1

Genomics Meets Biodiversity: Advances in Molecular Marker Development and Their Applications in Plant Genetic Diversity Assessment

Péter Poczai[1,2], Ildikó Varga[2], Neil E. Bell[1,3] and Jaakko Hyvönen[1]
[1]Plant Biology, University of Helsinki, Helsinki
[2]Department of Plant Science and Biotechnology, Georgikon Faculty
University of Pannonia, Keszthely
[3]Botanical Museum, University of Helsinki, Helsinki
[1,3]Finland
[2]Hungary

1. Introduction

Genetic diversity is the fundamental source of biodiversity – the total number of genetic characters contributing to variation within species. In other words it is the measure that quantifies the variation found within a population of a given species. Genetic diversity among individuals reflects the presence of different alleles in the gene pool, and hence different genotypes within populations. Genetic diversity should be distinguished from genetic variability, which describes the tendency of genetic traits found within populations to vary (Laikre et al., 2009). Since the beginning of the 20th century, the study of genetic diversity has been the major focus of core evolutionary biology. The theoretical metrics developed, such as genetic variance and heritability (Fisher, 1930; Wright, 1931), provided the quantitative standards necessary for the evolutionary synthesis. Further research has focused on the origin of genetic diversity, its maintenance and its role in evolution. Simple questions such as "who breeds with whom" initiated studies on the relatedness of populations. These investigations led to the formation of metapopulation theory, where a group of spatially separated populations of the same species interact at some level and form a coherent larger group (Hanski, 1998). The discovery of spatial structure in populations was a key element in the early concepts and models of population ecology, genetics and adaptive evolution (Wright, 1931; Andrewartha & Birch, 1954). How different levels of genetic variance affect the rate of evolutionary change within populations has also been intensively studied. Such changes were originally studied using phenotypic markers: variation among individual plants in traits, such as leaf shape or flower color (Ward et al., 2008). Subsequently the detection of genetic variation has become more sensitive, firstly through the utilization of variation in enzymes (allozymes) and then through PCR-based marker systems allowing direct examination of DNA sequence variation. The precise detection of genetic variation/diversity has greatly enhanced studies of evolution. There is no doubt that genetic variation influences the fitness of individuals, and that this is reflected in natural selection. In this regard, individual genotypes must vary in ecologically important ways. Does this mean that differences in genetic diversity levels will have predictable ecological consequences? The answer is no, because only one portion of genetic diversity is connected to ecological factors, i.e. adaptation. Ecological adaptation is a significant factor for example, in range expansion of plant species. Plants with different genotypes conferring the highest levels of fitness are expected to survive and reproduce better, shifting the gene pool over time towards higher frequencies of the alleles making up the more successful genotypes (Ward et al., 2008). This can be seen through Fisher's (1930) example: when increase in fitness is allowed, genetic diversity can increase the population growth rate, but only if the population is not regulated by other factors and if it is experiencing directional selection. This is due to the simple fact that individuals with different genetic traits can interact in unpredictable ways. Despite the presence of genetic variation in ecologically important traits, relatively little is known about the range of potential ecological effects of genetic diversity for population dynamics, species interactions and ecosystem processes (Hughes et al., 2008). This has led to the rise of the field molecular ecology, which is integrating the application of molecular population genetics, phylogenetics and genomics to answer ecological questions. These different disciplines are united in an attempt to study genetic-based questions of ecology, e.g. species-area relationships, species delimitation and conservation assessment.

Information about plant genetic diversity is necessary for the development of appropriate strategies in conservation biology as well as in many other applied fields. From a basic evolutionary standpoint, genetic diversity is assumed to be crucial for the evolutionary potential of a species. Research programs that aim to investigate population structure provide evolutionary insights into the demographic patterns of diverse organisms (Milligan et al., 1994). Interest in genetic diversity has also risen in applied fields of biology such as agriculture, where the technique of multiline sowing of varietal mixes within a single field has long been known to increase crop yields (Smithson & Lenne, 1996) and to decrease damage by pests and pathogens (Cantelo & Sanford, 1984). In some crops (e.g. rice) this is applied on a large scale to maximize yield by minimizing damage by pathogens (Zhu et al., 2000). Information on genetic diversity and population structure also assists plant breeding in the selection of parents for crossing, providing a more rational basis for expanding the gene pool, and for identifying materials that harbor genes of value for plant improvement. Furthermore, knowledge of population structure of genetic resources is necessary for the development of strategies for appropriate conservation of genetic diversity. Therefore, in recent decades a primary activity in genetic resource management has been to characterize the structure of diversity within crop species. Increasing attention has also been given to dissecting and understanding diversity in relation to genes underlying important agronomic traits in a number of crops. Molecular phylogenetics and genetic diversity analysis can help to clarify the taxonomic identity and evolutionary relationships of the wild relatives of crop species. These methods can also help prevent misidentification and carefully plan effective germplasm management strategies.

Variability and genetic diversity are important factors in evolution and also in applied sciences because they determine the responses of a given organism to, for example, to environmental stress, natural selection and susceptibility to different diseases. In fact, reactions to different types of stress (biotic or abiotic) receive the most attention in genetic diversity research. In an anthropocentric world the diversity of our own species also affects the study of other organisms, because human genetic variation determines sensitivity to toxins and different drugs, leading to the development of personalized medicine (Jain 2002). This has initiated the International HapMap project (http://hapmap.ncbi.nlm.nih.gov/).

To accurately study genetic diversity and to answer questions relating to organisms, some basic terminology is required. Thus before discussing genetic diversity assessment in detail, we will define some frequently used concepts and terms:

Allele frequency: the measure of the relative proportion of all copies of a gene, i.e., the frequency of an allele at a particular genetic locus. Allele frequency is used to represent genetic diversity at different scales (e.g. population, species).

Allelic richness: the average number of alleles per locus.

Biodiversity: the degree of variation of life forms within a given ecosystem, biome or entire planet. It is **not** synonymous with **species diversity**.

Coalescent theory: a model in population genetics often referred to as backward **genetic drift**, which attempts to trace alleles of a gene shared by members of a population to a single ancestral copy. Standard coalescence assumes non-overlapping generations, random reproduction, constant population size and random mating (no population structure). However, advanced analytic models allow extension of these parameters.

Evolutionary fitness: the probability that a population or species will not become extinct. Evolutionary fitness has more relevance to evolution and population control than **genetic fitness**, which describes the reproductive success of a genotype.

Founder effect: a special case of a population bottleneck, occurring when a new population is established by a random and small number of individuals from a larger population. The founder population forming the new colony from randomly sampled alleles of the original population may poorly represent the original population.

Gene flow: also called migration, is the transfer of genes from one population to another. This may happen through mobility (e.g. transfer of pollen from one location to another). Gene flow is a very important source of genetic variation, since it enables the free "flow" of genes into a population where they did not previously exist. It can also lead to the combination of gene pools or to the reduction of genetic variation between two populations.

Gene pool: the complete set of unique alleles in a species or population.

Genetic drift: the change in the frequency of an allele in a population due to random events. It is one of the basic mechanisms of evolution affecting the genetic makeup of a population. In contrast to **natural selection** it is an entirely random process.

Genotypic richness: the number of genotypes within a population. It can be measured using molecular markers as the number of haplotypes.

Natural selection: one of the key mechanisms of evolution. Unlike **genetic drift** it is a non-random process, although not constant and varying through time. It initiates the selection of biological traits which consequently become more frequent in a population.

Population bottleneck: is an event in which a significant percentage of the genetic diversity of a population is lost. It occurs when population size is reduced for at least one generation. Through time this could lead to recovery or extinction.

Population fragmentation: a form of population segregation, most often caused by habitat fragmentation. The fragmentation of a population may have significant effects on its genetic diversity through mechanisms including **genetic drift**, reduced **gene flow** and inbreeding depression.

Population structure: heterogeneity in allele frequencies across a population caused by limited gene flow.

Species diversity: an index incorporating the number of species in a given area and their relative abundance.

Subpopulation: a portion of a total population that experiences limited gene flow from other parts of the population, such that its allele frequencies evolve independently to some degree.

2. Why does genetic diversity matter?

The genetic characteristics of a population describe its structure as formed by the interplay of genetic drift, gene flow and natural selection, and perhaps also spatial distribution. In the case of threatened species, the assessment of these characteristics can help in developing accurate strategies for maintaining and preserving species diversity. This may be crucial in the case of endangered species. Ishwaran and Erdelen (2006) correctly note that the diversity of life and living systems are necessary conditions for human development, although genetic diversity assessment projects are costly and time consuming. However, we have many reasons to make efforts to maintain biodiversity. According to Wilcove and Master (2008) biodiversity (i) has an aesthetic and moral justification; moreover (ii) wild species provide products and "services" essential to human welfare; (iii) certain species are indicators of environmental health; (iv) certain species may be keystones of functioning ecosystems; and lastly (v) studies of wild organisms have led to several scientific breakthroughs. Genetic diversity is also important for the applied fields of science and even for economic growth. Plant evolution under domestication has led to increased productivity, but at the same time, domestication has dramatically narrowed the genetic basis of crop species. Due to lack of genetic diversity in crop gene pools, efforts have been made to explore the gene pools of wild species for potential utilization in meeting the future challenges of plant breeding. Thus the main aim of breeding programs nowadays is to trace diversity and to find new traits – particularly genes conferring resistance to diseases and pests – present in wild genetic resources. This is done in order to maintain current levels of agricultural productivity. Good examples can be found in solanaceous plants, including many success stories as well as some cautionary tales. In the Solanaceae, crops such as potato, tomato and pepper have greatly benefited from the use of wild relatives in breeding programs (Albrecht et al., 2010). Virtually all of the disease resistance genes introduced into modern tomato varieties are derived from related wild species (Rick & Chetelat, 1995), demonstrating how a successful breeding program can benefit from the genetic diversity of wild species. But what would have happened without the remarkable improvement in the potato gene pool? This leads to the other side of the story. During the 1840s a lack of genetic diversity in Irish potato cultivars caused by the monocultural production of basically a single variety led to one of the biggest disasters in human history. How could low genetic diversity cause the deaths of millions of people within a very short space of time? To understand this we need to look closely at the diet of Irish people in the 19th century. Potato was introduced to Ireland in the 17th century and quickly became the staple food of the poor. It was also known as the garden crop of the gentry (Gráda, 2000). Nutritionally, potato was excellent and it was cheap to produce. Most commonly eaten with added milk or butter it formed the basis of a diet with ample protein, carbohydrates, minerals and energy. Around 1700, an average farmer would have eaten one meal with potatoes every day, while by the 1840s this had risen to three meals per day (Whelan, 1997). In an expanding agrarian economy this represented a serious potato dependency. In 1810, a new variety called "Lumper Potato" was introduced to Ireland (Whelan, 1997). This cultivar quickly became widespread, since it required minimum work and tolerated poor soil while maintaining high yields (Whelan, 1997). Lumper Potato dominated when a new disease, late blight, caused by the oomycete *Phytophthora infestans* (Fig.1), started to spread in Ireland. A plant pathogen more closely related to brown algae and diatoms than to fungi (Beakes et al., 2008), *P. infestans* can destroy susceptible crops within 10 to 14 days in favorable weather conditions and in the presence of an inoculum source (Lebecka, 2008). After a few days the whole plant collapses and infected tubers rapidly degrade into a pungent mush. Tubers that appear healthy may also turn into slime at a later stage, during storage. Even today, this pathogen ranks economically as the most important disease of potato and tomato.

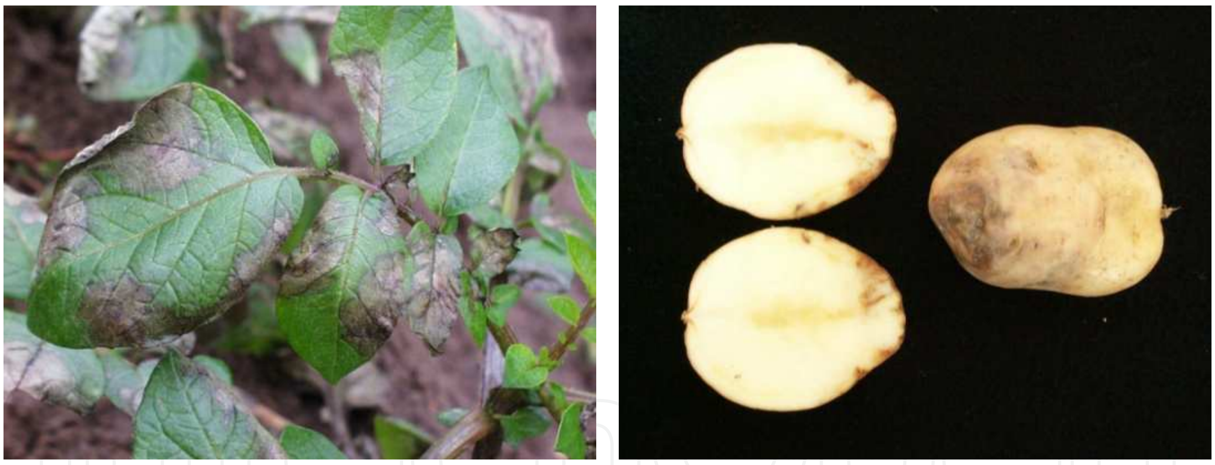


Fig. 1. The symptoms of late blight (*Phytophthora infestans*) infection in potato leaves and tubers. Photos kindly provided by Dr. Zsolt Polgár.

Prior to the arrival of *Phytophthora infestans* only two other potato diseases were known in Ireland (Bourke, 1964). The disease was first observed in September 1845 on Lumper potatoes. It spread quickly and destroyed a vast proportion of the yield in that same year. In subsequent years until 1850 the disease recurred each season, destroying yields to such an extent that people did not have sufficient seed potatoes to plant, far less to eat (Fraser, 2003). Between 1845 and 1850 approximately 2 million people lost their lives (Cousens, 1960) and 1–1.5million emigrated (Boyle & Gráda, 1986). Later in the 19th century the disease spread throughout the whole of Europe. Current studies suggest a three-step introduction process for the pathogen, including a first migration from central Mexico to South America several centuries ago, a migration from South America to the USA around 1841-42, and then a third and final migration to Europe from either South America, the USA or both in 1843-44 on ships carrying guano (Andrivon, 1996).

The epidemic had catastrophic social and economic consequences. Besides death and emigration it resulted in economic losses of 800 million US dollars (equaling 24 billion recalculated for today's market values, O'Rourke, 1994). The Great Famine also had an impact on the regional genetic structure of Ireland, resulting in a rapid decrease in population size in a short period and thus increasing the possibility of genetic drift (Relethford et al., 1997). These experiences tell us that low genetic variation can lead to disaster in ecosystems as well as in society, economy and everyday life. These consequences would not have been so severe if farmers had planted potatoes with more genetic variation.

3. The molecular toolbox: Recent technical advances in genetic diversity assessment

Understanding the forces that influence natural variation within and among populations has been a major objective of evolutionary biology for decades (Csilléry et al., 2010). It can be problematic, as natural populations may have complex demographic histories: size and range changes over time can lead to bottlenecks, fusions and expansions that leave signatures on the genetic composition of the population (Avise, 2004). Detection and analysis of genetic diversity can help us to understand the molecular basis of various biological phenomena in these natural systems. Established molecular databases providing a large volume of information on a range of different markers have potential to help in uncovering the complexity of the demographic and adaptive processes acting in natural populations. The widespread availability of different molecular markers and increased computing power has fostered the development of sophisticated methods and techniques that have begun to fulfill these expectations (Csilléry et al., 2010). Since the entire plant kingdom cannot presently be covered under sequencing projects, molecular markers and their correlation to phenotypes provide us with requisite landmarks for elucidation of genetic variation. In recent years there has been a trend away from arbitrarily amplified dominant (AAD) markers towards gene-targeted functional markers in genetic diversity assessment. Traditional genetic techniques, e.g. RAPDs (random amplified polymorphic DNA), SSRs (simple sequence repeats) and AFLPs (amplified fragment length polymorphism) are now routinely used in ecological, evolutionary, taxonomic, phylogenetic and, most importantly, in genetic diversity studies of plant species. These techniques are well established and their advantages as well as their limitations are well known. In recent years, a new class of advanced techniques has emerged, primarily derived from combination of earlier basic techniques. Genomic databases such as NCBI's GenBank have become primary sources for marker development. Based on the characterization of plant genes and gene families, new methods have been developed to analyze genetic diversity based on genomic database mining. Following this, many recent studies have suggested that polymorphism in functional regions of the genome should be exploited to achieve better estimates of genetic relationships that are relevant for conservation purposes. A wide variety of DNA arrays has been developed to meet these goals. Advanced marker techniques tend to combine adventitious features of several basic techniques, while also incorporating modifications in methodology to increase sensitivity and resolution in order to detect genetic discontinuity and distinctiveness. These advanced techniques utilize specific classes of genome elements such as (retro) transposons and cytochrome P450 genes and thus reveal genetic variation through increased genome coverage. They differ from each other with respect to important features such as genomic abundance, level of polymorphism detected, locus specify, reproducibility, technical requirements and cost. Depending on requirements, modifications have been made to these techniques, leading to second and third generations of advanced molecular markers. Major advanced techniques will be summarized here.

3.1 Targeting fingerprinting markers (TFMs)

3.1.1 Conserved DNA-Derived Polymorphism (CDDP)

The technique developed by Collard & Mackill (2009a) uses short primers to generate useful genetic markers across functional domains of well-characterized plant genes. It targets short conserved gene sequences present in the plant genome in multiple copies. Primers are specifically designed to anneal to these genes to generate polymorphic banding patterns detected on agarose gels. Theoretically any type of conserved gene region or plant gene family can be tagged using this technique. Collard & Mackill (2009a) described a set of primers that target well characterized plant genes involved in responses to abiotic and biotic stress or plant development. CDDP can easily generate functional markers (FM) related to a given plant phenotype, which is advantageous in many genetic studies. The conserved nature of priming sites in these gene regions makes the technique transferable to a wide variety of species. Since highly conserved DNA regions share the same priming site, but differ in their genomic distribution, variation can be detected as length polymorphism within these regions. The technique is based on single long primer amplifications with a high annealing temperature, which improves reproducibility. CDDP also differs in several ways from a recently developed technique called CoRAP (see 3.1.2.). Firstly, in CDDP a single primer is used in the PCR reaction, while CoRAP uses primer pairs. However, there have also been attempts to combine primers in CDDP reactions to amplify polymorphic regions representing DNA stretches between two identical or very similar conserved primer binding sites (Poczai et al., 2011). Secondly, CoRAP uses expressed sequence tags (ESTs) for specific taxa to detect polymorphism. Finally, the fragments generated by CoRAP are separated on polyacrylamide gels. This may seem to be only a minor technical difference, but the rationale is that CDDP generates fragments in the 200-1,500 bp range, while with CoRAP polymorphism manifests in small size differences in the banding patterns which require polyacrylamide gels to detect. The reproducibility of the technique has proved to be high compared to traditional RAPD (Williams et al., 1990). However, some primers problems can occur, suggesting that primer length and high annealing temperature may not ensure complete reproducibility. This indicates that scoring of banding patterns should be based on replicates and results should be treated cautiously. CDDP markers have not been widely applied in genetic diversity studies, due to their recent development and easy confusion with the other markers mentioned. They have been used to fingerprint rice varieties (Collard & Mackill, 2009a) and to assess the genetic diversity of *Solanum dulcamara* L. germplasm in Europe (Poczai et al., 2011). In these studies CDDP proved to be useful for achieving these goals.

3.1.2 Cytochrome P450 Based Analogues (PBA)

Cytochrome (Cyt) P450 mono-oxygenases are widely found in animals, plants and microorganisms (Shalk et al., 1999). In embryophytes they play important roles in oxidative detoxification and in the biosynthesis of secondary metabolites (Kessmann et al., 1990), while many P450 gene families have been found in various plant species. Sequence diversity of P450 gene-analogues in different plant species has been studied and it has been reported that such analogues can be used as genetic markers for diversity studies in plants, at both functional and genome-wide scales (Somerville & Somerville, 1999). Based on these findings a technique called Cytochrome P450 Based Analog (PBA) markers was developed by Yamanaka et al. (2003). Further, data mining of the genome of the model plant *Arabidopsis thaliana* (L.) Heynh. has resulted in the development of number of primer-sets derived from Cyt P450 genes, which could be used as universal tools for the assessment of genome-wide diversity in diverse plant species lacking relevant genetic markers (Yamanaka et al., 2003). Since Cyt P450 genes are widely distributed within the plant genome they can be universally utilized to create polymorphic fingerprints to characterize genetic diversity within and among populations of a wide verity of plant species. The genomic annotation of *Arabidopsis* revealed that out of the ~ 29,000 genes in the genome, nearly 0.9% (272 genes and 26 pseudogenes) are putative Cyt P450 genes (Riechmann et al., 2001). This indicates that these genes are very diverse, providing an opportunity to utilize them in diversity assessment. In the method developed by Yamanaka et al. (2003) universal primer pairs, designed to anneal to specific conserved exon regions of Cyt P450 genes, are arbitrarily paired. Forward and reverse primers flanking the intron regions are then used to initiate PCR amplification. Based on the random distribution of Cyt P450 genes in the genome, the resulting banding patterns will reflect polymorphism based on the variation found across the targeted (pseudo)genes. The cross-species amplification and transferability of PBAs was reported and verified in 52 different taxa from 28 families (Yamanaka et al., 2003). The generated PBA markers have diverse applications in population biology and have already been used to characterize genetic diversity in variety of species (e.g. Ahmad et al., 2009; Gilani et al., 2009; Panwar et al., 2010).

3.1.3 Intron-Targeting Polymorphism (IT)

The close proximity of introns to exons makes them especially well suited for the detection of length polymorphism in their structure that can be utilized for generating powerful new markers for genetic diversity analysis. This can be effectively achieved using the technique of intron-targeting (IT; Choi et al., 2004; Poczai et al., 2010). The basic principle of IT relies on the fact that intron sequences are generally less conserved than exons and they display polymorphism due to length and/or nucleotide variation in their alleles. Primers designed to anneal in conserved exons to amplify across introns can reveal length polymorphism in the targeted intron. Such primers can be designed for any organism using the available sequences of known genes or by exploiting expressed sequence tag (EST) records from the NCBI database. This method was first applied by Choi et al. (2004) to construct a linkage map of the legume *Medicago truncatula* Gaertn., so it represents a technological transfer from applied biology. However, it was further used to examine genetic diversity in *Solanum nigrum* L. (Poczai et al., 2010) and *S. dulcamara* (Poczai et al., 2011) populations. The results confirm that IT markers are suitable for characterizing genetic diversity. Moreover, the designed primer is transferable across species of the same genus, or in some cases possibly family. The successful transferability and cross-species amplification capacity of IT markers will depend on the conservation of exon-intron junctions and gene structures across related genomes in the genus or family. If the shared syntenies of the targeted genes as well as their sequence features are relatively conserved, the primers can be transferred easily between higher taxonomic levels. This is valuable for generating functional markers directly related to gene regions and facilitating the discovery of specific markers linked to a given phenotype. It is also possible to tag specific genes related to environmental factors that could have useful applications in molecular ecology in the broader sense. This is because IT uses primers based on allele sequences of functionally characterized genes, and thus specific banding patterns corresponding to plant phenotypes can be identified (Cernák, 2008; Cernák et al., 2008; Gizaw, 2011). However, development of such markers depends on the availability of robust genomic databases holding several target sequences for IT marker development. Functional gene characterization criteria might be limiting factors, since it is not possible to establish gene functions in a molecular ecological sense for all genes. The crucial question is whether useful allelic variation can be identified for all genes of ecological relevance in the targeted organism.

3.1.4 Start Codon Targeted (SCoT) Polymorphism

Molecular markers from the transcribed region of the genome can offer potential for various applications in plant genotyping as they reveal polymorphisms that might be directly related to gene function. A novel marker system called Start Codon Targeted Polymorphism (SCoT) was described by Collard & Mackill (2009b), based on the observation that the short conserved regions of plant genes are surrounded by the ATG translation start codon (Sawant et al., 1999). The technique uses single primers designed to anneal to the flanking regions of the ATG initiation codon on both DNA strands. The generated amplicons are possibly distributed within gene regions that contain genes on both plus and minus DNA strands. The utility of primer pairs in SCoTs was advocated by Gorji et al. (2011). SCoT markers are generally reproducible, and it is suggested that primer length and annealing temperature are not the sole factors determining reproducibility (Gorji et al., 2011). They are dominant markers, however, while a number of co-dominant markers are also generated during amplification, and thus they could be used for genetic diversity analysis (Collard & Mackill, 2009b). This has been validated through study of genetic diversity among rice varieties (Collard & Mackill, 2009b). Further studies have utilized the technique to investigate genetic relationships in a number of plant species (Xiong et al., 2011; Murthy, 2011). These studies show that SCoTs can be used solely, or in combination with other techniques, to assess genetic diversity and to obtain reliable information about population processes and structure across different plant families.

3.1.5 Sequence-Related Amplified Polymorphism (SRAP) and Targeted Region Amplified Polymorphism (TRAP)

With primers targeting short recognition sites a large number of polymorphisms can be generated. In these cases almost any primer can initiate PCR amplification. However, due to the nature of PCR itself, short primer pairs will amplify many sequences distributed throughout the genome. This is a feature of the widely used RAPD technique (Williams et al., 1990). The main disadvantages of this technique are poor consistency, low reproducibility and multiplexing of output, which limit its use. The sequence-related amplified polymorphism technique (SRAP), developed by Li & Quiros (2001), also uses arbitrary primers to generate a specific banding pattern, originally detected on polyacrylamide gels. Compared to RAPD, the primers used are much longer (17-21 nt). The forward primers are designed to contain GC-rich sequences near the 3′-end, while reverse primers contain AT-rich sequences at the 3′end. This is based on the rationale that protein coding regions tend to contain GC-rich codons, while 3′UTRs frequently consist of AT-stretches (Lin et al., 1999). More specifically, "CCGG" sequences in the core of the forward primer target exon regions. Li & Quiros (2001) found that in their experiments with randomly selected bacterial artificial colonies (BACs), 66% of the produced sequences of exon regions contained the "CCGG" motif. Lin et al. (1999) noted that approximately one-third of the part of the *Arabidopsis* genome found on chromosomes 2 and 4 represents exon regions. With the inclusion of the "CCGG" core in the primer set, sequences containing this element are preferentially amplified. However, because exons are generally conserved and might fail to produce sufficient polymorphism, the reverse primer in SRAP is designed to contain a second core with the aforementioned "AATT" motif. This second core is normally a frequent element in promoters, introns and spacers (Lin et al., 1999). Since these regions are more variable between different individuals, the intrinsic dissimilarity incorporated in the primer sets makes it feasible to generate polymorphic bands based on introns and exons (Li & Quiros, 2001). The arbitrary primers also contain further modifications, as they have 10 bases at the 5′end called filter sequences that have no specific constitution. These are followed by the core sequences (CCGG for forward and AATT for reverse), while at the 3′end three selective nucleotides are added. The PCR profile is also modified to ensure specificity and high stringency. The profile consists of two parts: the early and late cycles. During the early cycles mismatch is tolerated by the use of a lower annealing temperature (35°C), enabling the production of many amplicons. The rationale behind this is that primer annealing to the DNA template depends on the matching level of both sequences, and amplification efficiency is determined by the effectiveness of primer binding capability (Li & Quiros, 2001). The low initial annealing temperature ensures the binding of both primers to sites with partial matches in the target DNA, creating a population of amplicons that contain the priming sites. During the late cycles at higher annealing temperature (50°C), the initially generated amplicons serve as the template rather than the genomic DNA, ensuring high reliability, efficiency and reproducibility due to perfect base pairing of primers with the template. Because mismatches are allowed in the early cycles, the 5′ends of the PCR primers are usually "forced" into the PCR products. This is similar to *in vitro* mutagenesis using PCR primers (Cadwell & Joyce, 1992). The limiting factors in amplification are the 3′end sequences of the primers, which must match perfectly during the PCR cycles (Telenius et al., 1992). Therefore polymorphisms are limited by the 3′ ends of the primers, leading to exclusive amplification of alleles matching the 3′ ends, while alleles with mutations in these regions will not amplify. With changes in the selective nucleotides at the 3′ ends, new polymorphic patterns can be observed. The technique has gained popularity due to several advantages: i) a large number of polymorphic fragments are amplified in each reaction, ii) there is no *a priori* need for information about sequences, iii) primers can be applied to any species, iv) it is cost effective and easy to perform, v) reproducibility is high, and vi) PCR products can be directly sequenced using the original primers without cloning. These advantages have led to its widespread application for genetic diversity analysis in diverse organisms (e.g. Sun et al., 2006; Feng et al., 2009; Song et al., 2010).

The Targeted Region Amplified Polymorphism technique (TRAP) developed by Hu & Vick (2003) is similar to SRAP, but it incorporates the advantage of the availability of sequence information. It uses information from partial sequences of candidate genes in a fixed primer, while an arbitrary primer is used to amplify the putative candidate gene regions. The PCR cycling conditions are similar to SRAP, with the priming and amplification procedure having the same rationale. Arbitrary primers incorporate the characteristics of SRAP primers, having either AT- or GC-motifs in their sequences, with 5′ end filter sequences and 3′ end selective nucleotides. The major difference lies in the other primer pair, called the fixed primer. This primer is specifically designed from ESTs of the analyzed organism to target candidate gene regions. The generation of fixed primers limits the use of this technique to species where ESTs are known, or requires the generation of new sequence information for primer development. Despite this limitation it has been widely used for multiple purposes in different plant species (e.g. Hu et al., 2005; Alwala et al., 2006; Hu et al., 2007).

3.1.6 Conserved Region Amplification Polymorphism (CoRAP)

The CoRAP technique (Wang et al., 2009) is based on the use of two primers, a fixed and an arbitrary one, to detect polymorphism. The fixed primer is derived from directly targeted ESTs in repository sequence databases (e.g. Genbank), while the arbitrary primer contains a core sequence motif (CACGC) commonly found in plant gene introns. This core sequence ensures the utilization of conserved intron sequences in plant genotyping while the fixed (conserved) primers target coding sequences, together generating highly reproducible and reliable fingerprints. Generally, fixed primers are designed specifically for the analyzed organism, meaning that *a priori* sequence knowledge is required. The technique is similar to TRAP (see 3.1.5.) but differs in the incorporation of sequence motifs in the arbitrary primer. Fixed primers derived from ESTs will have a specific binding site on the exon of the target sequence, while the arbitrary primers will bind to most of the introns during the PCR amplifications. If the distribution of these gene elements allows successful PCR, banding patterns resulting from a specific fingerprint will be amplified. Indels in these regions will certainly generate different distributions of amplified production. The closer the genetic relationship between the two individuals, the more similar the corresponding band patterns of the amplified PCR products will be (Wang et al., 2009). The technique has been used successfully to identify economically important traits in sugarcane (Khan et al., 2011).

3.2 Mobile element based molecular markers

The following techniques can be regarded as a special group of TFMs, since they all utilize mobile elements of the plant genome to generate fingerprints. Mobile element based markers have great potential as tools for investigating aspects of molecular ecology, including population structure, conservation genetics, the genetics of speciation, phylogeography and phylogeny (Ray, 2007). Currently they remain under-utilized tools in molecular ecology. In particular, one group of mobile elements, retrotransposons, provides an excellent basis for the development of markers systems. Retrotransposons replicate by successive transcription, reverse transcription and insertion of the new cDNA copies back into the genome, very much like retroviruses (Scheifele et al., 2009). The structure and replication strategy of retrotransposons give them several advantages as markers (Kalendar et al., 1999). Firstly they are ubiquitous, present in high copy numbers as highly heterogeneous populations that are widely dispersed on chromosomes and show insertional polymorphism both within and among plant taxa (Kumar et al., 1997). Secondly, active mobile elements produce new insertions in the genome, leading to polymorphism which can be detected and used to temporally order insertion events in a lineage (Kumar & Bennetzen, 1999). Thirdly, many types of mobile elements are widely distributed in the euchromatin domains of chromosomes, making it possible to generate markers linked to a given phenotype (Kenward et al., 1999). Fourthly, an important trait of mobile element markers is that they provide homoplasy-free characters for genetic diversity assessments (Ray et al., 2006). Character states are clearly derived from a common ancestor and they are almost invariably identical by descent, but not identical by state (Ray, 2007). Fifthly, their ancestral state is known and stable, which means that the ancestral state at any amplified locus is the absence of the element, and once the element is present it will almost invariably remain there indefinitely (Shedlock & Okada, 2000). Finally, mobile element based markers are expected to be co-dominant. However, despite the fact that they are extremely useful for population genetics, all mobile element based markers have the same drawback: difficultly of data interpretation and uncertainty about the true nature of the polymorphism. Specifically, the question may arise as to whether differences in banding patterns are due to the absence or presence of retrotransposons, or are caused by some other mechanism, e.g. indels or restriction site loss. Fortunately, advances in analytical methods and a number of successful studies indicate that these drawbacks can be overcome.

3.2.1 Inter-Retrotransposon Amplified Polymorphism (IRAP) and Retrotransposon-Microsatellite Amplified Polymorphism (REMAP)

Both IRAP and REMAP were first described by Kalendar et al. (1999) for generating specific fingerprints, i.e. distinctive fragment patterns in samples. They both target a group of retrotransposons that contain direct long terminal repeats (LTRs) of variable size from 100-5,000 bp (Kumar & Bennetzen, 1999). LTRs do not code for any protein but instead contain the promoters and terminators for transcription. IRAP primers anneal to these regions and amplify DNA segments between two LTR sequences. Either one or two primers specifically designed for LTRs can be used in the same PCR, but the results will be determined by the orientation of these regions. Retrotransposons spread through the generation of new copies, rather than moving directly by excising themselves from their original position and "jumping" to a new locus. These newly generated daughter copies are then integrated back into the genome, which can happen in either orientation (5′ to 3′ or 3′ to 5′). This leads to the generation of differently oriented gene (copy) clusters which may be found in head-to-head, tail-to-tail or head-to-tail arrangements. These can be repeat type daughter sequences (head-to-tail), when the gene copy of the mobile element is directly repeated after the original copy. In other words, the two copies are in the same strand and are transcribed in the same direction. Head-to-head or bidirectional gene (copy) pairs are inverted orientations and describe a genomic locus where two adjacent genes (copies) are divergently transcribed from opposite strands of DNA, with the region between two transcription start sites (TSSs) being shared by the gene and used as a putative bidirectional promoter. In other words, the two adjacent genes are in two different strands separated by a short intergenic distance and are oriented in a divergent (-+) transcription configuration, meaning that they are transcribed away from each other. The tail-to-tail inverted orientation is the opposite of head-to-head orientation, where the two genes (copies) are located in opposite strands and are transcribed towards each other. For head-to-head and tail-to-tail arrangements, only a single primer is necessary to generate IRAP products. For head-to-tail orientation, both 5′ and 3′ LTR primers must be used to amplify the intervening genomic DNA (Kalendar et al., 1999). The other technique, REMAP, exploits the polymorphisms among regions amplified between an anchored simple sequence repeat (SSR) and an LTR sequence. To achieve this, one specifically designed LTR primer is mixed with another arbitrarily chosen primer directly containing a simple repeat [e.g. $(CA)_n$, $(GA)_n$] plus an additional and randomly chosen anchoring nucleotide in the 5′ or 3′ end [e.g. $C(CA)_n$, $(GA)_nG$]. This technique can be regarded as a modified or extended version of the inter-simple sequences repeat (ISSR) technique, since one primer in a REMAP reaction is an anchored ISSR primer combined with an IRAP primer. IRAP and REMAP have been used individually and in combination to study genetic diversity in several plant genera (e.g. Bretó et al., 2001; Branco et al., 2007; Carvalho et al., 2010), because they produce a reliable and reproducible banding profile.

3.2.2 Retrotransposon-Based Insertion Polymorphisms (RBIP)

It is also possible to tag the presence or absence of a given complete retrotransposon insertion at a specific locus and to use this information as a genetic marker. A PCR-based approach has been developed to detect such individual insertions by Flavell et al. (1998), based on a method originally developed to detect transposition of the PDR1 element in pea (*Pisum sativum* L.). The technique generates co-dominant markers, because the different allelic states caused by the presence or absence of an insertion are scored at a single locus (Kumar & Hirochika, 2001). This is achieved by PCR amplification using primers flanking the exact insertion point of the retrotransposon. Primers are designed for both the 5′ and 3′ end regions flanking the targeted retrotransposon insertion and used to amplify a product if the insertion is absent (Grzebelus, 2006). Further primers are generated from the mobile element, mostly from LTRs, and used together with a primer from the flanking region to detect the presence of the insertion. The results are then scored on agarose gels or by dot hybridization. For agarose detection, flanking primers can be combined with element based primers in a triplex-PCR for genotyping, so that detection of the occupied and unoccupied site is performed in a single tube reaction. The main advantage of this method is that it can be fully automated and applied to many samples in a high throughput system, e.g. the tagged microarray marker (TAM) system (Flavell et al., 2003). This microarray based method enables scoring of thousands of DNAs for RBIP markers on a glass slide, based on fluorescent labeling (Kalendar, 2011). The method does not suffer from the drawbacks that all gel-based markers have, particularly band sharing from one gel to another. However, the insertions of RBIPs can also be detected on agarose gels. Because of the investment needed to sequence flanking regions it is a more costly technique than those based on other mobile element based markers. In addition, it can be technically more complicated. RBIP markers have been utilized in a number of plant taxa for genetic diversity analysis, mostly in crop species where it has been possible to take advantage of sequence information on genomic structure provided by molecular breeding programs (Vershinin et al., 2003; Vitte et al., 2004; Jing et al., 2010).

3.2.3 Retrotransposon-based Sequence-Specific Amplification Polymorphism (SSAP)

The principle of this method is to convert a subset of retrotransposon based insertion sites into a corresponding set of bands on a gel, by PCR amplification of the junctions between the transposon and the host genome (Grzebelus, 2006). This technique was developed by Waugh et al. (1997) to investigate the location of the BARE-1 retrotransposon in the barley genome. The basic principle is similar to the commonly used amplified fragment length polymorphism technique (AFLP; Vos et al., 1995) – it is an extension of AFLP with generation of mobile element based molecular markers. As in AFLP, the procedure begins with the digestion of genomic DNA using a rare cutting restriction enzyme paired with a frequently cutting one (usually *Mse*I and *Pst*I, or any other restriction enzyme). The sequence of the retrotransposon is known *a priori* and the transposon family, group and type is carefully chosen before starting the reactions. The restricted DNA is then "tagged" by the ligation of short double-stranded adapters with known sequences. The ligation phase is followed by a pre-selective PCR amplification with adaptor–homologous primers. This pre-amplification is useful to reduce the genome complexity by amplifying a subset of restriction fragments when work is performed with organisms having a large genome. This step also promotes higher reproducibility of the downstream SSAP banding pattern. Pre-PCR with selective bases reduces the number of bands amplified, and this step can also be introduced to reduce the restriction fragments further if necessary (Waugh et al., 1997). The next step is selective amplification with a retrotransposon specific primer, paired with either a rare or a frequent site adaptor homologous primer. Primers are usually designed for the LTR region, but could also correspond to an internal part of the mobile element. This generates an amplified fragment for each of the transposon host junctions. The size of the fragments are determined by the distance between the transposon insertion site and the adjacent restriction cut site and differences in insertion sites between genomes are easily visible as different banding patterns (Syed & Flavell, 2006). These can by visualized on polyacrylamide gels, or a transposon specific primer can be fluorescently labeled to ensure detection. SSAP systems commonly utilize retrotransposons with LTRs from the *Ty1-copia* or *Ty3-gypsy* retrotransposon groups. The technique has been successfully used in a number of plant groups for phylogenetic and genetic diversity studies (e.g. Tam et al., 2005; Luo & Chen, 2007; Moisy et al., 2008; Ragupaty et al., 2010). It is important to note that retrotransposon families for tagging must be chosen carefully, because they may differ in the amount of information (variation) that they provide.

4. Beyond molecular marker diversity: Statistical analysis of polymorphism data

The techniques discussed in the previous section produce a number of reliable and unambiguous sets of bands, easily detectable with gel electrophoresis or by high throughput systems with fluorescently labeled primers. Polymorphism data can be scored in presence/absence matrices manually or with the aid of specific software. However, because these techniques are based on the incorporation of genomic elements in the primer sets or else target specific regions in the genome, biases affecting the evaluation process can occur. Although many recently developed targeting methods detect large numbers of polymorphisms, not many studies to date have utilized them, largely due to their unfamiliarity. In many cases the drawbacks are unknown. These mainly affect the analysis of the banding patterns produced, largely depending on the nature of the methods and whether they generate dominant or co-dominant markers. Here we give a brief description of common analytical approaches that can be applied to statistically evaluate the polymorphism data produced by recently developed targeting methods. Until such time as further studies reveal major weaknesses of targeting fingerprinting techniques, we assume that these drawbacks do not cause serious biases in the data sets produced. Our descriptions are based on those provided by the developers of the discussed techniques as detailed in the original studies cited, or on our own experiences with the application of some of the marker systems (Poczai & Hyvönen, 2011; Poczai et al., 2011; Gorji et al., 2011). There are two features of targeting fingerprinting markers (TFM) that considerably constrain how they can be statistically analyzed. Firstly, polymorphic TFM loci are generally scored for two alleles, the "band-presence" allele and the "band-absence" allele. This makes each locus less informative than a multi-allelic microsatellite locus. However, the large number of TFMs available across the genome, or the specific exploration of multiple regions and their distributions, can counteract this drawback. Secondly, the produced banding patterns for each technique are treated during the evaluation process as dominant markers, and the methods are described accordingly. Thus, it is difficult to distinguish heterozygous individuals from homozygous ones with respect to the band-presence allele, unless exact genotypes can be inferred from pedigree studies (van Haeringen et al., 2002). However, this assumption may be challenged in future and methods appropriately modified if the technical nature of each marker system is properly recognized. To date only a few studies have been reported on TFMs, which may prove to be co-dominant (Poczai et al., 2010), or with mixed banding patterns of dominant and co-dominant markers (Gorji et al., 2011). In the following sections some commonly used basic statistical approaches for measuring genetic diversity will be discussed.

4.1 Band-based approaches

The easiest way to analyze banding patterns and to measure the diversity represented by them is to focus on band presence or absence and to compare it between samples. This method is routinely used at the level of the individual, and produces distances generated from the bands rather than taking into account the diversity of the population. The disadvantage is that it treats the data as polymorphism produced by the markers rather than describing the genetic diversity of the organism. Although the results correlate with genetic diversity, the indices often have no biological meaning in a population–genetic sense.

4.1.1 Measuring polymorphism

A locus is considered polymorphic when the band is present at a frequency of between 5% and 95%. The polymorphism represented in the banding patterns can be extracted simply by observing the total number of polymorphic bands (*PB*) and then calculating the percentage of polymorphic bands (*PP*) present in any individual based on this number. The value of a particular band position can also be measured by its similarity to the optimal condition, which is 50% of genotypes containing the band. This "band informativeness" (I_b) can be represented on a scale ranging from 0 to 1 according to the formula: $I_b = 1 - (2 \times |0.5 - p|)$, where p is the portion of genotypes containing the band.

4.1.2 Shannon's information index (*I*)

A coefficient that is often called the Shannon index of phenotypic diversity is widely applied to quantify the degree of band polymorphism provided by banding patterns. It can be calculated as $I = -\sum p_i \log_2 p_i$, where it is assumed that i) the population is large, ii) random mating takes place , iii) the population is isolated from others, and iv) the effects of migration, mutation and selection are not relevant. Due to biallelic coding, some specific features also need to be taken into account. Because of these drawbacks, allele-frequency statistics are commonly supplemented with another estimation that allows for the distortion from HWE (see 4.3). These methods depend on the extraction of allelic frequencies from the data, which can be accomplished using a number of different procedures (see Lynch & Milligan, 1994; Stewart & Excoffier, 1996; Zhivotovsky, 1999; Hill & Weir, 2004; Holsinger et al., 2002). The accurate estimation of frequencies essentially determines the accuracy of the different indices calculated for further quantification of genetic diversity.

4.1.3 Similarity coefficients

Band-based approaches usually utilize similarity or dissimilarity (the inverse of the previous one) coefficients. Many coefficients have been developed, but they all have common properties that they incorporate the pattern of matches and mismatches among their respective band presences and absences. We will just only discuss the three most widely applied indices in this section. Abbreviations will be used for both three as the following: a = the number of bands (1s) shared by both individuals; b = number of positions where individual i has a band, but j does not; c = number of positions where individual j has a band, but i does not; and n = the total number of bands (0s and 1s). The following measures are commonly used for comparison between individuals: i) **Jaccard similarity coefficient** or Jaccard index $J = a/(a + b + c)$ (Jaccard, 1908); ii) **Simple matching coefficient** or index $SM = (n - b - c)/n$ (Sokal & Michener, 1958); iii) **Sørensen-Dice index** or Nei & Li index $SD = 2a/2a + b + c$ (Dice, 1945; Sørensen, 1948; Nei & Li, 1979). The Jaccard coefficient (*J*) only takes into account the bands present in at least one of the two individuals. It is therefore unaffected by homoplasic absent bands (where the absence of the same band is due to different mutations). The simple-matching index (*SM*) maximizes the amount of information provided by the banding patterns considering all scored loci. Doubling the band absence and presence receive the same importance, which is not the case for frequent band absence homoplasy. The Neil & Li index (*SD*) doubles the weight for bands present in both individuals, thus giving more attention to similarity than dissimilarity.

4.2 Allele frequency based approaches

The methods discussed here measure variability by describing changes in allele frequencies for a particular trait over time. Using the presence/absence dominant data allele frequencies are estimated at each locus. This strategy is more population oriented, than band-based approaches; however, data is treated as dominant for each locus and thus allele frequencies are accessible only with preliminary assumptions made additionally. Moreover, they expect that the population is in Hardy-Weinberg equilibrium (HWE), it is assumed that the i) population is large, ii) random mating takes place , iii) the population is isolated from others, and iv) the effects of migration, mutation and selection are neglected. Due to biallelic coding some specific features should also be taken into account. Because of these drawbacks allele-frequency statistics are commonly supplemented with other estimation which allows the distortion from HWE (see 4.3). These methods depend on the extraction of allelic frequencies from the data. This can be done according to many procedures (see Lynch & Milligan, 1994; Stewart & Excoffier, 1996; Zhivotovsky, 1999; Hill & Weir, 2004; Holsinger et al., 2002). The accurate estimates of frequencies essentially influence the results of different indices calculated for further measurements of genetic diversity.

4.2.1 Allelic diversity (*A*)

One of the easiest ways to measure genetic diversity is to quantify the number of alleles present. Allelic diversity (*A*) is the average number of alleles per locus and is used to describe genetic diversity. When there is more than one locus, *A* is calculated as the number of alleles averaged across all loci: $A = n_i / n_l$ where n_i is the total number of alleles over all loci and n_l is the number of loci. The observed number of alleles (n_a) can also be given by counting the number of alleles per locus and taking the average of these counts. These values then can be compared with the effective number of alleles (n_e), which is the number of alleles needed to provide the same heterozygosity if all alleles were equally frequent. It is less sensitive to sample size and rare alleles, and calculated as $n_e = 1 / \sum p_i^2$ pi2 ability, so it is expected to provide information about the dispersal ability of the organism and the degree of isolation among populations.

4.2.2 Effective population size (N_e)

The concept of effective population size (N_e) plays an important role in conservation management. It provides a measure of the rate of genetic drift, the rate of genetic diversity loss and increase of inbreeding within a population. The effective population size by definition is the number of individuals that would give rise to the calculated inbreeding coefficient, loss of heterozygosity, or variance in allele frequency frequency if they behaved in the manner of an idealized population. (Frankham et al., 2002). In other words, it is the size of an idealized population that would lose genetic diversity (or become inbred) at the same rate as the actual population. It is a value which depicts the general core number individuals in a population who contribute offspring to the next generation. The effective population size is usually smaller than the absolute population size. The major factors effecting Ne are i) fluctuating population size, ii) breeding sex ratio, iii) overlap of generations and iv) spatial dispersion. There are many formulas to calculate effective population size that will not be discussed here due to their specific nature. In should be noted that the effective size of a population is an idealized number, since many calculations depend on the genetic parameters used and on the reference generation. Therefore, a single population may have many different effective sizes which are biologically meaningful but distinct from each other.

4.2.3 Heterozygosity (*H*)

Heterozygosity can be regarded as the average portion of loci with two different alleles at a single locus within an individual. It is commonly extended to the whole or a sub-portion of an entire population and differentiated into observed and expected heterozygosity. Expected heterozygosity (H_E) or Nei's gene diversity (*D*) is the expected probability that an individual will be heterozygous at a given locus, in multi-locus systems over the assayed loci. In other words, it is the estimated fraction of all individuals that would be heterozygous for any randomly chosen locus. It is often calculated based on the square root of the frequency of the null (recessive) allele as follows: $H_E = 1 - \sum_i^n p_i^2$, where p_i is the frequency of the i^{th} allele. Observed heterozygosity (H_O) is the portion of genes that are heterozygous in a population. It is calculated for each locus as the total number of heterozygotes divided by sample size. Typically values for H_E and H_O range from 0 (no heterozygosity) to nearly 1 (a large number of equally frequent alleles). Expected heterozygosity is usually calculated when describing genetic diversity, as it is less sensitive to sample size than observed heterozygosity. If H_O and H_E are similar (they do not differ significantly) mating in the populations is approximately random. If H_O is less then H_E, the population is inbreeding; if H_O exceeds H_E, the population has a mating system avoiding inbreeding.

4.2.4 *F*-statistics

In population genetics the most widely applied measurements besides heterozygosity are *F*-statistics, or fixation indices. These were originally designed to measure the amount of allelic fixation by genetic drift. *F*-statistics are used to describe the structure of the population at different levels. The measure of Wright's index is based on the comparison of frequencies of identical alleles within and between groups. It can describe the inbreeding (coefficient) of an individual relative to the total population (F_{IT}) or the inbreeding (coefficient) of an individual relative to the subpopulation (F_{IS}), and can also express the "fixation" (index) resulting from comparing subpopulations to the total population (F_{ST}). In other words, different *F*-coefficients explain the correlation of genes within individuals over all populations (F_{IT}), the correlation of genes of within individuals within populations (F_{IS}), and the correlation of genes of different individuals in the same population (F_{ST}). For a given species this can be interpreted as the overall inbreeding (F_{IT}), the inbreeding within taxa (F_{IS}) and the coefficient of co-ancestry (F_{ST}), which provides an estimate of interspecific genetic differentiation. The *F*-statistics are related to heterozygosity and genetic drift. Since inbreeding increases the frequency of homozygotes, as a consequence, it decreases the frequency of heterozygotes. In other words, population substructure results in a reduction of heterozygosity. The general result of genetic drift in small populations is a deficiency of heterozygotes and an excess of homozygotes. Moreover, mating between relatives (inbreeding) and genetic drift are related phenomena, which allow us to determine fixation indices when expected and observed heterozygosity are known (see 4.2.4). Thus the three indexes can be calculated as: $F_{IT} = 1 - (H_I/H_T)$; $F_{IS} = 1 - (H_I/H_S)$ and $F_{ST} = 1 - (H_S/H_T)$ where H_I is the average observed heterozygosity within each population, H_S is the average expected heterozygosity of subpopulations assuming random mating within each population and H_T is the expected heterozygosity of the total population assuming random mating within subpopulations and no divergence of allele frequencies among subpopulations. Fixation indices are interrelated as follows: $1 - F_{IT} = (1 - F_{ST})\ (1 - F_{IS})$ and thus $F_{ST} = (F_{IT} - F_{IS})/\ (1 - F_{IS})$. They vary from 0 (no isolation) to 1 (complete isolation), but in the case of a bias towards polymorphic loci, lower values (e.g. 0.5) may indicate complete isolation between populations.

4.2.5 Gene flow (*Nm*)

The parameter of gene flow (*Nm*) is the product of the effective size of individual populations (*N*) and the rate of migration among them (*m*). However, it is difficult to measure the rate of gene flow directly by tracking individuals and also to mathematically infer its value. Many estimates have been developed concerning particular types of gene flow that will not be discussed here. However, commonly used approach is to make an estimation based on the degree of genetic differentiation among populations by calculating *F*-statistics (detailed in 4.2.5). F_{ST} is related to population size and migration rate, based on a standard island model, which is not applicable to every natural population, but provides an easy way to calculate gene flow as the following: $Nm = (1/F_{ST} - 1)$. Gene flow among fragmented populations is related to dispersal ability, so it is expected to provide information about the dispersal ability of the organism and the degree of isolation among populations.

4.3 Bayesian estimates of genetic diversity

Classical population genetic methods are based on the calculation of various values as discussed in the previous sections. The principles of Bayesian inference rely on the perspective that both model parameters and data are random variables with a joint probability distribution. In general terms, it is a statistical inference method where the evidence is used to revise the ambiguity of parameters and predictions in probability models. In these estimates, the initial belief is termed the *prior*, while the modified belief is the *posterior*. The prior summarizes the initial information about the value of a parameter or hypothesis, before the data is analyzed to produce a probability distribution. The aim of Bayesian calculations is to calculate the *posterior* distribution of the parameters given the data. These functions are useful since no natural population can possibly meet all of the requirements of HWE. However in most cases, allele frequency based methods provide *a priori* knowledge of what dynamics may influence the population. Later, these calculations can be refined making no assumption about HWE and estimating the effective size of populations, levels of heterozygosity, and inbreeding using Bayesian methods. Given that calculation of allele frequencies based on dominant markers requires the assumption of HWE, a more direct estimate of the relative magnitude of population divergence can be obtained using the Bayesian approach. These methods do not assume that genotypes are in Hardy-Weinberg proportion within the population, while taking full advantage of the information provided by molecular markers (Holsinger & Wallace, 2004). Bayesian methods are favored for estimating parameters such as mutation rate (μ) and effective population size (N_e) in demographic models. Bayesian assignment methods are also used to study population differences and to infer parameters in genealogical models concerning coalescent theory, and to detect selection. Other areas of Bayesian computation include inferring changes in population size, analyzing population structure, and even identifying haplotypes through population samples. In many cases, Bayesian methods can address the question of interest more directly than classical approaches. Consequently they are an essential part of genetic diversity estimation and population genetics. Bayesian inference allows the incorporation of complicated models that can be studied, and biologically relevant parameters that can be estimated with the incorporation of prior information. In addition to allele frequency based metrics, the posterior distribution of almost every classically used parameter can be estimated using Bayesian models. In this regard they have Bayesian analogues, e.g. Bayesian fixation index, θ^B (equivalent of Wright's F_{ST}), Bayesian inbreeding, f (the estimate of F_{IS}), etc. A detailed discussion of available Bayesian methods in genetic diversity assessment is beyond the scope of this chapter due to their complexity, but they have been reviewed for example by Shoemaker et al. (1999) and Beaumont & Rannala (2004).

5. Phylogenetic methods and phylogenetic diversity

Besides genetic diversity estimates it is often important in molecular ecological studies to represent the genetic structure of the population(s) examined. Such information is extremely valuable because it can address many important issues such as estimation of migration, identification of conservation units, or even resolution of biogeographical patterns (Mantel et al., 2005). When geographical locations are available for populations and sampling allows the detection of spatial structure, then tree-based methods should be considered. For many taxa the demarcation of populations is problematic when there is no *a priori* knowledge of population structure (Hollingsworth & Ennos, 2004). In these cases and in general, tree-based methods allow graphical representation of the relatedness of individuals and exploration of information on the (spatial) structure of populations. After identifying population structure the results can be evaluated in the light of biological phenomena, e.g. ancient vicariance, recent dispersal or other biogeographical patterns, genetic differentiation or even migration-drift equilibrium. These methods might be used to describe the history of a population over a long or a short time period depending on the design of the study. Various population processes can be traced and identified, such as range expansion or contraction (Tang et al., 2008), responses to environmental or climatic conditions including climate change (Schmidt & Jensen, 2000; Jump et al., 2006), sources of colonization and origins of founder populations (Gladieux et al., 2011). The simultaneous evaluation of the results of tree-based methods coupled with the distribution of individual markers within and among identified groups provides some information on the distribution of genetic diversity (Bonin et al., 2007). Genetic diversity is assumed to be higher in refugial areas or in meeting zones, while a large number of separate fragments might indicate ancient divergence (Schonswetter & Tribsch, 2005). More recent advances allow the addressing of questions such as “how many populations exist in the study area?” or “is this individual a migrant from another population?” (see Mantel et al., 2005). In the following section some basic tree-based methods will be briefly discussed in relation to the above mentioned goals.

5.1 Phylogenetic analyses

All life on Earth is connected through shared history. Reconstruction of this history, phylogeny, is one of the most difficult challenges in contemporary biology. However it is well worth the effort, due to the considerable benefits that robust phylogenetic hypotheses can provide for diverse fields of biology, both basic and applied. Because even distantly related organisms are surprisingly similar genetically and share the same regulatory genes, it would be counterproductive not to take advantage of the full range of variation produced by this great experiment conducted over billions of years. Robust phylogenetic hypotheses are essential tools for sampling this variation efficiently. In recent decades, phylogenetics has been revolutionized by (i) a re-evaluation of its fundamental theoretical premises, beginning with the seminal publications of Hennig (1950, 1966), (ii) the routine use of molecular sequence level data, (iii) the development of efficient new algorithms, and (iv) the availability of parallel computing resources allowing the use of large datasets (and the study of distant history). The latter two developments especially have enabled studies to be undertaken (e.g. Goloboff et al., 2009) that would not have been feasible only a few years ago. We are now able to provide robust phylogenetic hypotheses that can be used as reliable starting points for further studies that benefit practically all fields within the life sciences.

Phylogenetic analyses are now routinely performed and are an integral part of systematics. The methods and programs used are diverse, and it is beyond the scope of this treatment to list them comprehensively or to discuss the pros and cons of various fundamentally different methodologies. The primary division in contemporary approaches is in the amount of background knowledge they incorporate. Classical methods are non-statistical, and use parsimony as an optimality criterion. This approach is largely based on the numerous original contributions by Farris et al. (e.g. Kluge & Farris, 1969; Farris, 1970 and Farris et al., 1970), as summarized by Farris (1983). What is assumed *a priori* is only descent with modification. A fundamentally different approach, following Felsenstein (1973) and based on maximum likelihood, incorporates explicit statistical assumptions in the analyses. These model-based methods also include those based on Bayesian argumentation, pioneered by Edwards (1970) and Farris (1973) and lately further developed by e.g. Li et al. (1996) and Mau & Newton (1997). However, both of these approaches are set apart from purely algorithmic distance methods (e.g. Michener & Sokal, 1957) by their discrete nature, i.e. they operate directly on characters rather than on distances calculated from them (Page & Holmes, 1998).

The search for optimal trees, irrespective of the optimality criterion used, is computationally demanding for real data sets and discrete characters (the problem is NP-hard), and thus only heuristic searches are feasible. The advent of molecular characters has exposed another, sometimes equally difficult task in phylogenetic analyses, namely alignment. As succinctly noted by Higgs & Attwood (2005) „...the problems of multiple alignment and tree construction are inextricably linked." The most thorough treatment of this and of the related problems has been provided by Wheeler (e.g. 1996, 1999, 2003a, b).

The methods listed above are further united in that the trees produced are connected graphs without cycles. While macroevolution can be efficiently described using this type of divergent tree, it would be unrealistic to ignore the merging of lineages, i.e. hybridization, as one source of contemporary diversity. Although the methods described above can only indirect hint at hybridization, recently methods that also attempt to consider networks as phylogenetic hypotheses have been developed (e.g. Bryant & Moulton, 2004; Huson & Bryant, 2006). Their development is, however, still in its infancy. It should be noted that finding an optimal network is computationally an extremely demanding task – even for a pair of trees it has been shown to be NP-hard (Bordewich & Semple 2005, 2006). Thus the currently available methods are purely algorithmic, i.e. they do not even attempt to find the optimal solution.

5.2 Phylogenetic diversity

Traditionally there has been a conceptual gulf between research into genetic diversity and research into phylogeny. This partially reflects the differing concerns and working methods of biologists interested in genetic structure at and below the species level, where relationships are often reticulate, and those looking at variation between monophyletic groups, where hierarchical (tree-like) relationships predominate. Because conservationists and others interested in quantifying diversity have often come from ecological backgrounds, the phylogenetic structure of genetic variation in populations, taxa and ecosystems has frequently been neglected. A knowledge of this structure is essential for understanding what the genetic diversity within a given species (or, more ideally, a least-inclusive taxonomic unit, LITU; Pleijel & Rouse, 2000) is likely to represent as part of the total diversity within any more inclusive group. Thus it is also essential for any informed conservation strategy where the units of interest are discrete populations, species or higher-level taxa, or for optimizing any search strategy for genetic resources within a group of LITUs or higher-level taxa.

Faith (1992) showed how the sum of the branch lengths on a minimum-spanning path between any set of terminals on a cladogram could be used to calculate a metric he termed Phylogenetic Diversity (PD) for that set. As a simple example, take an isolated species with a very long branch that is sister to a clade of 20 species, also with a long branch but with the branches separating the species all very short, perhaps due to a recent, rapid adaptive radiation event. The PD of the first species would simply be its branch length, while the total PD of its sister clade would be the sum of the branch lengths connecting all terminals within that clade (including the single stem branch). While the much more speciose clade of 20 species might have a somewhat higher PD than its isolated sister species, it would only be necessary to sample one or two members of the clade to account for a very high proportion of that total PD, whereas neglecting to sample the single isolated species would miss a large amount of PD. Assuming that variation in the data used to construct the cladogram adequately reflects total genetic variation among the species, then PD effectively measures the proportion of diversity represented by any subset of terminals relative to any other by precisely accounting for shared diversity (represented as synapomorphy on the cladogram). In other words, maximizing PD when selecting a subset of genetic diversity to conserve or to explore for genetic resources is a way of maximizing the amount of hierarchical variation being considered and avoiding repeated sampling of similar or identical diversity derived from shared ancestry.

From a genetic conservation perspective, PD is important not only because species in isolated, monotypic lineages usually represent much more unique genetic diversity than ones in younger, more speciose groups, but also because the very nature of these species means that they may actually be disproportionately threatened by extinction (Purvis et al., 2000). Isaac et al. (2007) devised a measure they termed Evolutionary Distinctiveness (ED), which represents the relative contribution of an individual species to a clade's PD (with the standard PD measure, shared PD is included in the PD of any subset of terminals) . They further showed how extinction risk (derived from Red List categories; IUCN, 2011) could be combined with ED to identify species that are both Evolutionarily Distinct and Globally Endangered ("EDGE species"). Significantly, they also showed that ED scores (and by extension EDGE scores) are robust to clade size, missing species and poor phylogenetic resolution, issues that have been seen as obstacles to the widespread application of PD in practical biodiversity assessment in the past.

Often it is desirable to quantify genetic diversity on a geographic as well as a phylogenetic basis. Rosauer et al. (2009) provide a method of combining PD with weighted endemism measures (Williams et al., 1994; Crisp et al., 2001) to identify geographic areas in which high levels of PD are restricted. Phylogenetic Endemism (PE) effectively corrects for phylogenetic uniqueness when identifying high value areas based on levels of endemism. Thus when calculated using appropriate genetic data it will identify areas of high (phylo)genetic diversity where that diversity is less likely to occur elsewhere.

Just as standard genetic diversity is a measure of variation in a population or species, phylogenetic diversity measures variation between and amongst clades. It is the shared ancestry of clades (and the corresponding differentially shared genetic diversity) that makes a phylogenetic context essential for useful comparisons of genetic diversity between monophyletic taxa.

6. Acknowledgements

The authors would like to thank all colleagues who contributed to this study. This work was supported by the Hungarian Eötvös Research Grant (MÖB/90-1/2011) provided for Péter Poczai.

This study was supported by the TÁMOP-4.2.2/B-10/1-2010-0025 program.

7. References

Ahmad, S.M.; Hoot, S.B.; Qazi, P.H. & Verma, V. (2009) Phylogenetic patterns and genetic diversity of Indian *Tinospora* species based on chloroplast sequence data and cytochrome P450 polymorphisms. *Plant Systematics and Evolution* 281:87–96

Albrecht, E.; Escobar, M. & Chetelat, R.T. (2010) Genetic diversity and population structure in the tomato-like nightshades *Solanum lycopersicoides* and *S. sitens*. *Annals of Botany* 105: 535–554

Alwala, S.; Suman, A.; Arro, J.A.; Vermis, J.C. & Kimbeng, C.A. (2006) Target region amplification polymorphism (TRAP) for assessing genetic diversity in sugarcane germplasm collections. *Crop Science* 46:448–455

Andrewartha, H.G. & Birch, L.C. (1954) *The distribution and abundance of animals*. Unniversity of Chicago Press, Chicago

Andrivon, D. (1996) The origin of *Phytophthora infestans* populations present in Europe in the 1840s: a critical review of historical and scientific evidence. *Plant Pathology* 45:1027–1035

Avise, J.C. (2004) *Molecular markers, natural history and evolution*. Sinauer Associates

Beakes, G. W. & Sekimoto, S. (2008) The Evolutionary Phylogeny of Oomycetes—Insights Gained from Studies of Holocarpic Parasites of Algae and Invertebrates, In: *Oomycete Genetics and Genomics: Diversity, Interactions, and Research Tools,* Lamour, K. & Kamoun, S.; pp.1–25, John Wiley & Sons, Inc., Hoboken, NJ, USA

Beumont, M.A. & Rannala, B. (2004) The Bayesian revolution in genetics. *Nature Reviews* 5:251-261

Bonin, A.; Ehrich, D. & Mantel, S. (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology* 16:3737–3758

Bordewich, M. & Semple, C. (2005) On the computational complexity of the rooted subtree prine and regraft distance. *Annals of Combinatorics* 8:409–423

Bordewich, M. & Semple, C. (2006) Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics* 155:914–928

Bourke, P.M.A. (1964) Emergence of potato blight, 1843-1846. *Nature* 203:805–808

Boyle, P.P. & Gráda, C.Ó. (1986) Fertility trends, excess mortality, and the Great Irish Famine. *Demography* 23:543–562

Branco, J.S.C.; Vieira, E.A.; Malone, G.; Kopp, M.M.; Malone, E.; Bernardes, A.; Mistura, C.C.; Carvalho, F.I.F. & Oliveira, C.A. (2007) IRAP and REMAP assessments of genetic similarity in rice. *Journal of Applied Genetics* 48:107–113

Bretó, M.P.; Ruiz, C.; Pina, J.A. & Asíns, M.J. (2001) The diversification of *Citrus clementina* Hort. ex Tan., a vegetatively propagated crop species. *Molecular Phylogenetics and Evolution* 21:285–293

Bryant, D. & Moulton, V. (2004) NeighborNet: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21:255–265

Cadwell, R.C. & Joyce, G.F. (1992) Randomization of genes by PCR mutagenesis. *Genome Research* 2:28–33

Cantelo, W.W. & Sanford, L.L. (1984) Insect population response to mixed and uniform plantings of resistant and susceptible plant material. *Environmental Entomology* 13:1443–1445

Carvalho, A.; Guedes-Pinto, H.; Martins-Lopes, P. & Lima-Birto, J. (2010) Genetic variability of Old Portuguese bread wheat cultivars assayed by IRAP and REMAP markers. *Annals of Applied Biology* 156:337–345.

Cernák, I. (2008) *Development of molecular markers to the potato virus Y (PVY) extreme resistance gene (Ry_{sto}) originating from the wild potato species Solanum stoloniferum*. PhD thesis, University of Pannonia, Keszthely, Hungary

Cernák, I.; Decsi, K.; Nagy, S.; Wolf, I.; Polgár, Z.; Gulyás, G.; Hirata, Y. & Taller, J. (2008) Development of a locus-specific marker and localization of the Ry_{sto} gene based on linkage to a catalase gene on chromosome XII in the tetraploid potato genome. *Breeding Science* 58: 309–314

Choi, H.K.; Kim, D.; Uhm, T.; Limpens, E.; Lim, H.; Mun, J-H.; Kalo, P. et al. (2004) A sequence-based genetic map of *Medicago truncalata* and comparison of marker colinearity with *M. sativa*. *Genetics* 166:1463–1502

Collard, B.C.Y. & Mackill, D.J. (2009a) Conserved DNA-derived polymorphism (CDDP): a simple and novel method for generating DNA markers in plants. *Plant Molecular Biology Reporter* 27:558–562

Collard, B.C.Y. & Mackill, D.J. (2009b) Start Codon Targeted (SCOT) polymorphism: A simple novel DNA marker technique for generating gene-targeted markers in plants. *Plant Molecular Biology* 27:86–93

Cousens, S.H. (1960) Regional death rates in Ireland during the Great Famine, from 1846 to 1851. *Population Studies* 14:55–74

Crisp, M.; Laffan, S; Linder, H & Monro, A. (2001) Endemism in the Australian flora. *Journal of Biogeography* 28:183–198

Csilléry, K.; Blum, M.G.B.; Gaggiotti, O.E. & François, O. (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution* 25:410–418

Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology* 26: 297–302.

Edwards, A.W.F. (1970) Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society Series B* 32: 155–174

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61:1-10

Farris, J.S. (1970) A method for computing Wagner trees. *Systematic Zoology* 19: 83–92

Farris, J.S. (1973) A probability model for inferring evolutionary trees. *Systematic Zoology* 22: 250–256

Farris, J.S. (1983) *The logical basis of phylogenetic analysis*. In: Platnich, N.I. & Funk, V.A. (Eds.) *Advances in Cladistics Vol. 2*. pp. 7-36, Columbia University Press, New York.

Farris, J.S.; Kluge, A. G. & Eckardt, M.J. (1970) A numerical approach to phylogenetic systematics. *Systematic Zoology* 19: 172–189

Felsenstein, J. (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on descrete characters. *Systematic Zoology* 22: 240–249

Feng, N.; Sue, Q.; Guo, Q.; Zhao, R. & Guo, M. (2009) Genetic diversity and population structure of *Celosia argentea* and related species revealed by SRAP. *Biochemical Genetics* 47:521–532

Fisher, R.A. (1930) *The genetical theory of natural selection*. Oxford University Press, Oxford

Flavell, A.J.; Bolshakov, V.N.; Booth, A.; Jing, R.; Russell, J.; Ellis, T.H.N. & Isaac, P. (2003) A microarray-based high throughput molecular marker genotyping method: the tagged microarray marker (TAM) approach. *Nucleic Acids Research* 31:e115

Flavell, A.J.; Knox, M.R.; Pearce, S.R. & Ellis, T.H.N. (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *The Plant Journal* 16:643–650

Frankham, R.; Briscoe, D.A. & Ballou, J.D. (2002) *Introduction to conservation genetics*. Cambridge University Press, UK

Fraser, E.D.G. (2003) Social vulnerability and ecological fragility: building bridges between social and natural sciences using the Irish Potato Famine as a case study. *Conservation Ecology* 7:9

Gilani, S.A.; Kikuchi, A. & Watanabe, K.N. (2009) Genetic variation within and among fragmented populations of endangered medical plant, *Withania coagulans* (Solanaceae) from Pakistan and its implications for conservation. *African Journal of Biotechnology* 8:2948–2958

Gizaw, S.A. (2011) *Molecular marker development and genetic mapping in potato (Solanum tuberosum) genome of their use in breeding*. MSc thesis, University of Pannonia, Keszthely, Hungary

Gladieux, P.; Giraud, T.; Kiss, L.; Genton, B.J.; Jonot, O. & Shykoff, J.A. (2011) Distinct invasion sources of common ragweed (*Ambrosia artemisiifolia*) in Eastern and Western Europe. *Biological Invasions* 13:933–944

Goloboff, P.A.; Catalano, S.A.; Mirande, J.M.; Szumik, C.A.; Arias, J.S.; Källersjö, M. & Farris, J.S. (2009) Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics* 25:211–230

Gorji, A.M.; Poczai, P.; Polgar, Z. & Taller, J. (2011) Efficiency of arbitrarily amplified dominant merkers (SCoT, ISSR and RAPD) for diagnostic fingerprinting in tetraploid potato. *American Journal of Potato Research* 88:226–237

Gráda, C.Ó. (2000) *The Great Irish Famine*. Cambridge University Press, Cambridge, UK

Grzebelus, D. (2006) Transposon insertion polymorphism as a new source of molecular markers. *Journal of Fruit and Ornamental Plant Research* 14:21–29

van Haeringen, W.A.; Den Bieman, M.G.; Lankhorst, A.E.; van Lith, H.A. & van Zutphen, L.F. (2002) Application of AFLP markers for QTL mapping in the rabbit. *Genome* 45:924–921

Hanski, I. (1998) Metapopulation dynamics. *Nature* 396:41–49

Hennig, W. (1950) *Grundzüge einer Theorie der Phylogenetischen Systematik*. Deutscher Zentralverlag, Berlin.

Hennig,W. (1966) *Phylogenetic Systematics*. University of Illinois Press, Urbana.

Higgs, P.G. & Attwood, T.K. (2005) *Bioinformatics and molecular evolution*. Blackwell, London.

Hill, W.G. & Weir BS (2004) Moment estimation of population diversity and genetic distance from data on recessive markers. *Molecular Ecology* 13: 895–908

Holsinger, K.E. & Wallace, L.E. (2004) Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (Orchidaceae). *Molecular Ecology* 13:887–894

Holsinger, K.E.; Lewis, P.O. & Dey, D.K. (2002) A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology* 11:1157–1164

Hu, J. & Vick, B.A. (2003) Target region amplification polymorphism: a novel marker technique for plant genotyping. *Plant Molecular Biology Reporter* 21:289–294

Hu, J.; Mou, B. & Vick, B.A. (2007) Genetic diversity of 38 spinach (*Spinacia oleracea* L.) germplasm accessions and 10 commercial hybrids assessed by TRAP markers. *Genetic Resources and Crop Evolution* 54:1667–1674

Hu, J.; Ochoa, O.E.; Truco, M.J. & Vick, B.A. (2005) Application of the TRAP technique to lettuce (*Lactuca sativa* L.) genotyping. *Euphytica* 144:225–235

Hughes, A.R.; Inouye, B.D.; Johnson, M.T.J; Underwood, N. & Vellend, M. (2008) Ecological consequences of genetic diversity. *Ecology Letters* 11:609–623

Huson, D.H. & Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23:254–267

Isaac, N.J.B.; Turvey, S.T.; Collen, B.; Waterman, C. & Baillie, J.E.M. (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE* 2: e296.

IUCN (2011) The IUCN Red List of Threatened Species. http://www.iucnredlist.org.

Jaccard, P. (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise Des Sciences Naturelles* 44:223–270

Jain, K.K. (2002) Personalized medicine. *Current Opinion in Molecular Therapeutics* 4:548–558

Jing, R.; Vershinin, A.; Grezbyta, J.; Shaw, P.; Sm*ýkal, P.; Marshall, D.; Ambrose, M.J. et al. (2010) The genetic diversity and evolution of field pea (Pisum) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evolutionary Biology 10:44*

Jump, A.S.; Hunt, J.M.; Martínez-Izquirdo, J.A. & Pe*ñuelas, J. (2006) Natural selection and climate change: temperature-linked spatial and temporal trends in gene frequency in Fagus sylvatica. Molecular Ecology 15:3469–3480*

Kalendar, R. (2011) The use of retrotransposon-based molecular markers to analyze genetic diversity. *Field and Vegetable Crops Research* 48:261–274

Kenward, K.D.; Bai, D.; Ban, M.R. & Brandle, J.E. (1999) Isolation and characterization of *Tnd*-1, a Retrotransposon marker linked to black root resistance in tobacco. *Theoretic and Applied Genetics* 98:387–395

Kessmann, H.; Choudhary, A.D. & Dixon, A. (1990) Stress responses in alfalfa (*Medicago sativa* L.). III. Induction of medicarpin and cytochrome P450 enzyme activities in elicitor-treated cell suspension cultures and protoplasts. *Plant Cell Reports* 9:38–41

Khan, M.S.; Yadav, S.; Srivastava, S.; Swapna, M.; Chandra, A. & Singh, R.K. (2011) Development and utilization of conserved-intron scanning marker in sugarcane. *Australian Journal of Botany* 59:38–45

Kluge, A.G. & Farris, J.S. (1969) Quantitative phyletics and the evolution of Anurans. *Systematic Zoology* 18: 1–32

Kumar, A. & Bennetzen, J.L. (1999) Plant retrotransposons. *Annual Review of Genetics* 33:479–532

Kumar, A. & Hirochika, H. (2001) Applications of retrotransposons as genetic tools in plant biology. *Trends in Plant Science* 6:127–134

Kumar, A.; Pearce, S.R.; McLean, K.; Harrison, G.; Helsop-Harrison, J.S.; Waugh, R. & Flavell, A.J. (1997) The *Ty*1-*copia* group of retrotransposons in plants: genomic organization, evolution, and use as molecular markers. *Genetica* 100:205–217

Laikre, L.; Allendorf, F.W.; Aroner, L.C.; Baker, C.S.; Gregovich, D.P.; Hansen, M.M.; Jackson, J.A. et al. (2009) Neglect of genetic diversity in implementation of the conservation on biological diversity. *Conservation Biology* 24:86–88

Lebecka, R. (2008) Host-pathogen interaction between *Phytophthora infestans* and *Solanum nigrum*, *S. villosum* and *S. scabrum*. *European Journal of Plant Pathology* 120: 233–240

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

Li, S.; Pearl, D.K. & Doss, H. (2000) Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of American Statistical Association* 95:493–508

Lin, X.; Kaul, S.; Rounsley, S.; Shea, T.P.; Benito, M-I.; Town, C.D.; Fujii, C.Y. et al. (1999) Sequencing and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402:761–768

Lou, Q. & Chen, J. (2007) *Ty*1-*copia* retrotransposon-based SSAP marker development and its potential in the genetic study of cucurbits. *Genome* 50:802–810

Lynch, M. & Milligan, B.G. (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* 3:91–99

Mantel, S.; Gaggiotti, O.E. & Waples, R.S. (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution* 20:136–142

Mau, B. & Newton, M. (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* 6: 122–131

Michener, C.D. & Sokal, R.R. (1957) A quantitative approach to a problem in classification. *Evolution* 11:130–162

Milligan, B.G.; Leebens-Mack, J. & Strand, A.E. (1994) Conservation genetics: beyond the maintenance of marker diversity. *Molecular Ecology* 3:423–435

Moisy, C.; Garrison, K.E.; Meredith, C.P. & Pelsy, F. (2008) Characterization of ten novel *Ty*1/*copia*-like retrotransposon families of the grapevine genome. *BMC Genomics* 9:469

Murthy, V. (2011) *Population genetic analysis of common ragweed (Ambrosia artemisiifolia L.) in Europe using DNA-based molecular markers*. MSc thesis, University of Pannonia, Keszthely, Hungary

Nei, M. & Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences USA* 76:5269–5273

O'Rourke, K. (1994) The economic impact of the famine in the short and long run. *American Economic Review* 84:309–313

Page, R.D.M. & Holmes, E.C. (1998) *Molecular evolution*. Blackwell, London.

Panwar, P.; Saini, R.K.; Sharma, N.; Yadav, D. & Kumar, A. (2010) Efficiency of RAPD, SSR and Cytochrome P450 gene based marker sin accessing genetic variability amongst finger millet (*Eleusine coracana*) accession. *Molecular Biology Report* 37:4075–4082

Pleijel, F. & Rouse, G.W. (2000) Least-inclusive taxonomic unit: a new taxonomic concept for biology. *Procedings of the Royal Society of London B* 267:627-630

Poczai, P. & Hyvönen, J. (2011) On the origin of *Solanum nigrum*: can networks help? *Molecular Biology Reports* 38:1171–1185

Poczai, P.; Cernák, I.; Gorji, A.M.; Nagy, S.; Taller, J. & Polgár, Z. (2010) Development of intron targeting (IT) markers for potato and cross-species amplification in *Solanum nigrum* (Solanaceae). *American Journal of Botany* 97:e142–e145

Poczai, P.; Varga, I; Bell, N.E. & Hyvönen, J. (2011) Genetic diversity assessment of bittersweet (*Solanum dulcamara*, Solanaceae) germplasm using conserved DNA-derived polymorphism and intron-targeting markers. *Annals of Applied Biology* 159:141–153

Purvis, A.; Agapow, P.M.; Gittleman, J.L. & Mace, G.M. (2000) Nonrandom extinction and the loss of evolutionary history. *Science* 288:328–330

Ragupathy, R.; Banks, T. & Cloutier, S. (2010) Molecular characterization of the Sasanda LTR copia retrotransposon family uncovers their recent amplification in *Triticum aestivum* (L.) genome. *Molecular Genetics and Genomics* 283:255–271

Ray, D.A. (2007) SINEs of progress: mobile element applications to molecular ecology. *Molecular Ecology* 16:19–33

Ray, D.A.; Xing, J.; Salem, A-H. & Batzer, M.A. (2006) SINEs of a nearly perfect character. *Systematic Biology* 55:928–935

Relethford, J.H., Crawford, M.H. & Blangero, J. (1997) Genetic drift and gene flow in post-famine Ireland. *Human Biology* 69:443–465

Rick, C.M. & Chetelat, R.T. (1995) Utilization of related wild species for tomato improvement. *Acta Horticulturae* 412:21–38

Riechmann, J.L.; Heard, J.; Martin, G.; Reuber, L.; Jiang, C-Z.; Keddie, J.; Adam, L et al. (2001) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290:2105–2110

Rosauer, D.; Laffan, S.W.; Crisp, M.D.; Donnellan, S.C. & Cook, L.G. (2009) Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology* 18:4061–4072

Sawant, S.V.; Singh, P.K.; Gupta, S.K.; Madnala, R. & Tuli, R. (1999) Conserved nucleotide sequences in highly expressed genes in plants. *Journal of Genetics* 78:123–131

Scheifele, L.Z.; Cost, G.J.; Zupancic, M.L.; Caputo, E.M. & Boeke, J.D. (2009) Retrotransposon overdose and genome integrity. *Proceedings of the National Academy of Sciences of the United States of America* 106:13927–13932

Schmidt, K. & Jensen, K. (2000) Genetic structure and AFLP variation of remnant populations in the rare plant *Pedicularis palustris* (Scrophulariaceae) and its relation to population size and reproductive components. *American Journal of Botany* 87:678–689

Schonswetter, P. & Tribsch, A. (2005) Vicariance and dispersal in the alpine perennial *Bupleurum stellatum* L. (Apiaceae). *Taxon* 54:725–732

Shalk, M.; Nedelkina, S.; Schoch, G.; Batard, Y. & Werck-Reichhart, D. (1999) Role of unusual amino-acid residues in the proximal and distal heme regions of a plant P450, CYP73A1. *Biochemistry* 38:6093–6103

Shedlock, A.M. & Okada, N. (2000) SINE insertions: powerful tools for molecular systematics. *BioEssays* 22:148–160

Showmaker, J.S.; Painter, I.S. & Weir, B.S. (1999) Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics* 15: 354-358

Smithson, J.B. & Lenne, J.M. (1996) Varietal mixes: a viable strategy for sustainable productivity in subsistence agriculture. *Annals of Applied Biology* 128:127–158

Sokal, R.R. & Michener, C.D. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38:1409–1438.

Somerville, C. & Somerville, S. (1999) Plant functional genomics. *Science* 285:380–383

Song, Z.; Li, X.; Wang, H. & Wang, J. (2010) Genetic diversity and population structure of *Salvia miltiorrhiza* Bge in China revealed by ISSR and SRAP. *Genetica* 138:241–249

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter* 5:1–34.

Stewart, C.N. & Excoffier, L. (1996) Assessing population genetic structure and variability with RADP data: application to *Vaccinium macrocarpon* (American Cranberry). *Journal of Evolutionary Biology* 9:153–171

Sun, S-J.; Gao, W.; Lin, S-Q; Zhu, J.; Xie, B-G. & Lin, Z-B. (2006) Analysis of genetic diversity in *Ganoderma* population with a novel molecular marker SRAP. *Applied Microbiology and Biotechnology* 72:537–543

Syed, N.H. & Flavell, A.J. (2006) Sequcne-specific amplification polymorphisms (SSAPs): a multi-locus approach for analyzing transposon insertions. *Nature Protocols* 1:2746–2752

Tam, S.M.; Mhiri, C.; Vogelaar, A.; Kerkveld, M.; Pearce, S.R. & Grandbastien, M.A. (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. *Theoretic and Applied Genetics* 110:819–831

Tang, S.; Dai, W.; Li, M.; Zhang, Y.; Geng, Y.; Wang, L. & Zhong, Y. (2008) Genetic diversity of relictual and endangered plant *Abies ziyuanensis* (Pinaceae) revealed by AFLP and SSR markers. *Genetica* 133:21–30

Telenius, H.; Carter, N.P.; Beb, C.E.; Nordenskjold, M.; Ponder, B.A. & Tunnacliffe, A. (1992) Degenerate oligonucleote-primered PCR: general amplification of target DNA by single degenerate primer. *Genomics* 13:718–725

Vershinin, A.V.; Allnutt, T.R.; Knox, M.R.; Ambrose, M. J. & Ellis, T.H.N. (2003) Transposable elements reveal the impact of introgression, rather than transposition, in *Pisum* diversity, evolution and domestication. Molecular Biology and Evolution 20:2067–2075

Vitte, C.; Ishii, T.; Lamy, F.; Brar, D. & Panaud, O. (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L). *Molecular Genetics and Genomics* 272:504–511

Vos, P.; Hogers, R.; Bleeker, M.; Reijans, M.; van de Lee, T.; Hornes, M.; Friters, A. et al. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23:4407–4414

Wang, Q.; Zhang, B. & Lu, Q. (2009) Conserved region amplification polymorphism (CoRAP) a novel marker technique for plant genotyping in *Salivia miltiorrhiza*. *Plant Molecular Biology Reporter* 27:139–143

Ward, S.M.; Gaskin, J.F. & Wilson, L.M. (2008) Ecological genetics of plant invasions: what do we know? *Invasive Plant Science and Management* 1:98–109

Wheeler, W.C. (1996) Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* 12: 1–9

Wheeler, W.C. (1999) Fixed character states and the optimization of molecular sequence data. *Cladistics* 15: 379–385

Wheeler, W.C. (2003a) Implied alignment. *Cladistics* 19:261–268

Wheeler, W.C. (2003b) Iterative pass optimization. *Cladistics* 19:254–260

Whelan, K. (1997) *The atlas of the Irish rural landscape*. Cork University Press, Cork

Williams, J.G.K.; Kubelik, A.R.; Livak, K.J.; Rafalski, J.A. & Tingey, S.C. (1990) DNA polymorphsims amplified by arbitrary primers are useful as genetic markers. *Nucleaic Acids Research* 18:6531–6535

Williams, P.H.; Humphries, C.J.; Forey, P.L.; Humphries, C.J. & Vane-Wright, R.I. (1994) Biodiversity, taxonomic relatedness, and endemism in conservation, In: *Systematics and Conservation Evaluation,* Forey, P.L., Humphries, C.J. & Vane-Wright, R.I.; p. 438, Oxford University Press, Oxford

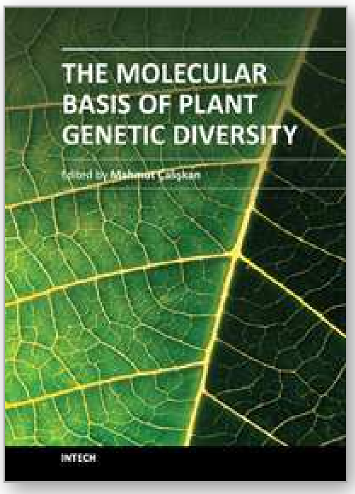
Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16:97–159

Xiong, F.; Zhong, R.; Han, Z.; Jiang, J.; He, L.; Zhuang, W. & Tang, R. (2011) Start codon targeted polymorphism for evaluation of functional genetic variation and relationships in cultivated peanut (*Arachis hypogaea* L.) genotypes. *Molecular Biology Reports* 38:3487–3494

Yamanaka, S.; Suzuki, E.; Tanaka, M.; Takeda, Y.; Watanabe, J.A. & Watanabe, K.N. (2003) Assessment of cytochrome P450 sequences offers a useful tool for determining genetic diversity in higher plant species. *Theoretic and Applied Genetics* 108:1–9

Zhivotovsky, L. (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* 8:907–913.

Zhu, Y.; Chen, H.; Fan, J.; Wang, Y.; Li, Y.; Chen, J.; Fan, J.; Yang, S.; Hu, L.; Leung, H.; Mew, T.W.; Teng, P.S.; Wang, Z. & Mundt, C.C. (2000) Genetic diversity and disease control in rice. *Nature* 406:718–722


THE MOLECULAR BASIS OF PLANT GENETIC DIVERSITY
Edited by Mahmut Çalışkan
INTECH

The Molecular Basis of Plant Genetic Diversity
Edited by Prof. Mahmut Caliskan

ISBN 978-953-51-0157-4
Hard cover, 374 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

The Molecular Basis of Plant Genetic Diversity presents chapters revealing the magnitude of genetic variations existing in plant populations. Natural populations contain a considerable genetic variability which provides a genomic flexibility that can be used as a raw material for adaptation to changing environmental conditions. The analysis of genetic diversity provides information about allelic variation at a given locus. The increasing availability of PCR-based molecular markers allows the detailed analyses and evaluation of genetic diversity in plants and also, the detection of genes influencing economically important traits. The purpose of the book is to provide a glimpse into the dynamic process of genetic variation by presenting the thoughts of scientists who are engaged in the generation of new ideas and techniques employed for the assessment of genetic diversity, often from very different perspectives. The book should prove useful to students, researchers, and experts in the area of conservation biology, genetic diversity, and molecular biology.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Peter Poczai, Ildiko Varga, Neil E. Bell and Jaakko Hyvonen (2012). Genomics Meets Biodiversity: Advances in Molecular Marker Development and Their Applications in Plant Genetic Diversity Assessment, The Molecular Basis of Plant Genetic Diversity, Prof. Mahmut Caliskan (Ed.), ISBN: 978-953-51-0157-4, InTech, Available from: http://www.intechopen.com/books/the-molecular-basis-of-plant-genetic-diversity/genomics-meets-biodiversity-advances-in-molecular-marker-development-and-their-applications-in-plant

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821