRs to investigate the genetic constitution of tetraploid hybrids in banana. Prior breeding schemes did not appreciate the occurrence of diploid gametes that would allow a broader range of breeding strategies, and SSRs were used to infer the ploidy level of the gametes leading to these hybrids.

Geneflow between crops and weeds

Gene flow among crops and weeds has long been considered a common occurrence in nature, and is a continuing source of useful new variation in landraces (van Raamsdonk and van der Maesen 1996). Artificial gene exchange also forms the basis of modern plant breeding which is replete with wide crosses, sometimes even between genera (Harlan and DeWet 1971). In 1996, the first commercial transgenic crops were introduced after review and approval from governmental regulatory agencies in the US and around the world. Since then, gene flow has become a much more topical issue. The US National Research Council (2002) concluded that transgenic plants are not "different in kind" from crops bred by other means, but contend that certain transgenes may be novel and therefore a special source of concern. Molecular markers are key tools to investigate gene flow, with implications for a wide array of economic, environmental, food safety and social concerns.

Environmental concerns relate to possible increased weediness and the survival of rare populations. For example, Ellstrand et al. (1999) conclude that gene flow from traditionally bred crops to weedy relatives has been implicated in the increased weediness in wild relatives of 7 of the 13 most important crops worldwide. They cite one extreme example of a weedy rye derived from natural hybridization between cultivated rye and a wild rye species. In California the weedy hybrid is so dominant that farmers have abandoned efforts to grow rye there for human consumption (National Academy of Sciences 1989). Another concern is the possible extinction, or reduction in fitness, of wild local populations resulting from transgenic gene flow (Ellstrand and Elam 1993; Levin et al. 1996; Rhymer and Simberloff 1996). Crops often contain genes that theoretically reduce fitness of individuals within populations in the wild, and many crops do not survive long in the wild. Extinction of related wild populations may result from outbreeding depression - a reduction in fitness following hybridization among individuals in populations (Templeton 1986; Waser 1993). Another concern is "swamping" of locally rare species with transgenes through repeated bouts of introgression (Ellstrand and Elam 1993). For example, hybridization was documented between cultivated rice and Taiwanese wild rice, resulting in a progressive loss of wild rice traits in native populations in Taiwan (Kiang et al. 1979). In less than 80 years, gene flow from cultivated rice, coupled with environmental factors, drove the populations of Taiwanese wild rice to near extinction.

Studies measuring pollen movement out of a source planting into larger surrounding fields indicates that pollen moves very short

distances, typically 5–20 m (e.g. Conner and Dale 1996; Llewellyn and Fitt 1996). Pollen sink studies, however, show vast underestimations of such gene transfer, with pollen measured at distances well over 1000 m or more (e.g. Arias and Rieseberg 1994; St. Amand et al. 2000; Rieger et al. 2002). However, these pollen flow studies are short-term, and gene flow is also possible by transgenics remaining in a field through the soil seed bank, and seed movement through transport from field to market. Individual studies and reviews are beginning to accept the reality of genes frequently being dispersed from crop to crop, and crop to weed, assuming compatible recipient wild and cultivated species are present within some reasonable distance, sometimes measured to distances over 1 km (Dale 1992; Hancock et al. 1996; Snow and Palma 1997; Linder et al. 1998; Ellstrand et al. 1999; Ellstrand 2001; St. Amand 2004). Rieseberg and Burke (2001) argue that gene flow is not as important as selection coefficients in determining maintenance of a gene in nature, and suggest that advantageous transgenes will likely be maintained.

Following are examples of gene flow studies facilitated by molecular markers. Rabinowitz et al. (1990) tested hypotheses of gene flow between wild and cultivated potato. By use of population-specific isozyme markers they were able to document high levels of natural gene flow in experimental field plots in the Andes. They used these data to suggest that there was extensive gene flow among other cultivated and wild species. Skogsmyr (1994) reported high frequencies of transgenic pollen dispersal up to 1000 m from a field trial of transgenic potatoes. Conner and Dale (1996) pointed out procedural problems in the design of the experiment and questioned if the geneflow was artefactual; the question remains unresolved. However, potatoes are pollinated by buzz-pollinators, and gene flow from such insects is possible in other insect pollinated species.

Beets are self-incompatible wind pollinated species and therefore obligately outcrossing. Weed beets pose a serious problem for sugar beet cultivation, and traditionally have been controlled only by manual removal. Transgenic herbicide resistant beets have been considered as a way to control this problem. Desplanque et al. (1999) showed, through the use of RFLP and microsatellite markers that weed beets in northern France were intermediates between cultivated sugar beets and wild beets in southwestern France. They attributed the origin of weed beet infesting cultivated sugar beet fields to accidental and recurrent hybridization between cultivated beet and wild beets during the production of commercial seeds in southwestern France. Desplanque et al. (2002) showed that herbicide resistant sugar beets could transfer their herbicide resistance to co-occurring weed beet populations, rapidly reducing

the effectiveness of the transgene. Saeglitz et al. (2000) also showed wide pollen dispersal in sugar beets (greater than 200 m), even when "containment" border plants (in this case hemp) were used, through the use of cytoplasmically male sterile receptor plants, and PCR screening with probes specific to a transgene. Bartsch and Ellstrand (1999) investigated the origin and gene flow among cultivated and weed beets from California, and showed, through allozyme analysis, germplasm conforming to cultivated beet, sea beet (different subspecies of *Beta vulvaris* L.) naturalized *B. macrocarpa* Guss., and evidence of gene flow among all types.

In contrast to beet, cultivated barley is a self-pollinating species and would be expected to have less transgene escape. Ritala et al. (2002) measured, via PCR screening with probes specific to a transgene, pollen-mediated dispersal of barley transgenes via cross-fertilization in barley at distances of 1, 2, 3, 6, 12, 25, 50 and 100 m from the donor plots. The number of seeds obtained from male-sterile heads diminished rapidly with distance and only a few seeds were found at distances of 50 and 100 m. Molecular genetic analysis revealed that all seeds obtained from male-sterile heads at a distance of 1 m were transgenic, as anticipated. However, only 3% of the distant seeds (50 m) actually carried the transgene, while the remaining resulted from fertilization with non-transgenic background pollen. This background pollen was mainly due to pollen leakage in some male-sterile heads. In normal male-fertile barley, the cross-fertilization frequency with transgenic pollen varied from 0 to 7% at a distance of 1 m, depending on weather conditions on the heading day. They concluded that because of competing self-produced and non-transgenic background pollen, the possibility of cross-pollination is very low between a transgenic barley field and an adjacent field cultivated with normal barley. Adequate isolation distances and best management practices are needed for cultivation of transgenic barley.

The above studies document introgression in early generations. Brubaker and Wendel (1994) document longer-term persistence of cultivar genes into wild populations of cotton. Linder et al. (1998), using RAPDs, documented the long persistence (perhaps up to 40 years) of cultivated sunflower specific markers in adjacent wild populations of this species.

Oilseed rape (both *Brassica rapa* L. and *B. napus* L.) are pollinated by wind and insects. Jørgensen and Andersen (1994) documented hybridization, using isozymes, RAPDs, morphological markers and chromosome counts, between cultivated oilseed rape (*B. napus*) and adjacent populations of *B. campestris* L. The former is an amphidiploid (2n = 38) and the latter is one of its diploid (2n =

20) progenitors. Hybrids were commonly produced, bidirectionally, between both crop and weed. Backcrossing of the hybrids to weedy *B. campestris* was documented with an isozyme marker, supporting an avenue for transgene escape.

Timmons et al. (1995) detected *Brassica napus* pollen distributed 2.5 km from its source. Natural hybrids between *B. rapa* and *B. napus* are documented in the British Isles (Harberd 1975). Field studies show these species to readily hybridize (Jørgensen et al. 1996). Extensive hybridization of these 2 species, in a field where they co-occurred for 11 years, was supported by AFLP data (Hansen et al. 2001). Rieger et al. (2002) screened 48 million individual canola (*B. napus*) plants. This was possible with a transgenic canola line resistant to acetolactate synthase (ALS) inhibiting herbicide. Seeds were collected from 63 conventional canola fields growing near herbicide-resistant fields in New South Wales, Victoria and South Australia. At crop maturity, 10 stratified samples totalling at least 100 000 seeds were taken from each of three locations in each field of conventional canola. These were parallel to the source field and taken at the edge nearest to the source field, the middle, and the edge furthest from the source field. Collected seed samples were screened with a lethal discriminating dose of the ALS. The results show that, in most cases, gene flow via pollen movement occurs between canola fields. However, even adjacent commercial canola fields in Australia will have much less than 1% gene flow. Resistance was detected up to, but not beyond three km from the source.

Despite theoretical concerns about transgene release, in the ten years since transgenics have been used in thousands of products, there has not yet been a confirmed case of a person suffering any ill effects from consuming food derived from genetically modified plants. Also, the only long-term experiment comparing the performance of four transgenic crops (oilseed rape, potato, maize and sugar beet) found no proof that the transgenics are more invasive or persistent than their non-transgenic conventional counterparts (Crawley et al. 2001). This experiment involved field comparisons over 10 years in 12 different habitats. The transgenic traits were the herbicide glufosinate tolerance for oilseed rape and maize, the herbicide glyphosate ('Roundup') tolerance in sugar beet, and two types of transgenic potato expressing either the insecticidal Bt toxin or a pea lectin. It is important to note that these experiments involved transgenic traits (resistance to herbicides or insects) that were not expected to increase fitness in natural habitats, or to increase fitness only somewhat depending on the level of biological control. This is in contrast to other traits that would have greater fitness impact such as cold, drought, or metal tolerance, improved

nutrient uptake, or altered development. These results cannot be extrapolated to all transgenic plants in every environment, but they are noteworthy for the length of the experiment and diversity of crops and habitats. A similar conclusion was made by Salisbury (2000) with oilseed rape.

Autotetraploid vs. allotetraploid inheritance

Autopolyploid refers to the multiplication of the chromosome set from a single species and allopolyploid refers to the multiplication from chromosome sets of different species. Some use the concept to refer to different genetic backgrounds within species as well. Knowledge of genome constitutions and inheritance patterns are essential in designing crosses for genetic improvement. Genetic data from codominant markers can clearly distinguish between allo- and autopolyploidy by progeny classes observed between genetically characterized crosses. A diploid organism with two different alleles at a locus (A and a) produces one heterozygote class (Aa). Selfing of this individual will produce a progeny array of 1AA:2Aa:1aa. However with tetrasomic inheritance, in an autotetraploid, three different classes of heterozygotes can be produced (AAaa), (Aaaa) and (AAAa). Selfing of an individual having the (AAaa) genotype will result in a progeny array of 1AA AA:8AAAa:18AAaa:8Aaaa:1aaaa. In contrast, preferential pairing of genomes in an allotetraploid having AA on one chromosome pair and aa on the other chromosome pair would produce only AAaa progeny resulting in "fixed" heterozygosity that would not be expected in an autotetraploid (Warnke et al. 1998).

Creeping bentgrass (*Agrostis palustris* Huds.) is a major cultivated turfgrass species. It is a tetraploid (2n = 4x = 28), and poorly characterized genetically. Warnke et al. (1998) screened 650 clones of *A. palustris* with 12 enzyme systems to assess polymorphism and selected putative duplex allele states (i.e. AAaa) that provided the clearest differentiation between disomic and tetrasomic inheritance in self and testcross progenies. They provided strong evidence that *A. palustris* is an allotetraploid via segregation patterns at four loci that exhibited sufficiently clear segregation patterns to infer this state.

Molecular diversity and heterosis

Heterosis, or hybrid vigour, originally referred to the selective superiority of heterozygotes regarding continuously variable characters of size, yield and vigour. The concept was expanded to include adaptive, selective, or reproductive advantage (Dobzhansky 1950). The biological basis of heterosis remains unknown, but

Tsaftaris and Shull (1995) summarized studies of RFLP genetic mapping in maize to suggest that a few major qualitative trait loci scattered throughout the genome explain some of the attributes of heterosis. Heterosis is important to breeders because it is commonly assumed that the cross of diverse rather than similar parents enhances breeding success. Much of the genetic characterization of crop germplasm has the practical goal of discovering genetically diverse lines for breeding (Mumm and Dudley 1994; Abo-elwafa et al. 1995; Maughan et al. 1995; Dubreuil et al. 1996; Yang et al. 1996; Menkir et al. 1997). Smith et al. (1990) showed a strong correlation of genetic distance and heterosis in maize. Melchinger et al. (1994) reported that genetic distances of inbred lines of maize, as assayed via RFLP data, were not correlated with heterosis for yield. Paz and Veilleux (1997) reported that in crosses between a diploid cultivated potato species with diploid clones of other complex interspecific hybrids of potato, the greatest total tuber yield was associated with diverse parents as measured by RAPDs. Manjarrez-Sandoval et al. (1997) compared RFLP similarity estimates of the parents, to yield in soyabean, and also found a positive correlation. Diers et al. (1996) examined the correlation of high RFLP variation to heterosis in mustards, and related this to the general combining ability of the parents (the ability of an accession to transfer a trait of interest to a diverse range of progeny). They found that genetic distance was related to heterosis in only some of the crosses, and concluded that genetic distance alone does not identify heterotic combinations consistently. The data, therefore, provide a suggestion of the importance of genetic diversity as measured by neutral markers to heterosis, but apparently the relationship is not constant in all cases and more research is needed to understand when his relationship will occur. The germplasm used in these studies may have a major effect on the correlation of genetic distance and yield. While Smith et al. (1990) used fairly elite (agronomically advanced) germplasm, other studies did not do so, and the effect of germplasm advancement needs to be investigated.

Current developments

The majority of marker technologies outlined in "Overview of molecular technologies", sample random genomic regions except those targeted to specific genes. Because the largest part of the genome consists of non-coding DNA, the sampled diversity through the use of conventional markers in general will be selectively neutral. For some questions facing genebank managers the selective neutrality of a marker is not problematic, and for some purposes it is even to be preferred (e.g. in studying the effect of random processes, such as genetic drift on genetic diversity; see "Crop breeding"). However, the representativeness of these markers to overall diversity remains an open question. Plant breeders are especially interested in diversity in agronomically important characters that may be expected to be influenced by selection. The extent to which neutral markers are representative of functional diversity will strongly depend on the extent to which the underlying genes have been influenced by selection. This may vary from crop to crop and among wild species, landraces and cultivars. Molecular markers that are able to quantify broad sense functional diversity may be very useful for the characterization and optimization of genebank collections.

In addition, knowledge about variation in specific (groups of) genes will contribute to the utilization of genetic resources. During the last decade large amounts of sequencing data have become available for various crops (as rice and tomatoes), and in the case of *Arabidopsis thaliana* (L.) Heynh. the entire genome has been sequenced. Determination of the sequence of expressed DNAs (Expressed Sequence Tags - ESTs) is possible through isolation of messenger-RNA(mRNA) and the construction of complementary DNA(cDNA) libraries. Gene functions may then be assigned to ESTs, for example by means of "differential display techniques" or through matching with known sequences that have already been assigned a function. Massive amounts of sequence data, possibly with genome location and gene function, have been stored in databases that are publicly available. Access to these data enables the exploitation of this knowledge, e.g. for the development of functional diversity markers and the identification of putative genes and the variation therein for traits of interest.

Because genetic resources collections may consist of large numbers of accessions, sample throughput is an important aspect in the application of molecular markers. Even the PCR-based techniques may still be time-consuming, particularly when polyacrylamide

gel electrophoresis, radiolabelling and manual scoring of autoradiograms are involved. Therefore, at the detection level, developments in the field of molecular marker technologies focus on gel-free, non-radioactive and highly automated experimental procedures.

The aforementioned developments in molecular marker applications have great potential for genebanks and are therefore discussed in more detail in the following sections.

Developments in marker techniques

Single Nucleotide Polymorphism (SNP)

The fact that in many organisms most polymorphisms result from changes in a single nucleotide position (point mutations), has led to the development of techniques to study single nucleotide polymorphisms (SNPs). Analytical procedures require sequence information for the design of allele-specific PCR primers or oligonucleotide probes. SNPs and flanking sequences can be found by library construction and sequencing or through the screening of readily available sequence databases. Once the location of SNPs is identified and appropriate primers designed, one of the advantages they offer is the possibility of high throughput automation. To achieve high sample throughput, multiplex PCR and hybridization to oligonucleotide microarrays or analysis on automated sequencers are often used to interrogate the presence of SNPs. Figure 7 outlines the analysis of SNPs using the SNaPshot method. SNP analysis may be useful for cultivar discrimination in crops where it is difficult to find polymorphisms, such as in the cultivated tomato. SNPs may also be used to saturate linkage maps in order to locate relevant traits in the genome. For instance, in *Arabidopsis thaliana* a high-density linkage map for easy to score DNA-markers was lacking until SNPs became available (Cho et al. 1999). To date, SNP markers are not yet routinely applied in genebanks, in particular because of the high costs involved.

Retrotransposon-based markers

Retrotransposons consist of long terminal repeats (LTR) with a highly conserved terminus, which is exploited for primer design in the development of retrotransposon-based markers. Retrotransposons have been found to comprise the most common class of transposable elements in eukaryotes, and to occur in high copy number in plant genomes. Several of these elements have been sequenced and were found to display a high degree of heterogeneity and insertional polymorphism, both within and between species. Because retrotransposon insertions are irreversible (Minghetti and Dugaiczyk 1993; Shimamura et al. 1997), they are considered

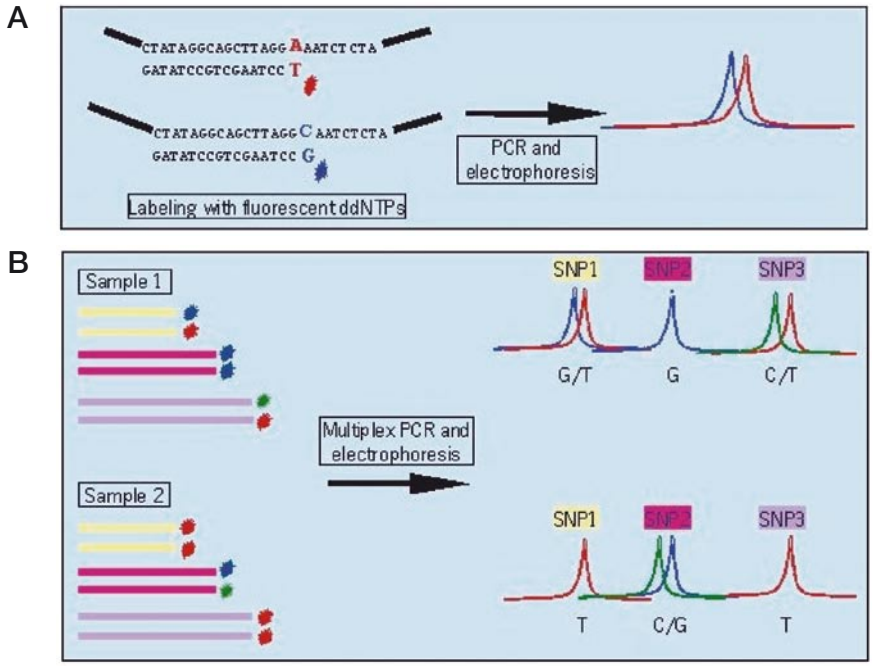


Figure 7. Principles of SNP analysis using the SNaPshot method. (A) Sequence data flanking the SNP are used to design a PCR primer with extension starting at the position of the SNP. The use of dideoxynucleotides (ddATP, ddCTP, ddGTP and ddTTP) ensures that primer extension only occurs at the SNP position. Using differently labeled dideoxynucleotides (A = black, C = green, G = blue, T = red) during PCR, extension products are tested for fluorescent signals by electrophoresis to determine which dideoxynucleotides are incorporated. (B) The use of unique primer sizes for different SNP loci allows multiplexing up to 15-fold during PCR. Nucleotide variation at the SNP position is detected by variation in incorporation of dideoxynucleotides. Note that heterozygotes display two different peaks and hence that SNPs are scored in a codominant manner. By courtesy of Gerard van der Linden (Plant Research International BV).

particularly useful in phylogenetic studies. In addition, their widespread occurrence throughout the genome can be exploited in gene mapping studies, and they are frequently observed in regions adjacent to known plant genes.

Several variations of retrotransposon-based markers exist. Sequence-Specific Amplified Polymorphism (S-SAP) is a dominant, multiplex marker system for the detection of variation in DNA flanking the retrotransposon insertion site. Retrotransposon containing fragments are amplified by PCR, using one primer designed from

the conserved terminus of the LTR and one based on the presence of a nearby restriction endonuclease site. Experimental procedures resemble those used for AFLP analysis and they are usually dominant markers. Compared to AFLP, S-SAP generally yields fewer fragments but higher levels of polymorphism (Waugh et al. 1997).

Inter-retrotransposon Amplified Polymorphism (IRAP) and Retrotransposon-Microsatellite Amplified Polymorphism (REMAP) are dominant, multiplex marker systems that examine variation in retrotransposon insertion sites. With IRAP, fragments between two retrotransposons are isolated by PCR, using outward-facing primers annealing to LTR target sequences. In the case of REMAP, fragments between retrotransposons and microsatellites are amplified by PCR, using one primer based on a LTR target sequence and one based on a simple sequence repeat motif. IRAP as well as REMAP fragments can be separated by high-resolution agarose gel electrophoresis (Kalendar et al. 1999).

Retrotransposon-Based Insertional Polymorphism (RBIP) is a codominant marker system that uses PCR primers designed from the retrotransposon and its flanking DNA to examine insertional polymorphisms for individual retrotransposons. Presence or absence of insertion is investigated by two PCRs, the first using one primer from the retrotransposon and one from the flanking DNA, the second using primers designed from both flanking regions. Polymorphisms are detected by simple agarose gel electrophoresis or by dot hybridization assays. A drawback of the method is that sequence data of the flanking regions is required for primer design. A major advantage of RBIP is that it can easily be automated, using gel-free procedures such as TaqMan™ or DNA chip technology (both explained in "Future challenges") in order to increase sample throughput (Flavell et al. 1998). RBIP markers have been used to characterize genetic diversity of international germplasm collections for pea, barley, tomato and pepper within the context of the European Union project "TEGERM" (http://www.biocenter.helsinki.fi/bi/tegerm/).

Functional diversity markers

The availability of sequence data of expressed DNA has enabled the development of markers that are physically associated with coding regions of the genome. ESTs are the result of sequencing cDNA clones and the information generated is generally stored in databases. These sequences can then be used for designing primers either to readily generate polymorphic markers or as a source of bands for CAPS markers. The raw sequence information will also aid in screening for the occurrence of microsatellite sequences (EST-SSR) or single nucleotide

polymorphisms (EST-SNP), after which markers can be developed that are targeted to transcribed regions of the genome. EST-SSRs have been applied successfully in the characterization of accessions of wheat (Eujayl et al. 2002) and barley (Thiel et al. 2003). EST-SNPs have been used to study functional diversity in maize (Rafalski 2002).

Complementary DNA can also be used as template for subsequent direct marker generation, for example through AFLP technology (cDNA-AFLP). cDNA-AFLP is commonly used in the identification of genetic polymorphisms between contrasting phenotypes under controlled conditions in order to facilitate the construction of linkage maps (e.g. Brugmans et al. 2002), or to identify candidate genes. In diversity studies, the application of cDNA-AFLP should be limited to the identification and comparison of specific gene-related patterns, as in general cDNA differences may be caused by differences in the developmental stage of plants and the environmental conditions rather than by existing DNA polymorphisms.

A simple PCR-based marker technique that targets coding sequences in the genome is Sequence-Related Amplified Polymorphism (SRAP). SRAP uses forward primers consisting of an unspecific filler sequence of ten bases, the sequence CCGG and three selective nucleotides. Reverse primers also contain a filler sequence, but are followed by the sequence AATT and three selective nucleotides. The CCGG sequence is used to target GC-rich regions, such as exons in open reading frames, while the AATT sequence on the reverse primers is aimed at AT-rich regions, such as promoters and introns. The generally conserved nature of exon sequences, combined with the generally variable nature of introns, promoters and spacers, enables SRAP analysis to generate polymorphic bands. The use of selective nucleotides on the PCR primers results in the amplification of subsets of open reading frames that display multilocus band profiles following appropriate labelling of a primer, polyacrylamide gel electrophoresis and autoradiography. In *Brassica oleracea* L. sequence analysis revealed that 45% of the SRAP fragments could be matched with known genes. Furthermore, SRAP analysis could easily be applied successfully to other crops, such as potato, rice, lettuce and apple (Li and Quiros 2001).

A related technique that uses EST sequence information is Target Region Amplification Polymorphism (TRAP). For TRAP analysis, a fixed primer designed from a targeted EST sequence is combined with an arbitrary primer having an AT- or CG-rich core sequence. For different plant species TRAP revealed multiple scorable fragments, and the technique may be well suited for determining the genotypes of germplasm and tagging genes for traits of interest (Hu and Vick 2003).

EST-based, cDNA-based and SRAP markers are common in that they target diversity in coding regions. The gene function of the targeted DNA

will usually be unknown. If gene sequence data are available, markers may be developed for particular (groups of) genes. This is the case, for example, in a recently developed strategy for analyzing groups of resistance genes - Resistance Gene Homologue Polymorphism (RGHP). With this methodology, groups of resistance genes are targeted by PCR using primers aimed at conserved domains of resistance genes, such as the Leucine Rich Repeat (LRR) or the Nucleotide Binding Site (NBS), both involved in resistance mechanisms (Chen et al. 1998; Sicard et al. 1999). In NBS-directed profiling (NBS-DP), one primer is targeted to a conserved sequence of the NBS, while the other primer is based on the presence of a nearby endonuclease restriction site. The highly conserved nature of these targets allows the NBS-DP primers to be used beyond the species level. Because resistance genes have often originated from gene duplications, variation may also be traced in analogs of the gene. Polyacrylamide gel electrophoresis and autoradiography are part of the NBS-DP technique, resulting in AFLP-like banding profiles that are scored as dominant markers. Variable fragments isolated from the gels and sequenced can be subjected to Basic Local Alignment Search Tool (BLAST) analysis in sequence databases in order to search for matches with sequences from known genes. The appealing feature of NBS-DP is that it may also detect variation in resistance genes that were thus far unknown (van der Linden et al. 2004).

A few studies show marker data to be correlated with such functional traits. Research was conducted in lettuce germplasm to evaluate the diversity of wild and cultivated material by means of markers derived from resistance genes of the NBS-LRR type (Sicard et al. 1999). The three markers used were highly correlated with the resistance phenotypes in lettuce and proved useful in differentiating accessions. This shows that markers based on functional genes may have some utility to build core collections, but this needs much further investigation.

Developments in detection techniques

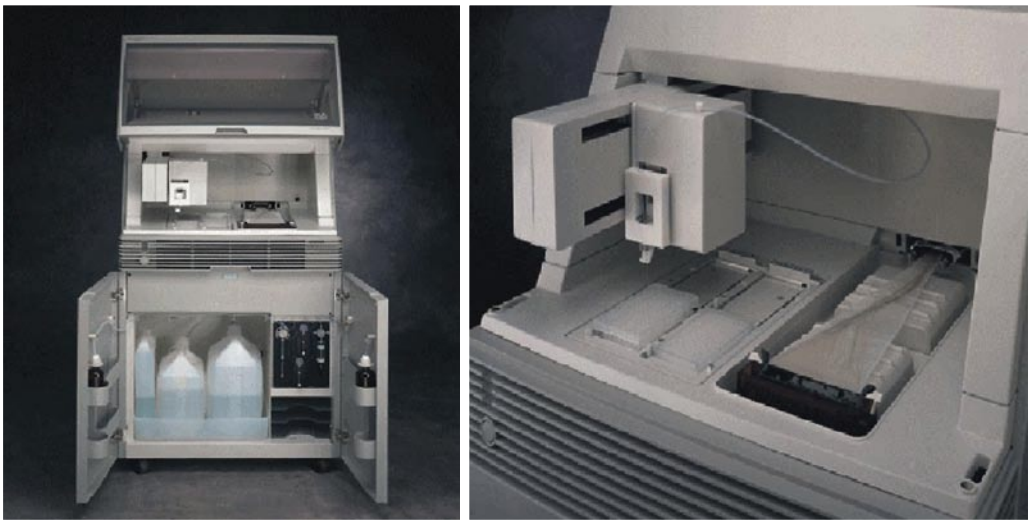
TaqMan™

TaqMan™ is a probe used to detect specific sequences in PCR products by employing the 5′→3′ exonuclease activity of the Taq DNA polymerase. The TaqMan™ probe (20–30 bp), disabled from extension at the 3′ end, consists of a site-specific sequence labeled with a fluorescent reporter dye and a fluorescent quencher dye. During PCR the TaqMan™ probe hybridizes to its complementary single strand DNA sequence within the PCR target. When amplification occurs, the TaqMan™ probe is degraded due to the 5′→3′ exonuclease activity of Taq DNA polymerase, thereby separating the quencher from the reporter during extension. Due to the release of the quenching effect on the reporter, the fluorescence intensity of the

reporter dye increases. During the entire amplification process this light emission increases exponentially, the final level being measured by spectrophotometry after termination of the PCR. Because increase of the fluorescence intensity of the reporter dye is only achieved when probe hybridization and amplification of the target sequence has occurred, the TaqMan™ assay offers a sensitive method to determine the presence or absence of specific sequences. Therefore, this technique is particularly useful in diagnostic applications, such as the screening of samples for the presence or incorporation of favourable traits, the detection of pathogens and diseases in plants and the screening of plant material for the presence of transgenic elements. The TaqMan™ assay allows high sample throughput because no gel electrophoresis is required for detection. When different probes are used which are able to discriminate between allelic variants, TaqMan™ behaves as a codominant marker. TaqMan™ is also referred to as fluorogenic 5′ nuclease assay (Holland et al. 1991; Lee et al. 1993).

Automated sequencers

DNA sequencing was initially carried out through radiolabeling, polyacrylamide gel electrophoresis and visual reading of gels (Figure 1). Throughput of DNA samples has increased substantially by the use of automated sequencers and automated data processing. Prior to running the samples on automated sequencers, tissue preparation, fluorescent labelling and PCR need to be performed. Depending on the type of sequencer used, preparation of polyacrylamide gels and manual loading of samples onto the gels may still be required. More advanced models do not require the preparation of gels. For example, in the ABI Prism 3700 DNA analyzers, fluorescently labelled samples are transferred by a robot from a microtitre plate to an electrophoresis chamber and run through a capillary system (Figure 8). A laser detection system measures the light emission of samples, either during individual or simultaneous electrophoresis. The 4300 System by LI-COR is a third generation instrument based on highly sensitive infrared fluorescence detection technology. The throughput of sample analysis is greatly enhanced compared to the manual methods, and can range from 450 000 bases to 2.8 million bases in a 24-hr day (for example, in a MegaBACE 4000 instrument). Apart from sequence analyses, automated sequencers are also being used for the analysis of microsatellites, AFLPs and SNPs. The high throughput capacity of these machines is achieved by the improved mechanical operation and detection systems, but also by allowing multiplexing of PCR reactions. Collected electronic data may then be processed with dedicated software packages (Figure 2).



(A) (B)

Figure 8. (A) ABI Prism 3700 DNA analyzer with opened doors and hood. The bottom part consists of a storage facility for water, buffers, etc. The opened hood gives access to the work surface. (B) Close-up picture of the work surface with opened electrophoresis chamber. Microtitre plates are loaded at the left, and the capillary system is located at the right. The robotic arm positioned above the work surface transfers the samples from the plates to the capillary system. At the end of the capillaries an optic device measures the light emission of the samples. By courtesy of Applied Biosystems.

Microarray or DNA chip technology

Microarray or DNA chip technology is a high throughput screening technique based on the hybridization between oligonucleotide probes (genomic DNA or cDNA) and either DNA or mRNA. Chips may consist of arrays of amplified DNA immobilized on miniature glass or nylon substrates that are then tested by hybridization to a series of fluorescently labelled probes. More commonly however, arrays of oligonucleotide probes are synthesized (e.g. Affymetrix GeneChips), followed by the exposure to fluorescently labelled PCR samples. Hybridization signals are determined by laser technology from which data on sequence variation is obtained (Figure 9). Microarrays now may comprise up to 250,000 features per square centimetre; however, technical advances are likely to further increase the extent of miniaturization in the future. Applications of the DNA chip technology include diagnostics, mutation and polymorphism detection, gene discovery, gene expression and gene mapping (Lemieux et al. 1998; Ramsay 1998; Gibson 2002).

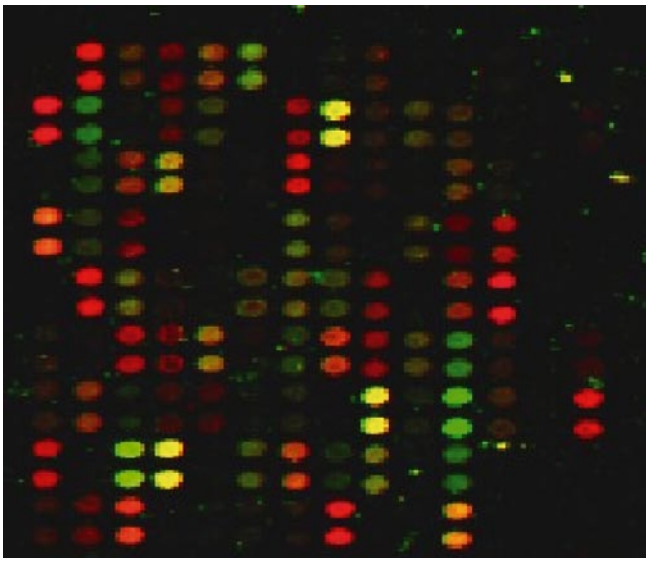


Figure 9. Example of the results from the differential display technique using DNA chip technology. A series of DNA probes are tested for hybridization to DNA samples with different fluorescent labels. Differences in gene expression between the samples are derived from determination of the red:green ratio. By courtesy of Asaph Aharoni (Plant Research International BV).

Wide application of DNA chips, such as the detection of SNPs between genotypes, has not yet found its way to genebanks, particularly because of the laborious sequencing methods and high costs involved. However, a recent development in the application of DNA chips to the analysis of genetic polymorphisms is the Diversity Array Technology (DArT) introduced by Jaccoud et al. (2001). First, a so-called 'diversity panel' is created consisting of arrays of DNA fragments originating from a group of diverse genotypes. To construct a diversity panel, total genomic DNA is isolated from a group of genotypes, after which the DNA is pooled and digested by restriction enzymes. Enzyme-specific adapters are then ligated to the fragments, followed by reduction of the genomic complexity through PCR using primers with selective nucleotides. Subsequently, the DNA fragments are molecularly cloned, the inserts amplified by PCR and the amplified fragments (amplicons) arrayed onto DNA chips. Thus, a diversity panel consists of a large subset of DNA fragments derived from total DNA of a group of genotypes, immobilized on a solid substrate. Second, organisms or groups of organisms belonging to the gene pool from which the diversity panel was derived can be genetically fingerprinted by testing for hybridization to the arrayed fragments. Two approaches are being used; in the first approach, two genomic samples are converted to so-called "representations" by isolation of DNA, followed by restriction of the DNA and reduction of the genomic complexity using the same procedures as in the development of the diversity panels. The amplicons of each representation are fluorescently labelled with a red or green dye, after which the representations are mixed and hybridized to the diversity panel. For each element of the array, the red/green signal ratio is determined. A significant deviation from a ratio of 1 indicates a difference in the presence of the fragment between the samples, and hence a genetic polymorphism. In the second approach, a single DNA sample is converted to a representation following the same procedures as in the first approach and labelled with green fluorescent dye. In addition,

fragments of the cloning vector that are common to all elements of the array are labelled with red fluorescent dye, after which the green and red fragments are hybridized to the diversity panel. Signal intensity ratios are then determined at each element of the array for each of the genotypes used to construct the diversity panel. Comparison of signal ratios at array elements between genotypes allows the identification of variation in the presence or absence of DNA fragments, and hence the identification of genetic polymorphisms. DArT does not require any sequence data and is considered an economical, high-throughput, robust marker detection technology with a high genome coverage that may have potential relevance to genebanks in germplasm characterization (Jaccoud et al. 2001). The data are similar to AFLPs.

Pyrosequencing™

Pyrosequencing™ refers to sequencing by synthesis, a simple to use technique for accurate analysis of DNA sequences. It is a novel sequencing method of relatively short DNA templates based on real-time (quantitative) pyrophosphate release, as outlined in the following. A DNA fragment consisting of a sequencing primer hybridized to a single stranded DNA template is incubated with the enzymes DNA polymerase, ATP sulfurylase, firefly luciferase and a nucleotide degrading enzyme. The deoxynucleotides dATP, dCTP, dGTP and dTTP are added sequentially in an iterative manner. In case of complementarity to the base of the DNA template, each time the DNA polymerase incorporates a nucleotide to the new DNA strand, pyrophosphate is released in equal molarity to that of the incorporated deoxynucleotide. Pyrophosphate is then used as a substrate for the enzyme ATP sulfurylase converting pyrophosphate into ATP. Subsequently, the concentration of ATP is detected by the enzyme luciferase as a visible and measurable real-time light signal. Between each addition of deoxynucleotides, unincorporated nucleotides and ATP are degraded by the nucleotide-degrading enzyme. Properties of this enzyme include a slower degradation of nucleotides than the nucleotide incorporation by the DNA polymerase and a slower degradation of ATP than ATP synthesis by the sulfurylase. Sequences of up to 20 bases, such as those in which SNPs are found, can be accurately determined by pyrosequencing™. A high throughput of samples can be achieved through a high degree of automation of the methods, e.g. by the use of high-density microtitre plates and micro-injector technology (Ronaghi et al. 1998). Pyrosequencing™can be successfully used for SNP and insertion/deletions detection, genotype identification and characterization, and in general for applications focusing on the variation of short DNA fragments. It can be used for accurate quantification of ratios of variant bases in a DNA sample.

Developments in functional genomics

Allele mining

Identification and access to allelic variation that affects the plant phenotype is of the utmost importance for the utilization of genetic resources, such as in plant variety development. Considering the huge numbers of accessions that are held collectively by genebanks, genetic resources collections are deemed to harbour a wealth of undisclosed allelic variants. The challenge is how to unlock this variation. Allele mining is a research field aimed at identifying allelic variation of relevant traits within genetic resources collections. For identified genes of known function and basic DNA sequence, genetic resources collections may be screened for allelic variation by e.g. the "tiling strategy" using DNA chip technology (e.g. Lemieux et al. 1998). In that approach the basic DNA sequence of a gene is spotted on a chip in the form of large series of sequence-overlapping probes consisting of 15–20 bases. Each base position in a fluorescently labelled sample is then interrogated for the presence of point mutations by monitoring hybridization signals with the spotted probes. Because the sequence of samples is determined in comparison with the primary composition of a gene, this method is also known as "re-sequencing". With this method new point mutations, in relatively large DNA fragments, can be detected. Once allelic variants of interest have been identified, the approach can be optimized by focusing on target sets of polymorphisms, for example by using SNP detection methods (see "Future challenges"). As an example, the tiling strategy has been used by the International Rice Research Institute (IRRI) to identify favourable alleles related to tolerance to biotic and abiotic stress factors in rice.

Association genetics

In allele mining studies as described in the previous section, allelic variation is analyzed for identified genes whose function and basic DNA sequence are known and whose map position in the genome will generally have been determined. On the contrary, in association genetic studies no prior information about the genes of interest is available, but associations between genetic markers and the considered traits are simply derived from observational research (Figure 10). Association genetics focuses on the identification of correlations between phenotypic traits and genetic markers with the aim to identify and locate the underlying genes in the genome (association mapping). Association genetics originated in human genetic studies focusing on the application of significant associations between marker data and diseases for mapping and diagnostic purposes (e.g. Peroutka 1997). Recently, association genetics has

gained increasing interest from plant geneticists (e.g. Buckler and Thornsberry 2002; Rafalski 2002). The rationale behind association genetics is that, in general, alleles at different loci are expected to be randomly associated into genotypes, or in other words, to occur in linkage equilibrium. The more adjacent two loci are, the lower the probability of chromosomal recombination occurring between the loci during meiosis. However, given sufficient numbers of regeneration cycles, once established linkage disequilibria will eventually be disrupted by recombination even in the case of tightly linked loci. Linkage disequilibria are therefore expected to be readily lost from populations, the rate of decay being determined by the recombination fraction and the number of generations elapsed. However, selection tends to accumulate favourably interacting alleles and hence opposes the decay of linkage disequilibria. Observed linkage disequilibria may therefore point towards adaptive significance and possible linkage of the underlying alleles. The interpretation of observed associations is by no means straightforward. For example, population sub-structuring and founder events may lead to irrelevant associations between loci, even if the markers and genes of interest are unlinked. Linkage disequilibria are also affected by the mating system and population history, and hence may vary between different kinds of populations (e.g. selfing vs. outcrossing populations or wild vs. cultivated material). In addition, in the case of complex traits, the presence or absence of a genetic marker may not be indicative for the expression of the phenotype. Theory and analysis regarding genetic linkage and association genetics are still under development (e.g. Ewens and Spielman 2001).

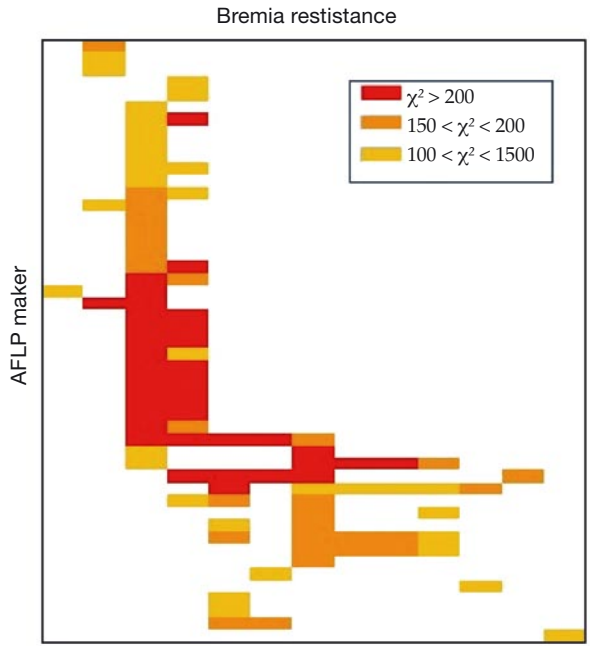


Figure 10. Significant associations between AFLP markers and resistance to different pathotypes of downy mildew in lettuce. Results are based on data from the EU biotech demonstration project "molecular markers for genebanks", directed to the characterization of an entire lettuce collection of about 2300 accessions with molecular markers. Data analyses included the relationship with existing evaluation data for about 1500 accessions of cultivated lettuce. Associations between markers and phenotypic traits were tested for significance by Chi-square values, different colours representing different levels of significance. Note that some markers are associated with resistance to different pathotypes and that some pathotypes are associated with multiple markers (Theo van Hintum, unpublished data).

Comparative genomics

Comparative genomics focuses on the integration of genome information derived from different species with the aim to obtain more insight in the genetic organization of traits through the identification of conserved mechanisms (e.g. Laurie and Devos 2002). This research field has emerged from previous findings of comparative mapping, by which conservation of tracts of collinear markers have been investigated not only among members of the same botanical family, but also among different species, and often led to better understanding of genome evolution. Comparative mapping has shown that a large proportion of the markers and genes are indeed located at comparable positions in the genome (synteny). Together with the availability of an increasing number of genome sequences, including those of known genes, the conservation of gene sequences and their functions among species have also been investigated and are used to further develop the knowledge obtained from previous genetic linkage maps for different species. Conserved Orthologous Set (COS) markers are conserved markers that may serve as anchors for map development (Fulton et al. 2002). Degenerate Oligonucleotide Primed-PCR (DOP-PCR) uses partially degenerated primers for polymorphism detection (Telenius et al. 1992). Because no prior sequence information is required, DOP-PCR is considered useful when crops are involved for which no, or limited sequence information is available. Comparison of the genetic maps of different species will reveal information on chromosome evolution and the common genetic control of traits in different organisms. Comparative maps are being developed for various important crops and are expected to facilitate the understanding of the genetic organization of traits in less studied organisms. This may also reveal novel alleles for relevant genes that subsequently may be exploited in crop improvement.

Future challenges

As outlined in the previous section, many interesting developments in the field of molecular biology and functional genomics are currently ongoing that are relevant to the genebank and user community. However, these developments are accompanied with new and exciting challenges, demanding involvement of expertise from other disciplines, including statistics, bioinformatics and economics.

Advances in biotechnology have resulted in a large variety of molecular marker systems and enhanced opportunities for automation of the majority of the techniques. Therefore, throughput of samples is expected to increase substantially in the near future, resulting in a wealth of information. This is expected to increase even further when the costs to apply the techniques decrease in the future. These developments are important to genebanks, considering the vast amounts of germplasm they maintain. Projects are already underway in which entire collections are screened with molecular markers. In the EU-funded project, "Molecular markers for Genebanks" (http://www.cgn.wur.nl/pgr/research/lettuce/) the lettuce collection of the Centre for Genetic Resources, The Netherlands was characterized with 3 AFLP primer combinations and 10 SSRs. Altogether a total of more than 8 000 DNA samples from about 2300 accessions were analyzed, resulting in nearly 2 million data points. In the EU-funded project, TEGERM (http://www.biocenter.helsinki.fi/bi/tegerm/) high throughput techniques for retrotransposon-based markers (see "Current developments") are used to screen the genetic diversity of several complete germplasm collections of pea, barley, tomato and pepper. The exploding amount of information will inevitably pose serious problems on data access, data analysis and presentation of results. Adequate information systems are currently lacking that can store the data efficiently, that can analyze the data properly in coherence with passport and evaluation data, and that can visualize the results of analyses in a meaningful way. This will frustrate or at least slow down any further progress towards the better conservation and exploitation of genebank collections. During the last years several bioinformatics projects addressing these issues were initiated, such as the EU-funded project "GENE-MINE" (http://www.gene-mine.org/) and the BBSRCfunded project "GERMINATE" (http://bioinf.scri.sari.ac.uk/germinate/).

Economics is getting more attention in the genebank community, assessing the costs of maintaining and characterizing many

accessions against their benefits. Insight in the costs and benefits of different genebank operations will allow genebanks to evaluate the efficiency of their strategies and will aid in the choices to be made (Engels and Visser 2003). For example, in the EU-funded project "ICONFORS" (http://www.igergru.bbsrc.ac.uk/iconfors/index.htm) genetic and economic knowledge is acquired that is essential to improve seed multiplication methodologies for genebanks maintaining *ex situ* seed collections of perennial European forage species. In an extensive experimental set-up, the effect is investigated of different multiplication strategies and sites within Europe on the genetic integrity of accessions of forage species. At the same time, detailed data are collected about all the costs involved in the regeneration procedures. Effects of the different strategies and regeneration sites on the genetic integrity will be evaluated against the necessary costs in order to optimize regeneration protocols.

Molecular markers are now common tools that are applied in many aspects of PGR management. However, like for all genebank operations, the use of molecular markers requires resources of time and funds that should not be taken for granted. For example, molecular markers may be used to rationalize collections for economic limitations. The success of such an approach will very much depend on the expected benefits relative to the investments necessary to collect the data. A case study is provided by McGregor et al. (2002) and van Treuren et al. (2004), using AFLPs that revealed 5% redundancy in a wild potato collection (314 accessions) of the Centre for Genetic Resources, The Netherlands. It was estimated that the cost of the AFLP data was 2.5 times as high as the savings in maintaining the collections for one regeneration cycle; thus, on the long term return of investments can only be expected only after three cycles of regeneration. This may seem not worth the effort. However, the study also revealed other relevant data that are difficult to express in economic terms. For example, the AFLP data identified several taxonomic misclassifications and incorrect origin data that were used to improve the documentation of this collection. From investigation of the AFLP data in relation to geographic data, general guidelines were derived for sampling strategies to lower the probability of sampling similar material in future collection missions. More detailed data on intra-accession variation appeared useful to optimize regeneration protocols. Therefore, the spin-off of molecular studies can be considerable, but these are difficult to value without proper economic theory. Theory about genebank economy is developing, but still in its infancy (Swanson 1996; Pardey et al. 2001; Sackville Hamilton et al. 2002; Engels and Visser 2003).

The huge numbers of accessions that are stored at present in genebanks contain a wealth of genetic diversity that is largely unexplored. Identifying this variation within genebank collections is a major challenge to make the genetic diversity for specific characters available for crop improvement. Allele mining techniques, association genetics and comparative genomics are promising developments for genebanks in order to achieve this goal. Close cooperation between the genebank community, molecular biologists, bio-informaticians and plant breeders is thereby needed, such as in the CGIAR Challenge Program, "Unlocking genetic resources in crops for the resource-poor" (http://www.cgiar.org/research/res_cppilot.html). As more sequence data become available for particular germplasm curators may store sequences as an additional service to compliment traditional services. This would require additional staff with bioinformatics expertise. . The supply of sequence and other molecular marker data might promote the use of germplasm by a wider clientele including plant molecular geneticists and biologists, and as such, contribute to the sustainable conservation of genetic resources. Linked with the interest of providing genomic data are also the relevance of good phenotyping. Germplasm users (breeders, plant physiologists, etc.) may be the best-positioned scientists to conduct appropriate phenotyping of germplasm. The advances in genomic research heavily depend on the availability of reliable phenotype data, so increased awareness of this important task should accompany any involvement of genebank staff in the use of modern technologies.

An important development in agriculture is the increasing interest in ecological farming. To meet the demands from this sector, genebanks may play a role in providing the genetic resources of interest, such as material with a genetically broader base or material less adapted to cultivated conditions. A clear and additional challenge for genebanks in this era of molecular advances is the consideration of the development of pre-breeding as part of their routine services for customers. The fact that molecular markers aid the identification of useful traits in germplasm, and their use in increasing efficiency in introgression in modern cultivars could stimulate more pre-breeding work to increase utilization of genebank collections.

Concluding remarks

A wide variety of new molecular marker technologies are available to assess genetic variation, and many of them are increasingly being applied to complement traditional approaches in germplasm and genebank management. Each technology has its strengths and weaknesses that need to be carefully considered in the light of the intended application. In diversity studies, microsatellites and AFLPs are often the preferred marker technologies. Many marker applications in taxonomy, genebank management and crop breeding rely on the assumption that the diversity assessed is to a large extent predictive for variation in qualitative and quantitative characters. However, this assumption should not be taken for granted. In diversity studies anonymous markers may be located in non-coding genomic regions. There are no few studies testing the assumed link between diversity and taxonomy prediction, and marker data analyses should be used with this caution in mind.

The wealth of sequencing data that are increasingly becoming available for more and more crops via whole-genome sequencing projects, and access to EST-databases, enables the development of markers targeting coding regions of the genome or even specific genes. Technological developments continue to increase sample throughput, which will facilitate large-scale genotyping of genetic resources. Allele mining, association genetics and comparative genomics are promising new approaches to obtain insight in the organization and variation of genes that affect relevant phenotypic traits. Exploiting these developments by combining expertise from several disciplines, including molecular genetics, statistics and bioinformatics is one of the main challenges facing genebanks. Streamlining collaborations among genebanks will greatly aid the exploitation of these techniques efficiently. Ultimately, the biggest challenge to the genebank community is to share materials and technologies to collectively optimize collection, characterization and use, to unlock the useful variation in the world's genebank collections.

Acknowledgments

Special thanks go to Jan Engels of the International Plant Genetic Resources Institute (IPGRI) for his role in encouraging the preparation of this publication and for numerous hours of revision to make a much better final version. We thank reviewers Glenn Brian and Richard Whitkus for review of an earlier version of this manuscript; Sarah Stephenson for editorial assistance; Gerard van der Linden, Applied Biosystems, and Asaph Aharoni and Theo van Hintum for permission to reproduce Figures 7, 8, 9, and 10, respectively. Thanks also to Toby Hodgkin, Ramanatha Rao, Luigi Guarino, Ehsan Dulloo and Prem Mathur for advice in the development of the outline; and IPGRI for financial support to publish this technical bulletin.

Names are necessary to report data. However, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable.

References

Abo-elwafa A, Murai K and Shimada T. 1995. Intra- and inter-specific variations in *Lens* revealed by RAPD markers. Theoretical and Applied Genetics 90:335–340.

Ahnert D, Lee M, Austin DF, Livini C, Woodman WL, Openshaw SJ, Smith JSC, Porter K and Dalton G. 1996. Genetic diversity among elite sorghum inbred lines assessed with DNA markers and pedigree information. Crop Science 36:1385–1392.

Akopyanz N, Bukanov N, Westblom TU and Berg DE. 1992. PCR-based RFLP analysis of DNA sequence diversity in the gastric pathogen *Helicobacter pylori*. Nucleic Acids Research 20:6221–6225.

Aldrich PR, Doebley J, Schertz KF and Stec A. 1992. Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:451–460.

Allard RW. 1970. Population structure and sampling methods. In Genetic resources in plants: their exploration and conservation (OHFrankel and E Bennett, eds.). F.A. Davis Company, Philadelphia, USA. pp. 97–107

Alonso-Blanco C, Peeters AJ, Koornneef M, Lister C, Dean C, van den Bosch N, Pot J and Kuiper MT. 1998. Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population. Plant Journal 14:259–271.

Anthony F, Combes MC, Astorga C, Bertrand B, Graziosi G and Lashermes P. 2002. The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. Theoretical and Applied Genetics 104:894–900.

Arias DM and Rieseberg LH. 1994. Gene flow between cultivated and wild sunflowers. Theoretical and Applied Genetics 89:655–660.

Arnheim N. 1983. Concerted evolution of multigene families. In Evolution of genes and proteins (M Nei and RK Koehn, eds.). Sinauer, Sunderland, Massachusetts, USA. pp.38–61

Avise JC. 2004. Molecular Markers, Natural History and Evolution, (2nd ed.). Sinauer Associates, Sunderland, Massachusetts, USA.

Backes G, Hatz B, Jahoor A, and Fischbeck G. 2003. RFLP diversity within and between major groups of barley in Europe. Plant Breeding 122:291–299.

Bailey CD, Carr TG, Harris SA and Hughes CE. 2003. Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Molecular Phylogenetics and Evolution 29:435–455.

Barrett SCH and Kohn JR. 1991. Genetic and evolutionary consequences of small population size in plants: implications for conservation. In Genetics and Conservation of Rare Plants (D.A. Falk and K.E. Holsinger, eds.). Oxford University Press, Oxford, UK. pp. 3–30.

Bartsch D and Ellstrand NC. 1999. Genetic evidence for the origin of California beets (genus *Beta*). Theoretical and Applied Genetics 99:1120–1130.

Baum DA and Donoghue MJ. 1995. Choosing among alternative "phylogenetic" species concepts. Systematic Botany 20:560–573.

Baurens F, Noyer JL, Lanaud C and Lagoda PJL. 1997. Assessment of a species-specific element (Brep 1) in banana. Theoretical and Applied Genetics 95:922–931.

Becker HC, Engqvist GM and Karlsson B. 1995. Comparison of rapeseed cultivars and resynthesized lines based on allozyme and RFLP markers. Theoretical and Applied Genetics 91:62–67.

Beckmann JS and Soller M. 1986. Restriction fragment length polymorphisms in plant genetic improvement. Plant Molecular Cell Biology 3:197–250.

Bernacchi D, Beck-Bunn T, Emmatty D, Eshed Y, Inai S, Lopez J, Petiard V, Sayama H, Uhlig J, Zamir D and Tanksley S. 1998. Advanced backcross QTL analysis of tomato. II. Evaluation of near-isogenic lines carrying single-donor introgressions for desirable wild QTL-alleles derived from *Lycopersicon hirsutum* and *L. pimpinellifolium*. Theoretical and Applied Genetics 97:170–180.

Berry A and Kreitman M. 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila meloganaster* on the east coast of North America. Genetics 134:869–893.

Bonierbale M, Beebe S, Tohme J and Jones P. 1995. Molecular genetic techniques in relation to sampling strategies and the development of core collections. In Report of the IPGRI workshop, 9–11 October 1995 (WG Ayad, T Hodgkin, A Jaradat and A Rao, eds.). International Plant Genetic Resources Institute, Rome, Italy. pp. 98–102.

Börner A, Chebotar S and Korzun V. 2000. Molecular characterization of the genetic integrity of wheat (*Triticum aestivum* L.) germplasm after long-term maintenance. Theoretical and Applied Genetics 100:494–497.

Botstein D, White RL, Skolnick M and Davis RW. 1980. Construction of a genetic map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32:314–331.

Bredemeijer GMM, Cooke RJ, Ganal MW, Peeters R, Isaac P, Noordijk Y, Rendell S, Jackson J, Röder MS, Wendehake K, Dijcks M, Amelaine M, Wickaert V, Bertrand L and Vosman B. 2002. Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theoretical and Applied Genetics 105:1019–1026.

Bretting PK and Duvick DN. 1997. Dynamic conservation of plant genetic resources. Advances in Agronomy 61:2–51.

Brickell CD, Baum BR, Hetterscheid WLA, Leslie AC, McNeill J, Trehane P, Vrugtman F and Wiersma J. 2004. International code of nomenclature for cultivated plants. Regnum Vegetabile 144:1–123.

Brown ADH. 1989a. The case for core collections. In The use of plant genetic resources (ADH Brown, OH Frankel, DR Marshall and JT Williams, eds.). Cambridge University Press, Cambridge, USA. pp. 136–156.

Brown AHD. 1989b. Core collections: a practical approach to genetic resources management. Genome 31:818–824.

Brown ADH, Brubaker CL and Grace JP. 1997. Regeneration of germplasm samples: wild versus cultivated plant species. Crop Science 37:7–13.

Brown JH and Lomolino MV. 1998. Biogeography (2nd ed.). Sinauer Associates, Inc. Sunderland, Massachusetts, USA.

Brown AHD and Marshall DR. 1995. A basic sampling strategy: theory and practice. In Collecting plant genetic diversity, technical guidelines (L Guarino, V Ramanatha Rao and R Reid, eds.). CAB International, Wallingford, UK. pp. 75–91.

Brown AHD and Munday J. 1982. Population-genetic structure and optimal sampling of land races of barley from Iran. Genetica 58:85–96.

Brubaker CL and Wendel JF. 1994. Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). American Journal of Botany 81:1309–1326.

Bruford MW and Wayne RK. 1993. Microsatellites and their application to population genetic studies. Current Biology 3:939–943.

Brugmans B, Fernandez del Carmen A, Bachem CWB, van Os H, van Eck HJ and Visser RGF. 2002. A novel method for the construction of genome wide transcriptome maps. Plant Journal 31:211–222.

Buckler ES and Thornsberry JM. 2002. Plant molecular diversity and applications to genomics. Current Opinions in Plant Biolology 5:107–111.

Burma BH. 1954. Reality, existence, and classification: a discussion of the species problem. Madroño 12:193–209.

Caetano-Anolles G. 1996. Fingerprinting nucleic acids with arbitrary oligonucleotide primers. Agro Food Industry Hi Tech 7:26–31.

Caetano-Anolles G, Bassam BJ and Gresshoff PM. 1991. DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. Biotechnology 9:553–557.

Caetano-Anolles G, Bassam BJ and Gresshoff PM. 1992. DNA fingerprinting: MAAPing out a RAPD redefinition? Biotechnology 10:937.

Cao WG, Hucl P, Scoles G and Chibbar RN. 1998. Genetic diversity within spelta and macha wheats based on RAPD analysis. Euphytica 104:181–189.

Castillo R and Spooner DM. 1997. Phylogenetic relationships of wild potatoes, *Solanum* series *Conicibaccata* (sect. Petota). Systematic Botany 22:45–83.

Cervera MT, Cabezas JA, Sancha JC, Martinez de Toda F and Martinez-Zapater JM. 1998. Application of AFLPs to the characterization of grapevine *Vitis vinifera* L. genetic resources. A case study with accessions from Rioja (Spain). Theoretical and Applied Genetics 97:51–59.

Chan KF and Sun M. 1997. Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. Theoretical and Applied Genetics 95:865–873.

Charters YM, Robertson A, Wilkinson MJ and Ramsay G. 1996. PCR analysis of oilseed rape cultivars (*Brassica napus* L. ssp. *oleifera*) using 5′-anchored simple sequence repeat (SSR) primers. Theoretical and Applied Genetics 92:442–447.

Chavarriaga-Aguirre P, Maya MM, Tohme J, Duque MC, Iglesias C, Bonierbale MW, Kresovich S and Kochert G. 1999. Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. Molecular Breeding 5:263–273.

Chen XM, Line RF and Leung H. 1998. Genome scanning for resistance-gene analogs in rice, barley, and wheat by high-resolution electrophoresis. Theoretical and Applied Genetics 97:345–355.

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW and Oefner PJ. 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nature Genetics 23:203–207.

Clausen AM and Spooner DM. 1998. Molecular support for the hybrid origin of the wild potato species *Solanum* × *rechei*. Crop Science 38:858–865.

Clegg MT 1993a. Chloroplast gene sequences and the study of plant evolution. Proceedings of the National Academy Science USA 90:363–367.

Clegg MT 1993b. Molecular evaluation of plant genetic resources. In Gene conservation and exploitation: Proceedings of the 20th Stadler genetics symposium held at the University of Missouri, Colombia, Missouri, USA. pp. 67–86.

Comes HP and Abbott RJ. 1998. The relative importance of historical events and gene flow on the population structure of a Mediterranean ragwort, *Senecio gallicus* (Asteraceae). Evolution 52:355–367.

Conner AJ and Dale PJ. 1996. Reconsideration of pollen dispersal data from field trials of transgenic potatoes. Theoretical and Applied Genetics 92:505–508.

Corriveau JL and Coleman AW. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. American Journal of Botany 75:1443–1458.

Cracraft J. 1989. Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In Speciation and its consequences (D Otte and JA Endler, eds.). Sinauer Associates, Inc., Sunderland, Massachusetts, USA. pp. 28–59.

Crawford DJ. 1990. Plant Molecular Systematics: Macromolecular Approaches. John Wiley and Sons, New York, USA.

Crawley MJ, Brown SL, Hails RS, Kohn DD and Rees M. 2001. Transgenic crops in natural habitats. Nature 409:682–683.

Cronn RC, Brothers M, Klier K, Bretting PK and Wendel JF. 1997. Allozyme variation in domesticated annual sunflower and its wild relatives. Theoretical and Applied Genetics 95:532–545.

Cronquist A. 1978. Once again, what is a species? In Biosystematics in agriculture (JA Romberger, ed.). Allenheld, Osman and Company, Montclair, New Jersey, USA. pp. 3–20.

Crossa J and Vencovsky R. 1994. Variance effective population size for two-stage sampling of monoecious species. Crop Science 37:14–26.

Crouch HK, Crouch JH, Jarret RL, Cregan PB and Ortiz R. 1998. Segregation at microsatellite loci in haploid and diploid gametes of *Musa*. Crop Science 38:211–217.

Dale PJ. 1992. Spread of engineered genes to wild relatives. Plant Physiology 100:13–15.

Daly DC, Cameron KM and Stevenson DM. 2001. Plant systematics in the age of genomics. Plant Physiology 127:1328–1333.

Dean RE, Dahlberg JA, Hopkins MS, Mitchell SE and Kresovich S. 1999. Genetic redundancy and diversity among 'orange' accessions in the US National Sorghum Collection as assessed with simple sequence repeat (SSR) markers. Crop Science 39:1215–1221.

Decker DS. 1988. Origin(s), evolution, and systematics of *Curcurbita pepo* (Cucurbitaceae). Economic Botany 42:4–15.

Degani C, Rowland LJ, Levi A, Hortynski JA and Galletta GJ. 1998. DNA fingerprinting of strawberry (*Fragaria* × *ananassa*) cultivars using randomly amplified polymorphic DNA (RAPD) markers. Euphytica 102:247–253.

Del Rio AH and Bamberg JB. 2002. Lack of association between genetic and geographical origin characteristics for the wild potato *Solanum sucrense*. American Journal of Potato Research 79:335–338.

Del Rio AH, Bamberg JB and Huamán Z. 1997a. Assessing changes in the genetic diversity of potato gene banks. 1. Effects of seed increase. Theoretical and Applied Genetics 95:191–198.

Del Rio AH, Bamberg JB, Huamán Z, Salas A and Vega SE. 1997b. Assessing changes in the genetic diversity of potato gene banks. 2. *in situ* vs. *ex situ*. Theoretical and Applied Genetics 95:199–204.

Desplanque B, Boudry P, Broomberg K, Saumitou-Laprade P, Cuguen J and van Dijk H. 1999. Genetic diversity and gene flow between wild, cultivated and weedy forms of *Beta vulgaris* L. (Chenopodiaceae), assessed by RFLP and microsatellite markers. Theoretical and Applied Genetics 98:1194–1201.

Desplanque B, Hautekèete N and van Dijk H. 2002. Transgenic weed beets: possible, probable, avoidable? Journal of Applied Ecology 39:561–571.

DeVicente MC and Tanksley SD. 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134:585–596.

Dice LR. 1945. Measures of the amount of ecologic association between species. Ecology 26:279–302.

Diers BW, Osborn TC and McVetty PBE. 1996. Relationship between heterosis and genetic distance based on Restriction Fragment Length Polymorphism markers in oilseed rape (*Brassica napus* L.). Crop Science 36:79–83.

Dillon SL, Lawrence PK and Henry RJ. 2001. The use of ribosomal ITS to determine phylogenetic relationships within *Sorghum*. Plant Systematics and Evolution 230:97–110.

Dobzhansky T. 1950. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseodobscura*. Genetics 35:288–302.

Doebley JF. 1989. Molecular evidence for a missing wild relative of maize and the introgression of its chloroplast genome into *Zea perennis*. Evolution 43:1555–1559.

Doebley JF. 1992. Molecular systematics and crop evolution. In Molecular Systematics of Plants (PS Soltis, DE Soltis and JJ Doyle, eds.). Chapman and Hall, New York, USA. pp. 202–222.

Doebley JF, Goodman MM and Stuber CW. 1984. Isoenzymatic variation in *Zea* (Gramineae). Systematic Botany 9:203-218.

Doldi ML, Vollmann J and Lelley T. 1997. Genetic diversity in soybean as determined by RAPD and microsatellite analysis. Plant Breeding 116:331–335.

Dos Santos JB, Nienhuis J, Skroch P, Tivang J and Slocum MK. 1994. Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. Theoretical and Applied Genetics 87:909–915.

Doyden JT and Slobobchikoff CN. 1974. An operational approach to species classification. Systematic Zoology 23:239–247.

Durham RE and Korban SS. 1994. Evidence of gene introgression in apple using RAPD markers. Euphytica 79:109–114.

Dubreuil P, Dufour P, Krejci E, Causse M, De Vienne D, Gallais A and Charcosset A. 1996. Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups. Crop Science 36:790–799.

Eernisse DJ and Kluge AG. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10:1170–1195.

Ehrlich PR and Raven PH. 1969. Differentiation of populations. Science 165:1228–1231.

Ellstrand NC. 2001. When transgenes wander, should we worry? Plant Physiology 125:1543–1545.

Ellstrand NC and Elam DR. 1993. Population genetic consequences of small population size: implications for plant conservation. Annual Review of Ecology and Systematics 24:217–242.

Ellstrand NC, Prentice HC and Hancock JF. 1999. Gene flow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics 30:539–563.

Engels JMM and Visser L. 2003. A guide to effective management of germplasm collections. IPGRI handbook for Genebanks No. 6. International Plant Genetic Resources Institute, Rome, Italy.

Ereshefsky M. 2001. The poverty of the Linnaean hierarchy. Cambridge University Press, Cambridge, UK.

Erskine W and Muehlbauer FJ. 1991. Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm. Theoretical and Applied Genetics 83:119–125.

Eshed Y, Gera G and Zamir D. 1996. A genome-wide search for wild-species alleles that increase horticultural yield for processing tomatoes. Theoretical and Applied Genetics 93:877–886.

Eshed Y and Zamir D. 1995. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147–1162.

Espejo-Ibañez MC, Sanchez MP, Sanchez MD and Yelamo MD. 1994. Isoenzymatic variability in seeds of some Spanish common beans (*Phaseolus vulgaris* L. Leguminosae): relation to their domestication centers. Biochemical Systematics and Ecology 22:827–833.

Eujayl I, Sorrells ME, Baum M, Wolters P and Powell W. 2002. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theoretical and Applied Genetics 104:399–407.

Ewens WJ and Spielman RS. 2001. Overview: locating genes by linkage and association. Theoretical Population Biology 60:135–139.

Excoffier L, Smouse PE and Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Fahima T, Roeder MS, Grama A and Nevo E. 1998. Microsatellite DNA polymorphism divergence in *Triticum dicoccoides* accessions highly resistant to yellow rust. Theoretical and Applied Genetics 96:187–195.

Falconer DS. 1981. Introduction to Quantitative Genetics (2nd ed.). Longman, London, UK.

Fang DQ, Roose ML, Krueger RR and Federici CT. 1997. Fingerprinting trifoliate orange germplasm accessions with isozymes, RFLPs, and inter-simple sequence repeat markers. Theoretical and Applied Genetics 95:211–219.

FAO. 1996. The State of the World's Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization, Rome, Italy.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27:401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Flavell AJ, Knox MR, Pearce SR and Ellis THN. 1998. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant Journal 16:643–650.

Frankel OH. 1984. Genetic perspectives on germplasm conservation. In Genetic Manipulation: Impact on Man and Society (W Arber, K Illmensee, WJ Peacock and P Starlinger, eds.). Cambridge University Press, Cambridge, UK. pp. 161–170.

Frankel OH and Brown AHH. 1984. Current plant genetic resources-a critical appraisal. In Genetics: new frontiers. Vol. 4. Oxford and IBH Publishing Co., New Delhi, India. pp. 1–11.

Frary A, Nesbitt TC, Frary A, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB and Tanksley SD. 2000. Fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88.

Fregene M, Bernal A, Duque M, Dixon A and Tohme J. 2000. AFLP analysis of African cassava (*Manihot esculenta* Crantz) germplasm resistant to the cassava mosaic disease (CMD). Theoretical and Applied Genetics 100:678–685.

Freville H, Justy F and Olivieri I. 2001. Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). Molecular Ecology 10:879–889.

Fry WE and Goodwin SB. 1997. Re-emergence of potato and tomato late blight in the United States. Plant Disease 81:1349–1357.

Fulton TM, van der Hoeven R, Eannetta NT and Tanksley SD. 2002. Identification, analysis and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467.

Garvin DF and Weeden NF. 1994. Isozyme evidence supporting a single geographic origin for domesticated tepary bean. Crop Science 34:1390–1395.

Geffroy V, Sevignac M, de Oliveira JCF, Fouilloux G, Skroch P, Thoquet P, Gepts P, Langinand T and Dron M. 2000. Inheritance of partial resistance against *Colletotrichum lindemuthianum* in *Phaseolus vulgaris* and co-localization of quantitative trait loci with genes involved in specific resistance. Molecular Plant-Microbe Interactions 13:287–296.

Ghislain M, Spooner DM, Rodríguez F, Villamon F, Núñez C, Vásquez C and Bonierbale M. 2004. Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. Theoretical and Applied Genetics 108:881–890.

Ghislain M, Zhang D, Fajardo D, Huamán Z and Hijmans RJ. 1999. Marker-assisted sampling of the cultivated Andean potato *Solanum phureja* collection using RAPD markers. Genetic Resources and Crop Evolution 46:547–555.

Giannattasio RB and Spooner DM. 1994. A reexamination of species boundaries and hypotheses of hybridization concerning *Solanum megistacrolobum* and *S. toralapanum* (*Solanum* sect. *Petota*, series *Megistacroloba*): molecular data. Systematic Botany 19:106–115.

Gibson G. 2002. Microarrays in ecology and evolution: a preview. Molecular Ecology 11:17–24.

Godwin ID, Aitken EAB and Smith LW. 1997. Application of inter simple sequence repeat (ISSR) markers to plant genetics. Electrophoresis 18:1524–1528.

Golembiewski RC, Danneberger TK and Sweeney PM. 1997. Potential of RAPD markers for use in the identification of creeping bentgrass cultivars. Crop Science 37:212–214.

Graner A, Ludwig WF and Melchinger AE. 1994. Relationships among European barley germplasm: II. Comparison of RFLP and pedigree data. Crop Science 34:1199–1205.

Graner A, Streng S, Kellermann A, Schiemann A, Bauer E, Waugh R, Pellio B and Ordon F. 1999. Molecular mapping and genetic fine-structure of the rym5 locus encoding resistance to different strains of the Barley Yellow Mosaic Virus Complex. Theoretical and Applied Genetics 98:285–290.

Grenier C, Deu M, Kresovich S, Bramel-Cox PJ and Hamon P. 2000. Assessment of genetic diversity in three subsets constituted from the ICRISAT *Sorghum* collection using random vs. non-random sampling procedures B. Using molecular markers. Theoretical and Applied Genetics 101:197–202.

Greuter W, Mcneill J, Barrie FR, Burdett HM, Demoulin V, Filgueiras TS, Nicolson DH, Silva PC, Skog JE, Trehane P, Turland NJ and Hawksworth DL (eds. and compilers). 2000. International Code of Botanical Nomenclature (St. Louis Code). Regnum Vegetabile 138:1–474.

Guilford P, Prakash S, Zhu JM, Rikkerink E, Gardiner S, Bassett H and Foster R. 1997. Microsatellites in *Malus* × *domestica* (apple): Abundance, polymorphism and cultivar identification. Theoretical and Applied Genetics 94:249–254.

Gupta M, Chyi Y-S, Romero-Severson J and Owen JL. 1994. Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theoretical and Applied Genetics 89:998–1006.

Gupta PK, Balyan HS, Sharma PC and Ramesh B. 1996. Microsatellites in plants: a new class of molecular markers. Current Science 70:45–54.

Hadrys H, Balick M and Schierwater B. 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. Molecular Ecology 1:55–63.

Hall BG. 2001. Phylogenetic trees made easy: a how-to manual for molecular biologists. Sinauer Associates, Inc., Sunderland, Massachusetts, USA.

Hallden C, Nilsson NO, Rading IM and Saell T. 1994. Evaluation of RFLP and RAPD markers in comparison of *Brassica napus* breeding lines. Theoretical and Applied Genetics 88:123–128.

Hamilton MB. 1994. *Ex situ* conservation of wild plant species: time to assess the genetic assumptions and implications of seed banks. Conservation Biology 8:39–49.

Hamon S, Dussert S, Deu M, Hamon P, Seguin M, Glaszmann JC, Grivet L, Chantereau J, Chevallier MH, Flori A, Lashermes P, Legnate H and Noirot M. 1998. Methodologies de gestion et de conservation des ressources genetiques. Genetic Selection and Evolution 30:S237–S258 (suppl.).

Hamon S, Dussert S, Noirot M, Anthony F and Hodgkin T. 1995. Core collections—accomplishments and challenges. Plant Breeding Abstracts 65:1125–1133.

Hamrick JL and Godt MJW. 1997. Allozyme diversity in cultivated crops. Crop Science 37:26–30.

Hancock JF, Grumet R and Hokanson SC. 1996. The opportunity for escape of engineered genes from transgenic crops. HortScience 31:1080–1085.

Hansen LB, Siegismund HR and Jørgensen RB. 2001. Introgression between oilseed rape (*Brassica napus* L.) and its weedy relative *B. rapa* L. in a natural population. Genetic Resources and Crop Evolution 48:621–627.

Harberd DJ. 1975. *Brassica*. In Hybridization and the flora of the British Isles (CA Stace, ed.). Academic Press, London,UK. pp. 137–139.

Harlan JR and de Wet JMJ. 1971. Toward a rational classification of cultivated plants. Taxon 20:509–517.

Hauser MT, Adhami F, Dorner M, Fuchs E and Glossl J. 1998. Generation of co-dominant PCR-based markers by duplex analysis on high resolution gels. Plant Journal 16:117–125.

Hawkes JG. 1990. The Potato: Evolution, Biodiversity, and Genetic Resources. Belhaven Press, Oxford, UK.

Hayashi K. 1992. PCR-SSCP: a method for detection of mutations. Genetic Analysis Techniques and Applications 9:73–79.

Hearne CM, Ghosh S and Todd JA. 1992. Microsatellites for linkage analysis of genetic traits. Trends in Genetics 8:288–294.

Heath DD, Iwama GK and Devlin RH. 1993. PCR primed with VNTR core sequences yields species specific patterns and hypervariable probes. Nucleic Acids Research 21:5782–5785.

Hennig W. 1950. Grundzüge einer Theorie der Phylogenetischen Systematik. Deutscher Zentralverlag, Berlin, Germany.

Hennig W. 1966. Phylogenetic Systematics (3rd ed.) (trans. DD Davis and R Zanderl). University of Illinois Press, Urbana Illinois, USA.

Hijmans RJ, Jacobs M, Bamberg JB and Spooner DM. 2003. Frost tolerance in wild potato species: unraveling the predictivity of taxonomic, geographic and ecological factors. Euphytica 130:47–59.

Hill M, Witsenboer H, Zabeau M, Vos P, Kesseli R and Michelmore R. 1996. PCR-based fingerprinting using AFLPs as a tool for studying genetic relationships in *Lactuca* spp. Theoretical and Applied Genetics 93:1202–1210.

Hillis DM. 1997. Phylogenetic analysis. Current Biology 7:R129–R131.

Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy and investigator bias. Systematic Biology 47:3–8.

Hillis DM, Moritz C and Mable BK. 1996. Molecular Systematics of Plants (2nd ed.). Sinauer Associates, Inc., Sunderland, Massachusetts. USA.

Hodgkin T, Brown AHD, van Hintum ThJL and Morales EAV. 1995. Core Collections of Plant Genetic Resources. John Wiley & Sons, Chichester, UK.

Holland PM, Abramson RD, Watson R and Gelfland DH. 1991. Detection of specific polymerase chain reaction product by utilizing the 5′→3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proceedings of the National Academy of Sciences USA. 88:7276–7280.

Hu J and Vick BA. 2003. Target region amplification polymorphism: a novel marker technique for plant genotyping. Plant Molecular Biology Reporter 21:289–294.

Huamán Z, Ortiz R and Gomez R. 2000. Selecting a *Solanum tuberosum* subsp. *andigena* core collection using morphological, geographical, disease and pest descriptors. American Journal of Potato Research 77:183–190.

Huang HW, Layne DR and Kubisiak TL. 2000. RAPD inheritance and diversity in pawpaw (*Asimina triloba*). Journal of the American Society of Horticultural Science 125:454–459.

Huang HW, Layne DR and Riemenschneider DE. 1998. Genetic diversity and geographic differentiation in pawpaw [*Asimina triloba* (L.) Dunal] populations from nine states as revealed by allozyme analysis. Journal of the American Society of Horticultural Science 123:635–641.

Hudson RR, Bailey K, Skarecky D, Kwaitowski J and Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics 136:1329–1340.

Huff DR. 1997. RAPD characterization of heterogeneous perennial ryegrass cultivars. Crop Science 37:557–564.

Hymowitz T, Singh RJ and Kollipara KP. 1998. The genomes of *Glycine*. Plant Breeding Reviews 16:289–319.

Jaccard P. 1908. Nouvelles recherches sur la distribution florale. Bulletin Société Vaudoise Sciences Naturelles 44:223–270.

Jaccoud D, Peng K, Feinstein D and Kilian A. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research 29:e25.

Jarne P and Lagoda PJL 1996. Microsatellites, from molecules to populations and back. Trends Ecology and Evolution 11:424–429.

Jeffreys AJ, Wilson V and Thein SL. 1985a. Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73.

Jeffreys AJ, Wilson V and Thein SL. 1985b. Individual-specific "fingerprints" of human DNA. Nature 316:76–79.

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, van de Wiel C, Bredemeijer G, Vosman B, Matthes M, Daly A, Brettschneider R, Bettini P, Buiatti M, Maestri E, Malcevschi A, Marmiroli N, Aert R, Volckaert G, Rueda J, Linacero R, Vazquez A and Karp A. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Molecular Breeding 3:381–390.

Jørgensen RB and Andersen B. 1994. Spontaneous hybridization between oilseed rape (*Brassica napus*) and weedy *B. camprestis* (Brassicaceae): a risk of growing genetically modified oilseed rape. American Journal of Botany 81:1620–1626.

Jørgensen RB, Anderson B, Landbo L, and Mikkelsen T. 1996. Spontaneous hybridization between oilseed rape (*Brassica napus*) and weedy relatives. In Proceedings of an International Symposium on *Brassicas*/Ninth Crucifer Genetics Workshop (JS Dias, I Crute, and AA Montiero, eds.). ISHS, Lisbon, Portugal. pp. 193–197.

Judd WS, Campbell CS, Kellogg EA and Stevens PF. 2002. Plant systematics: a phylogenetic approach (2nd ed.). Sinauer Associates, Inc., Sunderland, Massachusetts, USA.

Kalendar R, Grob T, Regina M, Suoniemi A and Schulman A. 1999. IRAP and REMAP: Two new retrotransposon-based DNA fingerprinting techniques. Theoretical and Applied Genetics 98:704–711.

Kardolus JP, van Eck HJ and van den Berg RG. 1998. The potential of AFLPs in biosystematics: a first application in *Solanum* taxonomy (Solanaceae). Plant Systematics and Evolution 210:87–103.

Karp A, Edwards KJ, Bruford M, Funk S, Vosman B, Morgante M, Seberg O, Kremer A, Boursot P, Arctander P, Tautz D and Hewitt GM. 1997a. Molecular technologies for biodiversity evaluation: opportunities and challenges. Nature Biotechology 15:625–628.

Karp A, Kresovich S, Bhat KV, Ayad WG and Hodgkin T. 1997b. Molecular tools in plant genetic resources conservation: a guide to the technologies IPGRI Technical Bulletin No. 2. International Plant Genetic Resources Institute, Rome, Italy.

Karp A, Seberg O and Buiatti M. 1996. Molecular techniques in the assessment of botanical diversity. Annals of Botany 78:143–149.

Kephart SR. 1990. Starch gel electrophoresis of plant isozymes: a comparative analysis of techniques. American Journal of Botany 77:693–712.

Kesseli R, Ochoa O and Michelmore R. 1991. Variation at RFLP loci in *Lactuca* spp. and origin of cultivated lettuce (*L. sativa*). Genome 34:430–436.

Kiang YT, Antonovics J and Wu L. 1979. The extinction of wild rice (*Oryza perennis-formosana*) in Taiwan. Journal of Asian Ecology 1:1–9.

Kiers AM, Mes THM, van der Meijden R and Bachmann K. 2000. A search for diagnostic AFLP markers in *Cichorium* species with emphasis on endive and chicory cultivar groups. Genome 43:470–476.

Khlestkina EK, Huang XQ, Quenum FJ-B, Chebotar S, Röder MS and Börner A. 2004. Genetic diversity in cultivated plants—loss or stability? Theoretical and Applied Genetics 108:1466–1472.

Koester RP, Sisco PH and Stuber CW. 1993. Identification of quantitative trait loci controlling days to flowering and plant height in two near isogenic lines of maize. Crop Science 33:1209–1216.

Kollipara KP, Singh RJ and Hymowitz T. 1997. Phylogenetic and genomic relationships in the genus *Glycine* Willd. based on sequences from the ITS region of nuclear rDNA. Genome 40:57–68.

Konieczny A and Ausubel FM. 1993. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant Journal 4:403–410.

Koopman WJM, Zevenbergen MJ and van den Berg RG. 2001. Species relationships in *Lactuca* s.l. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. American Journal of Botany 88:1881–1887.

Koornneef M and Stam P. 2001. Changing paradigms in plant breeding. Plant Physiolology 125:156–159.

Korzun V, Boerner A, Worland AJ, Law CN and Roeder MS. 1997. Applications of microsatellite markers to distinguish inter-varietal chromosome substitution lines of wheat (*Triticum aestivum* L.). Euphytica 95:149–155.

Kota R, Wolf M, Michalek W and Graner A. 2001. Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (*Hordeum vulgare* L.). Genome 44:523–528.

Kreiger M and Ross KG. 2002. Identification of a major gene regulating complex social behavior. Science 295:328–332.

Lamboy WF and Alpha CG. 1998. Using simple sequence repeats (SSRs) for DNA fingerprinting germplasm accessions of grape (*Vitis* L.) species. Journal of the American Society of Horticultural Science 123:182–188.

Lamboy WF, McFerson JR, Westman AL and Kresovich S. 1994. Application of isozyme data to the management of the United States national *Brassica oleracea* L. genetic resources collection. Genetic Resources and Crop Evolution 41:99–108.

Lamboy WF, Yu J, Forsline PL and Weeden NF. 1996. Partitioning of allozyme diversity in wild populations of *Malus sieversii* L. and implications for germplasm collection. Journal of the American Society of Horticultural Science 121:982–987.

Lanner HC, Bryngelsson T and Gustafsson M. 1997. Relationships of wild *Brassica* species with chromosome number 2n = 18, based on RFLP studies. Genome 40:302–308.

Lanner HC, Gustafsson M, Falt AS and Bryngelsson T. 1996. Diversity in natural populations of wild *Brassica oleracea* as estimated by isozyme and RAPD analysis. Genetic Resources and Crop Evolution. 43:13–23.

Laurie DA and Devos KM. 2002. Trends in comparative genetics and their potential impacts on wheat and barley research. Plant Molecular Biology 48:729–740.

Lee DJ, Berding N, Jackes BR and Bielig LM. 1998. Isozyme markers in *Saccharum* spp. hybrids and *Erianthus arundinaceus* (Retz.) Jeswiet. Australian Journal of Agricultural Research. 49:915–921.

Lee LG, Connell CR and Bloch W. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. Nucleic Acids Research 21:3761–3766.

Lee DJ, Reeves JC and Cooke RJ. 1996. DNA profiling and plant variety registration: 1. The use of random amplified DNA polymorphisms to discriminate between varieties of oilseed rape. Electrophoresis 17:261–265.

Lemieux B, Aharoni A and Schena M. 1998. Overview of DNA chip technology. Molecular Breeding 4:277–289.

Lenné JM and Wood D. 1991. Plant diseases and the use of wild germplasm. Annual Review of Phytopathology 29:35–63.

Levin DA. 2000. The origin, expansion, and demise of plant species. Oxford University Press, New York, New York, USA.

Levin DA, Francisco-Ortega J and Jansen RK. 1996. Hybridization and extinction of rare plant species. Conservation Biology 10:10–16.

Li G and Quiros CF. 2001. Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in *Brassica.* Theoretical and Applied Genetics 103:455–461.

Li Z, Pinson SRM, Stansel JW and Park WD. 1995. Identification of quantitative trait loci (QTLs) for heading date and plant height in cultivated rice (*Oryza sativa* L.). Theoretical and Applied Genetics 91:374–381.

Lin JJ, Kuo J, Jin M, Saunders DA, Beard HS, MacDonald MH, Kenworthy W, Ude GN and Matthews BF. 1996. Identification of molecular markers in soybean comparing RFLP, RAPD and AFLP DNA mapping techniques. Plant Molecular Biology Reporter 14:156–169.

Lincoln R, Boxshall G and Clark P. 1998. A dictionary of ecology, evolution and systematics (2nd ed.). Cambridge University Press, Cambridge, Massachusetts, USA.

Linder CR, Taha I, Seiler GJ, Snow AA and Rieseberg LH. 1998. Long-term introgression of crop genes into wild sunflower populations. Theoretical and Applied Genetics 96:339–347.

Link W, Dixkens C, Singh M, Schwall M and Melchinger AE. 1995. Genetic diversity in European and Mediterranean faba bean germ plasm revealed by RAPD markers. Theoretical and Applied Genetics 90:27–32.

Llewellyn D and Fitt G. 1996. Pollen dispersal from two field trials of transgenic cotton in the Namoi Valley, Australia. Molecular Breeding 2:157–166.

Lopez-Sese AI, Staub J, Katzir N and Gomez-Guillamon ML. 2002. Estimation of between and within accession variation in selected Spanish melon germplasm using RAPD and SSR markers to assess strategies for large collection evaluation. Euphytica 127:41–51.

Lu J, Knox MR, Ambrose MJ, Brown JKM and Ellis THN. 1996. Comparative analysis of genetic diversity in pea assessed by RFLP- and PCR-based methods. Theoretical and Applied Genetics 93:1103–1111.

Maass HI and Klaas M. 1995. Infraspecific differentiation of garlic (*Allium sativum* L.) by isozyme and RAPD markers. Theoretical and Applied Genetics 91:89–97.

Maass BL and Ocampo CH. 1995. Isozyme polymorphism provides fingerprints for germplasm of *Arachis glabrata* Bentham. Genetic Resources and Crop Evolution 42:77–82.

Maddison WP. 1995. Phylogentic histories within and among species. In Experimental and molecular approaches to plant biosystematics (PC Hoch and AG Stephenson, eds.). Missouri Botanical Garden, St. Louis, Missouri, USA. pp. 273–287.

Maddison WP, Donoghue MJ and Maddison DR. 1984. Outgroup analysis and parsimony. Systematic Zoology 33:83–103.

Mallet J. 2001. Species, concepts of. In Encyclopedia of Biodiversity (SA Levin, ed.). Academic Press, San Diego, California, USA. pp. 427–440.

Mallet J. 2004. Poulton, Wallace and Jordan: how discoveries in Papilio butterflies led to a new species concept 100 years ago. Systemetics and Biodiversity 1:441–452.

Mandolino G, De MS, Faeti V, Bagatta M, Carboni A and Ranalli P. 1996. Stability of fingerprints of *Solanum tuberosum* plants derived from conventional tubers and vitrotubers. Plant Breeding 115:439–444.

Manjarrez-Sandoval P, Carter Jr. TE, Webb DM and Burton JW. 1997. Heterosis in soybean and its prediction by genetic similarity measures. Crop Science 23:1443–1452.

Maquet A, Zoro Bi IZ, Delvaux M, Wathelet B and Baudoin JP. 1997. Genetic structure of a Lima bean base collection using allozyme markers. Theoretical and Applied Genetics 95:980–991.

Maquet A, Zoro Bi IZ, Rocha OJ and Baudoin JP. 1996. Case studies on breeding systems and its consequences for germplasm conservation. Genetic Resources and Crop Evolution 43:309–318.

Marita JM, Rodríguez JM and Nienhuis JM. 2000. Development of an algorithm identifying maximally diverse core collections. Genetic Resources and Crop Evolution 47: 515–526.

Martin C, Juliano A, Newbury HJ, Lu BR, Jackson MT and Ford Lloyd BV. 1997. The use of RAPD markers to facilitate the identification of *Oryza* species within a germplasm collection. Genetic Resources and Crop Evolution 44:175–183.

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M and Doebley J. 2002. Microsatellites in *Zea*—variability, patterns of mutations, and use for evolutionary studies. Theoretical and Applied Genetics 104:436–450.

Mau B, Neuton MA and Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo analysis. Biometrics 55:1–12.

Maughan PJ, Saghai Maroof MA and Buss GR. 1995. Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38:715–723.

May B. 1992. Starch gel electrophoresis of allozymes. In Molecular genetic analysis of populations: a practical approach (AR Hoelzel, ed.). Oxford University Press, Oxford, UK. pp. 1–27.

Mayden RL 1997. A hierarchy of species concepts: the denouement of the saga of the species problem. In Species: the units of biodiversity (MF Claridge, HA Dawson and MR Wilson, eds.). Chapman and Hall, New York, New York, USA. pp. 381–424

Mayer MS, Tullu A, Simon CJ, Kumar J, Kaiser WJ, Kraft JM and Muehlbauer FJ. 1997. Development of a DNA marker for Fusarium wilt resistance in chickpea. Crop Science 37:1625–1629.

Mayr E. 1942. Systematics and the origin of species. Columbia University Press, New York, USA.

McGregor CE, van Treuren R, Hoekstra R and van Hintum TJL. 2002. Analysis of the wild potato germplasm of the series *Acaulia* with AFLPs: implications for *ex situ* conservation. Theoretical and Applied Genetics 104:146–156.

Melchinger AE, Graner A, Singh M and Messmer MM. 1994. Relationships among European barley germplasm. I. Genetic diversity among winter and spring cultivars revealed by RFLPs. Crop Science 34:1191–1199.

Menkir A, Goldsbrough P and Ejeta G. 1997. RAPD based assessment of genetic diversity in cultivated races of sorghum. Crop Science 37:564–569.

Michener CD. 1963. Some future developments in taxonomy. Systematic Zoology 12:151–172.

Milbourne D, Meyer R, Bradshaw JE, Baird E, Bonar N, Provan J, Powell W and Waugh R. 1997. Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Molecular Breeding 3:127–136.

Miller DR and Rossman AY. 1995. Systematics, biodiversity, and agriculture. Bioscience 45:680–686.

Miller JT and Spooner DM. 1996. Introgression of *Solanum chacoense* (*Solanum* sect. Petota) upland populations reexamined. Systematic Botany 21:461–475.

Miller JT and Spooner DM. 1999. Collapse of species boundaries in the wild potato *Solanum brevicaule* complex (Solanaceae, *S.* sect. *Petota*): molecular data. Plant Systematics and Evolution 214:103–130.

Miller JC and Tanksley SD. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theoretical and Applied Genetics 80:437–448.

Minghetti PP and Dugaiczyk A. 1993. The emergence of new DNA repeats and the divergence of primates. Proceedings of the National Academy of Sciences USA 90:1872–1876.

Morgante M and Olivieri AM. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant Journal 3:175–182.

Morgante M, Hanafey H and Powell W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genome. Nature Genetics 30:194–200.

Mort ME and Crawford DJ. 2004. The continuing search: low-copy nuclear sequences for lower level plant molecular phylogenetic studies. Taxon 53:257–261.

Mumm RH and Dudley JW. 1994. A classification of 148 U.S. maize inbreds: I. Cluster analysis based on RFLPs. Crop Science 34:842–851.

Murphy JP and Philips TD. 1993. Isozyme variation in cultivated oat and its progenitor species, *Avena sterilis* L. Crop Science 33:1366–1372.

Nagaoka T and Ogihara Y. 1997. Applicability of inter-simple sequence repeat polymorphisms in wheat for use as DNA markers in comparison to RFLP and RAPD markers. Theoretical and Applied Genetics 94:597–602.

National Academy of Sciences. 1989. Field testing genetically modified organisms: framework for decisions. National Academy Press, Washington, D.C., USA.

National Research Council. 1972. Genetic vulnerability of major crops. National Academy of Sciences, Washington, D.C., USA.

National Research Council. 2002. Environmental effects of transgenic plants: the scope and adequacy of regulation. National Academy Press, Washington, D.C., USA.

Neale DB and Williams CG. 1991. Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Canadian Journal of Forest Research 21:545–554.

Nebauer SG, del Castillo-Agudo L and Segura J. 1999. RAPD variation within and among natural populations of outcrossing willow-leaved foxglove (*Digitalis obscura* L.). Theoretical and Applied Genetics 98:985–994.

Negash A, Tsegaye A, van Treuren R and Visser L. 2002. AFLP analysis of enset clonal diversity in south and southwestern Ethiopia for conservation. Crop Science 42:1105–1111.

Nei M and Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences USA 76:5269–5273.

Nevo E, Apelbaum Elkaher I, Garty J and Beiles A. 1997. Natural selection causes microscale allozyme diversity in wild barley and a lichen at 'Evolution Canyon', Mt. Carmel, Israel. Heredity 78:373–382.

Nevo E, Baum B, Beiles A and Johnson DA. 1998. Ecological correlates of RAPD DNA diversity of wild barley, *Hordeum spontaneum*, in the Fertile Crescent. Genetic Resources and Crop Evolution 45:151–159.

Nevo E, Golenberg E, Beiles A, Brown AHD and Zohary D. 1982. Genetic diversity and environmental associations of wild wheat, *Triticum dicoccoides*, in Israel. Theoretical and Applied Genetics 62:241–254.

Oberwalder B, Ruoss B, Schilde Rentschler L, Hemleben V and Ninnemann H. 1997. Asymmetric fusion between wild and cultivated species of potato (*Solanum* spp.)—detection of asymmetric hybrids and genome elimination. Theoretical and Applied Genetics 94:1104–1112.

O'Donoughue LS, Souza E, Tanksley SD and Sorrells ME. 1994. Relationships among North American oat cultivars based on restriction fragment length polymorphisms. Crop Science 34:1251–1258.

Olmstead RG. 1995. Species concepts and plesiomorphic species. Systematic Botany 20:623–630.

Olufowote JO, Xu Y, Chen X, Park WD, Beachell HM, Dilday RH, Goto M and McCouch SR. 1997. Comparative evaluation of within-cultivar variation of rice (*Oryza sativa* L.) using microsatellite and RFLP markers. Genome 40:370–378.

Palombi MA and Damiano C. 2002. Comparison between RAPD and SSR molecular markers in detecting genetic variation in kiwifruit (*Actinidia deliciosa* A. Chev). Plant Cell Reports 20:1061–1066.

Paran I and Michelmore RW. 1993. Development of reliable PCR based markers linked to downy mildew resistance genes in lettuce. Theoretical and Applied Genetics 85:985–993.

Parani M, Singh KN, Rangasamy S and Ramalingam RS. 1997. Identification of *Sesamum alatum* × *Sesamum indicum* hybrid using protein, isozyme and RAPD markers. Indian Journal of Genetics and Plant Breeding 57:381–388.

Pardey PG, Koo B, Wright BD, van Dusen ME, Skovmand B and Taba S. 2001. Costing the conservation of genetic resources: CIMMYT's *ex situ* maize and wheat collection. Crop Science 41:1286–1299.

Parker PG, Snow AA, Schug MD, Booton GC and Fuerst PA. 1998. What molecules can tell us about populations: choosing and using a molecular marker. Ecology 79:361–382.

Parsons BJ, Newbury HJ, Jackson MT and Ford-Lloyd BV. 1997. Contrasting genetic diversity relationships are revealed in rice (*Oryza sativa* L.) using different marker types. Molecular Breeding 3:115–125.

Parzies HK, Spoor W and Ennos RA. 2000. Genetic diversity of barley landrace accessions (*Hordeum vulgare* ssp. *vulgare*) conserved for different lengths of time in *ex situ* gene banks. Heredity 84:476–486.

Patterson TB and Givinish TJ. 2002. Phylogeny, concerted convergence, and phylogenetic niche conservatism in the core Liliales: insights from *rbc*L and *ndh*F sequence data. Evolution 56:233–252.

Paz MM and Veilleux RE. 1997. Genetic diversity based on randomly amplified polymorphic DNA (RAPD) and its relationship with the performance of diploid potato hybrids. Journal of the American Society of Horticultural Science 122:740–747.

Pejic I, Ajmone-Marsan P, Morgante M, Kozumplick V, Castiglioni P, Taramino G and Motto M. 1998. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs and AFLPs. Theoretical and Applied Genetics 97:1248–1255.

Penteado MID, deMiera LES and de laVega MP. 1996. Genetic resources of *Centrosema* spp: Genetic changes associated to the handling of an active collection. Genetic Resources and Crop Evolution 43:85–90.

Perez JA, Maca N and Larruga JM. 1999. Expanding informativeness of microsatellite motifs through the analysis of heteroduplexes: a case applied to *Solanum tuberosum*. Theoretical and Applied Genetics 99:481–486.

Peroutka SJ. 1997. The medical utility of genomics data in neuropsychiatry: mutational genetics versus association genetics. Current Opinions in Biotechnology 8:688–691.

Petersen L, Ostergard H and Giese H. 1994. Genetic diversity among wild and cultivated barley as revealed by RFLP. Theoretical and Applied Genetics 89:676–681.

Pflieger S, Lefebvre V, Caranta C, Blattes A, Goffinet B and Palloix A. 1999. Disease resistance gene analogs as candidates for QTLs involved in pepper–pathogen interactions. Genome 42:1100–1110.

Phippen WB, Kresovich S, Candelas FG and McFerson JR. 1997. Molecular characterization can quantify and partition variation among genebank holdings: a case study with phenotypically similar accessions of *Brassica oleracea* var. *capitata* L. (cabbage) 'Golden Acre'. Theoretical and Applied Genetics 94:227–234.

Poulton EB. 1904. What is a species? Proceedings of the Entomological Society of London, UK. 1903: lxxvii-cxvi.

Powell W, Machray GC and Provan J. 1996a. Polymorphism revealed by simple sequence repeats. Trends in Plant Science 1:215–222.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S and Rafalsky A. 1996b. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for gemplasm analysis. Molecular Breeding 2:225–238.

Prabhu RR, Webb D, Jessen H, Luk S, Smith S and Gresshoff PM. 1997. Genetic relatedness among soybean genotypes using DNA amplification fingerprinting (DAF), RFLP and pedigree. Crop Science 37:1590–1595.

Provan J, Kumar A, Shepherd L, Powell W and Waugh R. 1996. Analysis of intra-specific somatic hybrids of potato (*Solanum tuberosum*) using simple sequence repeats. Plant Cell Reports 16:196–199.

Queller DC, Strassmann JE and Hughes CR. 1993. Microsatellites and kinship. Trends in Ecology and Evolution 8:285–288.

Rabinowitz D, Linder CR, Ortega R, Begazo D, Murguia H, Douches DS and Quiros CF. 1990. High levels of interspecific hybridization between *Solanum sparsipilum* and *S. stenotomum* in experimental plots in the Andes. American Potato Journal 67:73–81.

Rafalski JA. 2002. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Science 162:329–333.

Ramsay G. 1998. DNA chips: State-of-the art. Nature Biotechnology 16:40–44.

Reed DH and Frankham R. 2001. How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution 55:1095–1103.

Reed DH and Frankham R. 2003. Correlation between fitness and genetic diversity. Conservation Biology 17:230–237.

Reedy ME, Knapp AD and Lamkey KR. 1995. Isozyme allelic frequency changes following maize (*Zea mays* L.) germplasm regeneration. Maydica 40:269–273.

Rhymer JM and Simberloff D. 1996. Extinction by hybridization and introgression. Annual Review of Ecology and Systematics 27:83–109.

Rick CM. 1963. Barriers to interbreeding in *Lycopersicon peruvianum*. Evolution 17:216–232.

Rick CM. 1973. Potential genetic resources in tomato species: clues from observations in native habitats. In Genes, enzymes, and populations (AM Srb, ed.). Plenum Press, New York, USA. pp. 255–269.

Rick CM. 1979. Biosystematic studies in *Lycopersicon* and closely related species in *Solanum*. In The biology and taxonomy of the Solanaceae. Linnean Society of London Symposium. Series 7 (JD Hawkes, RN Lester and AD Skelding, eds.). Academic Press, New York, USA. pp. 667–678 + 1 pl.

Rick CM and Fobes JF. 1974. Association of an allozyme with nematode resistance. Tomato Genetics Cooperative Report 24:25.

Riedel GE, Swanberg SL, Kuranda KD, Marquette K, LaPan P, Bledsoe P, Kennedy A and Lin BY. 1990. Denaturing gradient gel electrophoresis identifies genomic DNA polymorphism with high frequency in maize. Theoretical and Applied Genetics 80:1–10.

Rieger MA, Lamond M, Preston C, Powles SB and Roush RT. 2002. Pollen-mediated movement of herbicide resistance between commercial canola fields. Science 296:2386–2388.

Rieseberg LH. 1991. Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. American Journal of Botany 78:1218–1237.

Rieseberg LH. 1995. The role of hybridization in evolution: old wine in new skins. American Journal of Botany 82:944–953.

Rieseberg LH and Brouillet L. 1994. Are many plant species paraphyletic? Taxon 43:21–32.

Rieseberg LH and Burke JM. 2001. The biological reality of species: gene flow, selection and collective evolution. Taxon 50:47–67.

Rieseberg LH and Ellstrand NC. 1993. What can molecular and morphological markers tell us about plant hybridization? Critical Reviews in Plant Science 12:213–241.

Rieseberg LH, Sinervo B, Linder CR, Ungerer M and Arias DM. 1996. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. Science 272:741–745.

Riesner D, Steger G, Zimmat R, Owens RA, Wagenhofer M, Hillen W, Vollbach S and Henco K. 1989. Temperature-gradient gel electrophoresis of nucleic acids: analysis of conformational transitions, sequence variations, and protein-nucleic acid interactions. Electrophoresis 10:377–89.

Ritala A, Nuutila AM, Aikasalo R, Kauppinen V and Tammisola J. 2002. Measuring gene flow in the cultivation of transgenic barley. Crop Science 42:278–285.

Roa AC, Maya MM, Duque MC, Tohme J, Allem AC and Bonierbale MW. 1997. AFLP analysis of relationships among cassava and other *Manihot* species. Theoretical and Applied Genetics 95:741–750.

Röder MS, Wendehake K, Korzun V, Bredemeijer G, Laborie D, Bertrand L, Isaac P, Rendell S, Jackson J, Cooke RJ, Vosman B and Ganal MW. 2002. Construction and analysis of a microsatellite-based database of European wheat varieties. Theoretical and Applied Genetics 106:67–73.

Rodman J, Price RA and Karol K. 1993. Nucleotide sequences of the *rbc*L gene indicate monophyly of mustard oil plants. Annals of the Missouri Botanical Garden 80:686–699.

Rohlf FJ. 1992. NTSYS-pc, numerical taxonomy and multivariate system. Exeter Publishing, New York, USA.

Rokas A, Williams BL, King N and Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rollins RC. 1965. On the basis of biological classification. Taxon 14:1–6.

Ronaghi M, Uhlén M and Nyrén P. 1998. A sequencing method based on real-time pyrophosphate. Science 281:363.

Ronning CM and Schnell RJS. 1994. Allozyme diversity in a germplasm collection of *Theobroma cacao* L. Journal of Hereditity 85:291–295.

Ross H. 1986. Potato Breeding—Problems and Perspectives. Advances in Plant Breeding Supplement 13. Journal of Plant Breeding. Verlag Paul Parey, Berlin, Germany.

Ruiz de Galerreta JI, Carrasco A, Salazar A, Barrena I, Iturritxa E, Marquinez R, Legorburu FJ and Ritter E. 1998. Wild *Solanum* species as resistance sources against different pathogens of potato. Potato Research 41:57–68.

Russell JR, Fuller JD, Macaulay M, Hatz BG, Jahoor A, Powell W and Waugh R. 1997a. Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theoretical and Applied Genetics 95:714–722.

Russell JR, Fuller JD, Young G, Thomas B, Taramino G, Macaulay M, Waugh R and Powell W. 1997b. Discriminating between barley genotypes using microsatellite markers. Genome 40:442–450.

Saal B and Wricke G. 2002. Clustering of amplified fragment length polymorphism markers in a linkage map of rye. Plant Breeding 121:117–123.

Saeglitz C, Pohl M and Bartsch D. 2000. Monitoring gene flow from transgenic sugar beet using cytoplasmic male-sterile bait plants. Molecular Ecology 9:2035–2040.

Sackville Hamilton NR, Engels JMM, van Hintum TJL, Koo B and Smale M. 2002. Accession management. Combining or splitting accessions as a tool to improve germplasm management efficiency. IPGRI Technical Bulletin No. 5. International Plant Genetic Resources Institute, Rome, Italy.

Sackville Hamilton NR and Chorlton KH. 1997. Regeneration of accessions in seed collections: a decision guide. Handbook for Genebanks No. 5. International Plant Genetic Resources Institute, Rome, Italy.

Saitou N and Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425.

Salisbury PA. 2000. The myths of gene transfer: A canola case study. Plant Protection Quarterly 15:71–76.

Sanger F, Nicklen S and Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences USA 74:5463–5467.

Santacruz-Varela A, Widrlechner MP, Ziegler KE, Salvador RJ, Mallard MJ and Bretting PK. 2004. Phylogenetic relationships among North American popcorns and their evolutionary links to Mexican and South American popcorns. Crop Science 44:1456–1467.

Schierwater B and Ender A. 1993. Different thermostable DNA polymerases may apply to different RAPD products. Nucleic Acids Research 21:4647–4648.

Schittenhelm S, Gladis T and Rao VR. 1997. Efficiency of various insects in germplasm regeneration of carrot, onion and turnip rape accessions. Plant Breeding 116:369–375.

Schneider K, Borchardt DC, Schafer-Pregl R, Nagl N, Glass C, Jeppsson A, Gebhardt C and Salamini F. 1999. PCR-based cloning and segregation analysis of functional gene homologues in *Beta vulgaris*. Molecular Genetics 262:515–524.

Schneider K and Douches DS. 1997. Assessment of PCR-based simple sequence repeats to fingerprint North American potato cultivars. American Potato Journal 74:149–160.

Schoen DJ and Brown AHD. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proceedings of the NationalAcademy of Sciences USA 90:10623–10627.

Schoen DJ and Brown AHD. 1995. Maximising genetic diversity in core collections of wild relatives of crop species. Pp. 55–76 in Core Collections of Plant Genetic Resources (T Hodgkin, AHD Brown, TJL van Hintum and EAV Morales, eds.). John Wiley & Sons, Chichester, UK.

Schoen DJ, David JL and Bataillon TM. 1998. Deleterious mutation accumulation and regeneration of genetic resources. Proceedings of the National Academy of Sciences USA 95:349–399.

Schut JW, Qi X and Stam P. 1997. Association between relationship measures based on AFLP markers, pedigree data and morphological traits in barley. Theoretical and Applied Genetics 95:1161–1168.

Schlötterer C. 2004. The evolution of molecular markers—just a matter of fashion? Nature Reviews Genetics 5:63–69.

Scott MC, Caetano AG and Trigiano RN. 1996. DNA amplification fingerprinting identifies closely related *Chrysanthemum* cultivars. Journal of the American Society of Horticultural Science 121:1043–1048.

Second G. 1982. Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. Japanese Journal of Genetics 57:25–57.

Sharma SK, Dawson IK and Waugh R. 1995. Relationships among cultivated and wild lentils revealed by RAPD analysis. Theoretical and Applied Genetics 91:647–654.

Sharma SK, Knox MR and Ellis THN. 1996. AFLP analysis of the diversity and phylogeny of *Lens* and its comparison with RAPD analysis. Theoretical and Applied Genetics 93:751–758.

Shashidhara G, Hema MV, Koshy B and Farooqi AA. 2003. Assessment of genetic diversity and identification of core collection in sandalwood germplasm using RAPDs. Journal of Horticutural Science and Biotechnology 78:528–536.

Shenoy VV, Seshu DV and Sachan JKS. 1990. Shikimate dehydrogenase-1(2) allozyme as a marker for high seed protein content in rice. Crop Science 30:937–940.

Shimamura M, Yasue H, Oshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I and Okada N. 1997. Molecular evidence from retrotransposons that whales form a clade within even-toed ungulates. Nature 388:666–670.

Sicard D, Woo SS, Arroyo-Garcia R, Ochoa O, Nguyen D, Korol A, Nevo E and Michelmore R. 1999. Molecular diversity at the major cluster of disease resistance genes in cultivated and wild *Lactuca* spp. Theoretical and Applied Genetics. 99:405–418.

Singh RJ, Kollipara KP and Hymowitz T. 1998. The genomes of *Glycine canescens* F.J. Herm., and *G. tomentella* Hyata of Western Australia and their phylogenetic relationships in the genus *Glycine* Willd. Genome 41:669–679.

Singh AK and Smartt J. 1998. The genome donors of the groundnut/peanut (*Arachis hypogaea* L.) revisited. Genetic Resources and Crop Evolution 45:113–118.

Skogsmyr I. 1994. Gene dispersal from transgenic potatoes to conspecifics: a field trial. Theoretical and Applied Genetics 88:770–774.

Skroch PW, Dobert RC, Triplett EW and Nienhuis J. 1993. Polymorphism of the leghemoglobin gene in *Phaseolus* demonstrated by polymerase chain reaction amplification. Euphytica 69:177–183.

Skroch PW, Nienhuis J, Bebee S, Tohme J and Pedraza F. 1998. Comparison of Mexican common bean (*Phaseolus vulgaris* L.) core and reserve germplasm collections. Crop Science 38:488–496.

Small RL, Cronn RC and Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany 17:145–170.

Smartt J and Simmonds NW. 1995. Evolution of crop plants. Longman Scientific and Technical, Essex, UK.

Smith OS, Smith JSC, Bowen SL, Tenborg RA and Wall SJ. 1990. Similarities among a group of elite maize inbreds as measured by pedigree, F_1 grain yield, grain yield, heterosis and RFLPs. Theoretical and Applied Genetics 80:833–840.

Sneath PHA and Sokal RR. 1962. Numerical taxonomy: the principles and practice of numerical classification. W.H. Freeman and Company, New York.

Snow AA and Palma PM. 1997. Commercialization of transgenic plants: potential ecological risks. BioScience 47:86–96.

Sokal RR. 1985. The continuing search for order. American Naturalist 126:729–749.

Sokal RR and Crovello TJ. 1970. The biological species concept: a critical evaluation. American Naturalist 104:127–153.

Soltis DE, Soltis PE, Morgan DR, Swensen SM, Mullin BC, Down JM and Martin PG. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proceedings of the National Academy of Sciences USA 92:2647–2651.

Somers DJ and Demmon G. 2002. Identification of repetitive, genome-specific probes in crucifer oilseed species. Genome 45:485–492.

Song K, Liu P and Osborn TC. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proceedings of the National Academy of Sciences USA 92:7719–7723.

Sonnate G, Stockton T, Nodari RO, Becerra Velásquez VL and Gepts P. 1994. Evolution of genetic diversity during the domestication of common bean (*Phaseolus vulgaris* L.). Theoretical and Applied Genetics 89:629–635.

Spagnoletti-Zeuli PL, Sergio L and Perrino P. 1995. Changes in the genetic structure of wheat germplasm accessions during seed rejuvenation. Plant Breeding 114:193–198.

Spooner DM, Hetterscheid WLA van den Berg RG and Brandenburg W. 2003. Plant nomenclature and taxonomy: An horticultural and agronomic perspective. Horticulture Reviews 28:1–60.

Spooner D and Lara-Cabrera S. 2001. Sistemática molecular y evolución de plantas cultivadas. In Enfoques Contemporáneos para el estudio de la biodiversidad (HM Hernández, A García-Aldrete, F Alvarez and M. Ulloa, eds.). Instituto de Biología, UNAM/ Fondo de Cultura Económica, Mexico. pp. 57–114.

Spooner DM, Tivang J, Nienhuis J, Miller JT, Douches DS and Contreras-M. A. 1996. Comparison of four molecular markers in measuring relationships among the wild potato relatives *Solanum* section *Etuberosum* (Subgenus *Potatoe*). Theoretical and Applied Genetics 92:532–540.

Spooner DM, Sytsma KJ and Smith JF. 1991. A molecular reexamination of diploid hybrid speciation of *Solanum raphanifolium*. Evolution 45:757–764.

Stahlhut R, Park G, Petersen R, Ma W and Hylands P. 1999. The occurrence of the anti-cancer diterpene taxol in *Podocarpus gracilior* Pilger (Podocarpaceae). Biochemical Systematics and Ecology 27L:613–622.

St. Amand PC. 2004. Risks associated with genetically engineered crops. In Genetically modified crops: their development, uses, and risks (G H Liang and D Skinner, eds.). Haworth Press, Inc. Binghamton, New York, USA.

St. Amand PC, Skinner DZ and Peaden RN. 2000. Risk of alfalfa transgene dissemination and scale-dependent effects. Theoretical and Applied Genetics 101:107–114.

Staub JE, Serquen FC and Gupta M. 1996. Genetic markers, map construction, and their application in plant breeding. HortScience 31:729–741.

Steiner JJ, Piccioni E, Falcinelli M and Liston A. 1998. Germplasm diversity among cultivars and the NPGS crimson clover collection. Crop Science 38:263–271.

Steiner AM, Ruckenbauer P and Goecke E. 1997. Maintenance in genebanks, a case study: contaminations observed in the Nurnberg oats of 1831. Genetic Resources and Crop Evolution 44:533–538.

Steinmetz LM, Mindrinos M and Oefner PJ. 2000. Combining genome sequences and new technologies for dissecting the genetics of complex phenotypes. Trends in Plant Science 5:397–401.

Stevens PF. 1998. What kind of classification should the practicing taxonomist use to be saved? In Plant diversity in Malesia III: Proceedings of the 3rd International Flora Malesiana Symposium 1995 (J Drandsfield, MJE Coode and DA Simpson, eds.). Royal Botanic Gardens, Kew, UK. pp. 295–319

Struss D and Plieske J. 1998. The use of microsatellite markers for detection of genetic diversity in barley populations. Theoretical and Applied Genetics 97:308–315.

Stuber CW, Polacco M and Lynn Sr. M. 1999. Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Science 39:1571–1583.

Stuessy TF. 1990. Plant taxonomy: the systematic evaluation of comparative data. Columbia University Press, New York, USA.

Swanson T. 1996. Global values of biological diversity: the public interest in the conservation of plant genetic resources for agriculture. Plant Genetic Resources Newsletter 105:1–7.

Sytsma KJ and Hahn W. 1997. Molecular systematics: 1994-1995. Progress in Botany 58:470–499.

Takaiwa F, Oono K and Sugiura M. 1985. Nucleotide sequence of the 17S - 25S spacer region from rice rDNA. Plant Molecular Biology 4:355–364.

Tang X and Zhang W. 1992. Studies on pre-selection of dwarf apple seedlings by starch gel electrophoresis. Acta Horticulturae 317:29–34.

Tanksley SD and McCouch SR. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066.

Tanksley SD and Nelson JC. 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theoretical and Applied Genetics 92:191–203.

Tanksley SD and Orton TJ. 1983. Isozymes in plant genetics and breeding. Elsevier Science Publishers, Amsterdam, The Netherlands.

Tao R and Sugiura A. 1987. Cultivar identification of Japanese persimmon by leaf isozymes. HortScience 22:932–935.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BAJ and Tunnacliffe A. 1992. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 13:718–725.

Templeton AR. 1986. Coadaptation and outbreeding depression. in Conservation biology: the science of scarcity and diversity (M Soulé, ed.). Sinauer Press, Sunderland, Massachusetts, USA. pp. 105–116.

Templeton, AR. 1989. The meaning of species and speciation: a genetic perspective. In Speciation and its consequences (D Otte and JA Endler, eds.). Sinauer Associates, Inc., Sunderland, Massachusetts, USA. pp. 3–27

Thiel T, Michalek W, Varshney RK and Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics 106:411–422.

Thieme R, Darsow U, Gavrilenko T, Dorokhov D and Tiemann H. 1997. Production of somatic hybrids between *S. tuberosum* L. and late blight resistant Mexican wild potato species. Euphytica 97:189–200.

Thorman CE, Ferreira ME, Camargo LEA, Tivang JG and Osborn TC. 1994. Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. Theoretical and Applied Genetics 88:973–980.

Timmons AM, O'Brien ET, Charters YM, Dubbels SJ and Wilkonson MJ. 1995. Assessing the risks of wind pollination from fields of genetically modified *Brassica napus* ssp. *oleifera*. Euphytica 85:417–423.

Tohme J, Gonzalez DO, Beebe S and Duque MC. 1996. AFLP analysis of gene pools of a wild bean core collection. Crop Science 36:1375–1384.

Tsaftaris SA and Shull GH. 1995. Molecular aspects of heterosis in plants. Physiology of Plants 94:362–370.

Tsegaye S, Tesemma T and Belay G. 1996. Relationships among tetraploid wheat (*Triticum turgidum* L) landrace populations revealed by isozyme markers and agronomic traits. Theoretical and Applied Genetics 93:600–605.

Urrea CA, Miklas PN, Beaver JS and Riley RH. 1996. A codominant randomly amplified polymorphic DNA (RAPD) marker useful for indirect selection of bean golden mosaic virus resistance in common bean. Journal of the American Society of Horticultural Science 121:1035–1039.

Ursla FWM, Hayward MD and Kearsey MJ. 1997. Isozyme and quantitative traits polymorphisms in European provenances of perennial ryegrass (*Lolium perenne* L.). Euphytica 93:263–269.

van der Linden G, Smulders MJM and Vosman B. 2004. Motif-directed profiling: a glance at molecular evolution. In: Plant species-level systematics: new perspectives on pattern and process (FT Bakker, LW Chatrou, B Gravendeel and PB Pelser, eds.). Regnum Vegetabile 142. Koeltz, Königstein (in press).

van Hintum TJL. 1994. Comparison of marker systems and construction of a core collection in a pedigree of European spring barley. Theoretical and Applied Genetics 89:991–997.

van Hintum TJL, Boukema IW and Visser DL. 1996. Reduction of duplication in a *Brassica oleracea* germplasm collection. Genetic Resources and Crop Evolution 43:343–349.

van Hintum TJL, Brown AHD, Spillane C and Hodgkin T. 2000. Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy.

van Hintum TJL and Visser DL. 1995. Duplication within and between germplasm collections. II. Duplication in four European barley collections. Genetic Resources and Crop Evolution 42:135–145.

van Hintum TJL, von Bothmer R and Visser DL. 1995. Sampling strategies for composing a core collection of cultivated barley (*Hordeum vulgare* s. lat.) collected in China. Hereditas 122:7–17.

van Raamsdonk LWD and van der Maesen LJG. 1996. Crop-weed complexes: The complex relationship between crop plants and their wild relatives. Acta Botanica Neerlandica 45:135–155.

van Treuren R, Bijlsma R, Tinbergen JM, Heg D and van de Zande L. 1999. Genetic analysis of the population structure of socially organized oystercatchers (*Haematopus ostralegus*) using microsatellites. Molecular Ecology 8:181–187.

van Treuren R, Magda A, Hoekstra R and van Hintum TJL. 2004. Genetic and economic aspects of marker-assisted reduction of redundancy from a wild potato germplasm collection. Genetic Resources and Crop Evolution 51:277–290.

van Treuren R and van Hintum TJL. 2001. Identification of intra-accession genetic diversity in selfing crops using AFLP markers: implications for collection management. Genetic Resources and Crop Evolution. 48:287–295.

van Treuren R, van Soest LJM and van Hintum TJL. 2001. Marker-assisted rationalisation of genetic resources collections: a case study in flax using AFLPs. Theoretical and Applied Genetics 103:144–152.

van Valen L. 1976. Ecological species, multispecies and oaks. Taxon 25:233–239.

Varghese YA, Knaak C, Sethuraj MR and Ecke W. 1997. Evaluation of random amplified polymorphic DNA (RAPD) markers in *Hevea brasiliensis*. Plant Breeding 116:47–52.

Vignani R, Bowers JE and Meredith CPL. 1996. Microsatellite DNA polymorphism analysis of clones of *Vitis vinifera* 'Sangiovese'. Science and Horticulture 65:163–169.

Virk PS, Newbury HJ, Jackson MT and Ford-Lloyd BV. 1995. The identification of duplicate accessions within a rice germplasm collection using RAPD analysis. Theoretical and Applied Genetics 90:1049–1055.

Virk PS, Newbury HJ, Jackson MT and Ford-Lloyd BV. 2000a. Are mapped markers more useful for assessing genetic diversity? Theoretical and Applied Genetics 100:607–613.

Virk PS, Zhu J, Newbury HJ, Bryan GJ, Jackson MT and Ford-Lloyd BV. 2000b. Effectiveness of different classes of molecular marker for classifying and revealing variation in rice (*Oryza sativa*) germplasm. Euphytica 112:275–284.

Von Bothmer R and Seberg O. 1995. Strategies for the collecting of wild species. Pp. 93-111 in Collecting Plant Genetic Diversity, Technical Guidelines (L Guarino, V Ramanatha Rao and R Reid, eds.). CAB International, Wallingford, UK.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M and Zabeau M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23:4407–4414.

Wagner DB and Allard RW. 1991. Pollen migration in predominantly self-fertilizing plants: barley. Journal of Hereditity 82:302–304.

Wang Z, Weber JL, Zhong G and Tanksley SD. 1994. Survey of plant short tandem DNA repeats. Theoretical and Applied Genetics 88:1–6.

Warburton FE. 1967. The purposes of classifications. Systematic Zoology 16:241–245.

Warnke SE, Douches DS and Branham BE. 1998. Isozyme analysis supports allotetrapoloid inheritance in tetraploid creeping bluegrass (*Agrostis palustris* Huds.). Crop Science 38:801–805.

Waser NM. 1993. Population structure, optimal outbreeding, and assortative mating in angiosperms. In The natural history of inbreeding and outbreeding, theoretical and empirical perspectives (NW Thornhill, ed.). University of Chicago Press, Chicago, USA. pp. 173–199

Watanabe KN, Orrillo M, Vega S, Valkonen JPT, Pehu E, Hurtado A and Tanksley SD. 1995. Overcoming crossing barriers between nontuber-bearing and tuber-bearing *Solanum* species: towards potato germplasm enhancement with a broad spectrum of solanaceous genetic resources. Genome 38:27–35.

Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BBT and Powell W. 1997. Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular Genetics and Genomics 253:687–694.

Waycott W and Fort SB. 1994. Differentiation of nearly identical germplasm accessions by a combination of molecular and morphologic analyses. Genome 37:577–583.

Weaver KR, Callahan LM, Caetano Anolles G and Gresshoff PM. 1995. DNA amplification fingerprinting and hybridization analysis of centipedegrass. Crop Science 35:881–885.

Weising K, Nybom H, Wolff K and Kahl G. 2005. DNA fingerprinting in plants: principles, methods, and applications CRC Press, Boca Raton, Florida, USA.

Weising K, Winter P, Huttel B and Kahl G. 1998. Microsatellite markers for molecular breeding. Journal of Crop Production 1:113–143.

Welsh J and McClelland M. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research 18:7213–7218.

Wendel JF. 1995. Cotton. In Evolution of Crop Plants (J Smart and NW Simmonds, eds). Longman, Essex, UK. pp. 358–366

Wendel JF and Doyle JJ. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. In Molecular systematics of plants II: DNA sequencing (DE Soltis, PS Soltis and JJ Doyle, eds.). Kluwer Academic Publishers, Boston, USA. pp. 265–296

Westerbergh A and Doebley J. 2002. Morphological traits defining species differences in wild relatives of maize are controlled by multiple quantitative trait loci. Evolution 56:273–283.

Whitkus R, de la Cruz M, Mota Bravo L, Gomez Pompa A and De la Cruz M. 1998. Genetic diversity and relationships of cacao (*Theobroma cacao* L.) in southern Mexico. Theoretical and Applied Genetics 96:621–627.

Whitkus R, Doebley J and Wendel JF. 1994. Nuclear DNA markers in systematics and evolution. In DNA-based markers in plants (Advances in cellular and molecular biology of plants, vol. 1.) (RL Phillips and IK Vasil, eds.). Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 116–141.

Wiley EO, Brooks DR, Siegel-Causey D and Funk VA. 1991. The complete cladist: a primer of phylogenetic procedures. Museum of Natural History, Lawrence, KA, USA.

Williams CE, Yanagihara S, McCouch SR, Mackill DJ and Ronald PC. 1997. Predicting success of indica/japonica crosses in rice, based on a PCR marker for the S-5(n) allele at a hybrid-sterility locus. Crop Science 37:1910–1912.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA and Tingey SV. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18:6531–6535.

Witsenboer H, Vogel J and Michelmore RW. 1997. Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca spp*.). Genome 40:923–936.

Wolfe AD and Liston A. 1998. Contributions of PCR-based methods to plant systematics and evolutionary biology. In Plant molecular systematics II (DE Soltis, PS Soltis and JJ Doyle, eds.). Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 43–86

Wolfe KH, Li W-H and Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences USA 84:9054–9058.

Wolff K, Rogstad SH and Schaal BA. 1994. Population and species variation of minisatellite DNA in *Plantago*. Theoretical and Applied Genetics 87:733–740.

Xiao J, Grandillo S, Ahn SN, McCouch SR, Tanksley SD, Li J and Yuan L. 1996. Genes from wild rice improve yield. Nature 384:223–224.

Yabuno T. 1962. Cytotaxonomic studies on the two cultivated species and the wild relatives in the genus *Echinochola*. Cytologia 27:296–305.

Yamada T, Misoo S, Ishii T, Ito Y, Takaoka K and Kamijima O. 1997. Characterization of somatic hybrids between tetraploid *Solanum tuberosum* L. and dihaploid *S. acaule*. Breeding and Science 47:229–236.

Yang WP, De Oliveira AC, Godwin I, Schertz K and Bennetzen JL. 1996. Comparison of DNA marker technologies in characterizing plant genome diversity: variability in Chinese sorghums. Crop Science 36:1669–1676.

Young ND. 1999. A cautiously optimistic vision for marker-assisted breeding. Molecular Breeding 5:505-510.

Young WP, Schupp JM and Keim P. 1999. DNA methylation and AFLP marker distribution in the soybean genome. Theoretical and Applied Genetics 99:785–792.

Zamir D. 2001. Improving plant breeding with exotic genetic libraries. National Review of Genetics 2:983–989.

Zerega NJC, Ragone D and Motley TJ. 2004. Complex origins of breadfruit (*Artocarpus altilis*, Moraceae): implications for human migrations in Oceania. American Journal of Botany 91:760–766.

Zeven AC, Dehmer KJ, Gladis T, Hammer K and Lux H. 1998. Are the duplicates of perennial kale (*Brassica oleracea* L. var. *ramosa* DC.) true duplicates as determined by RAPD analysis? Genetic Resources and Crop Evolution 45:105–111.

Zhou Z, Bebeli PJ, Somers DJ and Gustafson JP. 1997. Direct amplification of minisatellite-region DNA with VNTR core sequences in the genus *Oryza*. Theoretical and Applied Genetics 95:942–949.

Zietkiewicz E, Rafalski A and Labuda D. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183.

Zoro Bi I, Maquet A, Degreef J, Wathelet B and Baudoin JP. 1998. Sample size for collecting seeds in germplasm conservation: the case of the Lima bean (*Phaseolus lunatus* L.). Theoretical and Applied Genetics 97:187–194.

Glossary

Agarose: A chain of sugar molecules that form the basis of agarose gels used for electrophoresis.

Allele mining: A research field directed to the identification of useful alleles within genetic resources collections.

Allozymes: Allelic forms of an enzyme that can be distinguished by gel electrophoresis (see isozyme).

Amplicons: DNA fragments amplified by PCR.

Amplified Fragment Length Polymorphism (AFLP): A molecular marker technique that targets variation in DNA restriction sites and in DNA restriction fragments.

Apospecies: A species concept based on cladistics that does not insist on monophyly. It recognizes species pairs, one a monophyletic daughter species (apospecies) and the other a paraphyletic progenitor species (plesiospecies).

Arbitrarily Primed Polymerase Chain Reaction (AP-PCR): A variant of the RAPD technique that uses longer arbitrary primers than RAPDs.

Association genetics: A research field directed to the identification of correlations between phenotypic traits and genetic markers with the aim to identify and locate the underlying genes in the genome.

Basic Local Alignment Search Tool (BLAST): A bioinformatics tool to find matches between a DNA sequence and known sequences stored in databases.

Bayesian analysis: A statistical approach for constructing phylogenetic trees related to maximum likelihood that operates on a priori weighting factors and probabilities (see maximum likelihood, parsimony).

Bootstrap analysis: A method in cladistic analysis to infer the "strength" or "confidence" of a branch on a phylogenetic tree, obtained by generating trees many times from a sample distribution of characters. Bootstrap values theoretically can vary from 0% (poor support) to 100% (excellent support).

Capillary electrophoresis: A technique to separate DNA fragments in an electric field, carried out within narrow tubes (capillaries).

cDNA-AFLP: A molecular marker technique performing AFLP analysis on cDNA.

Cladistic Species Concept: A philosophy and set of methods that use cladistic criteria to determine the limits of species.

Cladistics: A branch of biology that determines evolutionary branching orders or trees of descent based on derived similarities (see phenetics).

Cladogram: A branching phylogenetic tree of individuals or taxa, rooted on an outgroup(s) produced by a method that minimizes evolutionary changes (by parsimony, maximum likelihood, or other methods) of characters believed to be homologous among a group of organisms.

Cleaved Amplified Polymorphic Sequence (CAPS): A molecular marker technique that is based on the amplification of DNA fragments by PCR followed by DNA restriction of the fragments.

Codominant marker: A genetic marker for which all alleles are expressed when co-occurring in an individual.

Comparative genomics: A research field directed to the comparison of genomes from different species in order to obtain a better understanding of the evolution of species and the location and functioning of genes.

Complementary DNA (cDNA): In vitro generation of DNA constructed from mRNA.

Concatenated dataset: A combined dataset that connects together many individual datasets into one, like links in a chain, so that analyses can be done as a single unit.

Concerted evolution: A process whereby repetitive DNA families maintain one type of sequence within the repeat (become homogenized), through genetic mechanisms of unequal crossing over and gene conversion.

Congruence: Agreement in results of as taxonomic analysis; refers to both phenetic and cladistic results.

Conserved Orthologous Set Markers (COS): Conservative molecular markers that are used as anchors for map development in comparative genomics studies.

Core collection: A subset of the entire germplasm collection that incorporates a representative sample of variation with a minimum of redundancy. This is an attempt to bring the entire collection to a workable size for economic or space or other constraints, and to facilitate its use, by choosing accessions that represent its most representative or useful diversity.

Dendrogram: A branching diagrammatic representation of a set of individuals or taxa, constructed from overall similarity of a set of characters among organisms.

Degenerate Oligonucleotide Primed-PCR (DOP-PCR): A molecular marker technique that uses partially degenerated primers for polymorphism detection in comparative genomics studies.

Deoxynucleotides: The components of DNA: Adenine (A), Cytocine (C), Guanine (G) or Thymidine (T).

Directed Amplification of Minisatellite-region DNA (DAMD): A technique related to ISSR analysis that uses a single primer containing only the core motif of a minisatellite.

DNA Amplification Fingerprinting (DAF): A variant of the RAPD technique that uses shorter, 5–8 bp primers to generate a larger number of fragments.

DNA chip: Also known as microarray. A high throughput screening technique based on the hybridization between oligonucleotide probes and either DNA or mRNA, carried out on miniaturized reaction surfaces.

DNA sequencing: The determination of the sequence of deoxynucleotides in DNA fragments.

Diversity study: The use of morphological or molecular markers to assess the diversity of a set of related accessions.

Domestication syndrome: A set of similar traits that confer adaptation to the cultivated environment. Specific traits will vary among different crops.

Dominant marker: A genetic marker for which only a single allele is expressed when multiple alleles are co-occurring in an individual.

Eclectic species concept: A philosophy that species are defined and formed and maintained by a variety of morphological, interbreeding, ecological, and phylogenetic factors.

Ecological farming: A farming system that aims to develop an integrated, humane, environmentally and economically sustainable agricultural production system.

Ecological species concept: A philosophy that ecological constraints are the primary factor in forming and maintaining a species.

Electrophoresis: A technique to separate proteins or DNA fragments in an electric field.

EST-SNP: SNP analysis targeted to ESTs.

EST-SSR: Microsatellite analysis targeted to ESTs.

Exotic libraries: Collections of elite crop lines containing defined genomic regions from wild species, to provide pre-breeding material for modern varieties.

Expressed Sequence Tag (EST): A DNA sequence derived from transcribed regions of a genome.

Extant: Existing or living at the present time, in contrast to extinct (no longer living).

Fluorescent labeling: The labeling of probes or primers with fluorescent tags in order to enable detection of variation in DNA fragments.

Functional diversity: Genetic diversity as assessed by variation in transcribed regions of the genome that is known to be associated with a biological function.

Gene flow: The natural flow of genes within a population or from one population to another by interbreeding or migration.

Genetic distance: A measure to quantify the genetic relatedness between individuals or populations.

Genetic drift: Fluctuations in genetic variation between generations as a result of random processes.

Genomic DNA: The full complement of DNA contained in the genome of a cell or organism.

Geographic Information Systems (GIS): A set of computer software designed to capture, organize, store, and analyze geographically referenced (spatial) information.

Heterotic groups: Groups of germplasm that when crossed maximize heterosis. Heterosis is a phenomenon that heterozygotes in a population often have higher fitness than the homozygotes.

Heterozygosity: The condition of having one or more pairs of different alleles on homologous chromosomes.

Homologous: Characters that arise by common descent.

Homoplasy: A term in cladistic analysis that refers to the proportion of parallelisms and reversals on a phylogenetic tree. Also used for different DNA fragments of identical size that cannot be distinguished by gel electrophoresis.

Ingroup: A putatively monophyletic group that is the prime subject of a cladistic analysis.

Interbreeding species concept: A philosophy and set of methods that define species almost entirely on the ability of species to exchange genes naturally or artificially, as assessed by artificial crossing programs, studies of mechanisms to facilitate gene flow, and biological isolating mechanisms.

Internal Transcribed Spacer (ITS): A sequence of nuclear ribosomal DNA commonly examined for phylogenetic analysis consisting of two spacer regions intercalated between the 18S, 5.8S, and 26S genes.

Inter-retrotransposon Amplified Polymorphism (IRAP): A molecular marker technique that targets variation in retrotransposon insertion sites.

Inter Simple Sequence Repeat (ISSR): A molecular marker technique that targets variation in the DNA between two microsatellite loci.

Interspecific: Refers to studies among species.

Intraspecific: Refers to studies of taxa or populations within a species.

Isozymes: Enzymes that have the same chemical function as another enzyme but differ in structure as a result of different amino acid composition.

Leucine Rich Repeat (LRR): Conserved domain of disease resistance genes.

Linkage disequilibrium: The non-random association of the alleles from different loci in the gametes.

Linked markers: The presence of genes (or molecular probes) close to each other on chromosomes so that they are never completely independently assorted; the degree of linkage is greater the closer the genes or markers are on the same chromosome.

Long branch attraction: A phenomenon in cladistic analyses where strongly unequal rates of evolutionary change in different members of a group cause cladistics to produce incorrect trees.

Luciferase: An enzyme used to determine the concentration of ATP as a measurable real-time light signal in pyrosequencing.

Mantel test: A test that computes the linear correlation between two proximity matrices (dissimilarity or similarity), used in phenetics to test whether results from different analyses of the same taxa are similar or different.

Mapped markers: Molecular markers with known chromosomal locations.

Marker index: A way to assess the comparative degree of information provided by different molecular marker systems, as assessed by the product of heterozygosity and multiplex ratio.

Maximum likelihood: A set of methods used to construct cladograms based on certain evolutionary models of character state changes (see Bayesian analysis, parsimony).

Microarray: Also known as DNA chip. A high throughput screening technique based on the hybridization between oligonucleotide probes and either DNA or mRNA, carried out on miniaturized reaction surfaces.

Microsatellite: A molecular marker technique that targets tandem repeats of a small (1–6 base pairs) nucleotide repeat motif.

Minisatellites: A molecular marker technique that targets tandem repeats of a large (10–50 base pairs) nucleotide repeat motif.

Monophyletic: A group that includes an ancestral species and all of its descendants.

Morphological Species Concept: A philosophy and set of methods that define species entirely on morphological or anatomical characters.

mRNA: Messenger RNA.

Multiple Arbitrary Amplicon Profiling (MAAP): A collective term for techniques using single arbitrary primers, such as AP-PCR, DAF, and RAPD.

Multiplex ratio: A measure of the number of bands simultaneously analyzed per DNA marker assay (experiment), that is, the number of bands resolved on a particular gel.

Neutral marker diversity: Genetic diversity as assessed by variation in non-transcribed regions of the genome that are not known to be associated with a biological function.

Nominalistic Species Concept: A philosophy that questions the very existence of species, and believes that individuals or interbreeding populations are the only population system with any objective reality.

Nucleotide Binding Site (NBS): A conserved domain of disease resistance genes.

Nucleotide Binding Site-Directed Profiling (NBS-DP): A molecular marker technique that targets variation at disease resistance genes and analogues.

Oligonucleotide probe: A DNA fragment consisting of nucleotides that is used to detect the presence of its complementary sequence in a DNA sample by hybridization testing.

Orthologous: Characters that are homologous from a speciation event, that is, identical by descent.

Outgroup: Any group, used to root a phylogenetic tree in a cladistic analysis, which is not a member of the taxon group being studied.

Paralogous: Characters that have arisen as a result of gene duplication.

Paraphyletic: A non-monophyletic group containing some, but not all representatives of a taxon; said another way, an incomplete group of descendants from one common ancestor with one or more descendants missing.

Parsimony: A set of methods that assumes that the simplest solution is the most likely one. It is used to construct cladograms, and assumes that minimizing the number of character state changes on a tree is the best approximation of phylogenetic history. (see Bayesian analysis, maximum likelihood).

Pedigree: A list of ancestors (often displayed on a branching tree) based on known relationships from formal records of crosses.

Phenetics: A branch of biology that determines overall similarity of organisms, not evolutionary relationships (see cladistics).

Phenogram: A branching tree of individuals or taxa based on phenetics.

Plesiospecies: A species defined by a cladistic concept that does not insist on monophyly. It recognizes species pairs, one a monophyletic daughter species (apospecies) and the other a paraphyletic progenitor species (plesiospecies).

Polyacrylamide: A chemical that forms the basis of polyacrylamide gels used for electrophoresis.

Polymerase Chain Reaction (PCR): A common molecular technique used to generate numerous copies of specific DNA fragments in vitro.

Polyphyletic: A non-monophyletic group where the common ancestor is placed in another taxon; said another way, a group in which the members do not ultimately derive from one common ancestor, where the descendants of one or more other groups are included.

Pre-breeding/Pre-competitive breeding: The development of germplasm with a genetically broader base for utilization by breeders, such as the introduction of exotic germplasm into a cultivar.

Predictive component of taxonomy: The idea that inferences or predictions can be made from taxonomy.

Proximity matrices: A numerical measure of similarity (or dissimilarity) among objects, used in multivariate analysis, as assessed by various formulas.

Pyrophosphate: A chemical molecule that is released each time a deoxynucleotide is incorporated to the new DNA strand during pyrosequencing.

Pyrosequencing: A new sequencing method of relatively short DNA templates based on real-time (quantitative) pyrophosphate release.

Radioactive labeling: The labeling of probes or primers with radioactive tags in order to enable detection of variation in DNA fragments.

Random Amplified Polymorphic DNA (RAPD): A molecular marker technique that uses primers of random sequence to amplify DNA fragments by PCR.

Redundancy: Refers to identical or near identical items. In germplasm analyses it refers to (near) duplicate germplasm accessions.

Resistance Gene Homologue Polymorphism (RGHP): A group of molecular marker techniques that target groups of resistance genes by PCR using primers aimed at conserved domains of resistance genes.

Restriction Fragment Length Polymorphism (RFLP): A molecular marker technique that targets variation in DNA restriction sites and in DNA restriction fragments.

Re-synthesized/Recreated: In studies of hybrid origins of species, investigators may attempt to "re-synthesize" a hybrid by crossing its putative parents and compare the natural to the putative hybrid by morphological or molecular markers.

Retrotransposon-Based Insertional Polymorphism (RBIP): A molecular marker technique that targets variation in retrotransposon insertion sites.

Retrotransposon-Microsatellite Amplified Polymorphism (REMAP): A molecular marker technique that targets variation in retrotransposon insertion sites.

Selective Fragment Length Amplification (SFLA): A synonym used for AFLP.

Selective neutrality: The state in which genetic variation is influenced only by random processes.

Selective Restriction Fragment Amplification (SRFA): A synonym used for AFLP.

Selectively Amplified Microsatellite Polymorphic Locus (SAMPL): A variation of the AFLP technique that amplifies microsatellite loci by using a single AFLP primer in combination with a primer complementary to compound microsatellite sequences, which do not require prior cloning and characterization.

Sequence Characterized Amplified Region (SCAR): A molecular marker technique that amplifies DNA fragments by PCR using specific primers, designed from nucleotide sequences established from cloned RAPD fragments linked to a trait of interest.

Sequence-Related Amplified Polymorphism (SRAP): A molecular marker technique that targets subsets of open reading frames from coding sequences in the genome.

Sequence-Specific Amplified Polymorphism (S-SAP): A dominant, multiplex marker system for the detection of variation in DNA flanking a retrotransposon insertion site

Simple Sequence Repeat (SSR): A synonym used for microsatellites.

Single Nucleotide Polymorphism (SNP): A molecular marker technique to detect changes in a single nucleotide position.

Single Primer Amplification Reaction (SPAR): A techniques related to ISSR analysis that uses a single primer containing only the core motif of a microsatellite.

Single-Strand Conformation Polymorphism (SSCP): A molecular marker technique that uses PCR and gel electrophoresis of single-strand DNA to detect nucleotide sequence variation among amplified DNA fragments.

Sympatric: Refers to two or more populations that occupy overlapping geographic areas.

TaqMan™: A trademark term for a high throughput, closed tube assay to detect specific sequences in PCR products.

Target Region Amplification Polymorphism (TRAP): A molecular marker technique related to SRAP, but using a fixed primer designed from a targeted EST sequence.

Template DNA: The "template" DNA that is used for the initial reaction of a molecular marker analysis.

Tiling strategy: The detection of allelic variation at genes of known basic sequence by hybridization tests on microarrays using large series of sequence-overlapping probes.

Variable Number of Tandem Repeats (VNTR): A synonym of minisatellites.

FUTURE HARVEST

IPGRI is
a Future Harvest Centre
supported by the
Consultative Group on
International Agricultural
Research (CGIAR)

ISBN-13: 978-92-9043-684-3
ISBN-10: 92-9043-684-0