Chapter 3 in: **Which DNA Marker for Which Purpose?** Final Compendium of the Research Project **Development, optimisation and validation of molecular tools for assessment of biodiversity in forest trees** in the European Union DGXII Biotechnology FW IV Research Programme **Molecular Tools for Biodiversity**. Gillet, E.M. (ed.). 1999. URL http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm

Purposes - Classification and desired marker characteristics

Elizabeth M. Gillet[1,2]

[1] Institut für Forstgenetik und Forstpflanzenzüchtung, Universität Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

[2] Institut für Forstgenetik und Forstpflanzenzüchtung, Bundesforschungsanstaalt für Forst- und Holzwirtschaft, Sieker Landstrasse 2, 22927 Grosshansdorf, Germany

Email: egillet@gwdg.de

In the following, the purposes dealt with in the contributions to this compendium are classified, and special requirements that determine the suitability of genetic markers are discussed.

Inference of family relationships and mating system

Genetic markers provide tools for inferring both close family relationships and the realized mating system among the individuals of a community (or stand). Inference of one or both parents of an individual is termed "parentage analysis". If the (maternal) seed parent is known, as is the case when seed is collected directly from the seed tree, inference only of the (paternal) pollen donor or pollen parent is called "paternity analysis". DNA markers can also be used to identify full-sibs, *i.e.*, individuals that share both parents, as well as half-sibs, *i.e.*, individuals that share one parent (usually the seed parent) but have different pollen parents. Yet even in cases where such close genealogical ties either cannot be inferred or are not of primary interest, DNA markers can give information about the mating system that is realized in a stand.

The objective of parentage analysis is to identify both parents of an individual or, if the seed parent is known, the individual's pollen parent. In population genetics, this is done using the genotype of the individual, the genotype of the seed parent (if known), and the genotypes of all potential parents at a defined set of gene loci. All of the various methods available initially apply the "principle of exclusion", in that the alleles possessed by the individual at each of the involved loci permit all those (pairs of) potential parents to be ruled out that could not have contributed these alleles. If all but one (pair of) potential parents can be eliminated, then the remaining parent (pair) is unambiguously assigned. When more than one remains, which is very often the case, there are two ways of proceeding. One is to specify the genotypes more precisely by investigating more loci until

an unambiguous assignment is achieved. The other is to apply likelihood methods, most of which are based on models that make assumptions on the mating system. These may also require additional information, such as the spatial distance between two potential parents under the assumption of limited pollen dispersal distances (Chapter 4).

Reconstruction of sib relationships is based on the same reasoning. The identification of full-sibs may be more difficult than of a parent. Whereas an offspring will possess one of the alleles of its parent at each locus, two full-sibs need not share a single allele at a locus if both parents are heterozygous for different alleles. This situation is even more difficult for half-sibs. These problems can be alleviated if the offspring are known to stem from the same seed tree. Nevertheless, information from several gene loci is usually needed for such an analysis (Chapter 5).

The inference of family relationships can be useful for enforcing seed trade laws that aim at obtaining a seedlot that is representative for the genetic variation that is present in the harvested stand. Such laws, such as in Austria, may prescribe the harvesting of a minimum number of seed trees per seedlot. Genetic markers provide a tool for determining whether this minimum number is reached (Chapter 6).

The results of a parentage analysis can also be utilized for the model-based study of actual mating systems. Some of the models commonly investigated in forest population genetics are the models of random mating, mixed mating, inbreeding, assortative mating with respect to certain phenotypic traits, and distance-dependent pollen dispersal. In order to test these models using real data, the actual matings observed or inferred in a population can be used to calibrate the free model parameters (such as the allelic frequencies within the pollen cloud, the rate of self-fertilization, mating preferences, pollen dispersal functions, etc.).

Yet even when parentage analysis has not been performed, inferences on the realized mating system can be drawn on the basis of the observed distribution of the genotypes at one or more loci. In particular, certain patterns of disparities between the actual genotypic structure and the structure that would be expected under random mating are typical for particular mating systems. Inbreeding, which is considered to be undesirable in forest reproductive material, leads to an excess of homozygotes (Chapter 7). Other examples are differential individual ovule and pollen fertilities, which lead to sexual asymmetry between the ovule and pollen pools and thus to an excess of heterozygotes, and assortative mating , which can lead to seemingly erratic disparities between the actual and expected frequencies of the different genotypes.

The ideal marker for parentage analysis and reconstruction of sib relationships is one for which (1) the parental phenotypes are inferrable from the phenotypes of an offspring and (2) for which knowledge of the parental phenotypes in turn allows unambiguous identification of the parents as individuals. If the trait is a genetic marker, then these criteria are fulfilled if any two potential parents have different alleles for at least one of the loci. Thus suitable markers for parentage analysis are of the following types:

- a few single-locus markers, each showing codominance of gene action and possessing a large number of alleles (*e.g.* nuclear microsatellites);
- a large number of codominantly expressed single-locus markers with intermediate numbers of alleles (*e.g.* isoenzymes);
- a very large number of dominantly expressed (*i.e.*, presence/absence of bands) single-locus markers (*e.g.* AFLP);
- uniparentally inherited, polymorphic markers (*e.g.* chloroplast microsatellites) for identification of the parent of the transmitting gender.

The following four contributions to this compendium deal with the inference of family relationships and the mating system. All of them are based on results of studies in oak stands using nuclear microsatellites or AFLPs:

Chapter 4: Comparison of microsatellites and AFLP markers for parentage analysis - S. Gerber, S. Mariette, R. Streiff, C. Bodénès, A. Kremer

Chapter 5: DIG-labelled AFLPs in oaks - A DNA marker for reconstruction of full- or half-sib family relationships? - B. Ziegenhagen, V. Kuhlenkamp, R. Brettschneider, F. Scholz, B.R. Stephan, B. Degen

Chapter 6: Microsatellite analysis of anonymous seedlot samples from oak: a promising approach to monitor the number of different seed parents and pollen donors - C. Lexer, B. Heinze, S. Gerber, H. Steinkellner, B. Ziegenhagen, A. Kremer, J. Glössl

Chapter 7: Nuclear microsatellites as a tool in the genetic characterization of forest reproductive material. A case study in sessile oak (*Quercus petraea* Matt., Liebl.) - S. La Scala, R. Schubert, G. Müller-Starck, K. Liepe

Distribution of within-species genetic variation over space and time

The genetic variation within a species is usually expressed in terms of the different genetic types (alleles) that occur at each of a given set of observable loci and the relative frequencies of their presence among the individuals belonging to the species, that is, the genetic structure. The genetic structure can conceivably be expressed for the entire species. As a rule, however, it will vary locally among the different populations that are distributed over the species' range. For example, changes in genetic structure can occur when a new population is established from an older population due to the operation of genetic processes (*e.g.* adaptive selection, mutation, drift). Once the new population has become established, continued operation of these and additional genetic processes (*e.g.* selection, gene flow from other populations) may cause further changes in its genetic structure over time. The same applies to the older population. The result is spatial and temporal variation among the genetic structures of the species' populations.

The analysis of genetic structures at marker loci can allow historical inference of the processes that led to the present-day distribution of populations. Observation of the genetic structures of the single populations yields a pattern of genetic variation over the species' range. This in turn can be used to reconstruct the spatial movement of populations (*e.g.* post-glacial remigration pathways, artificial establishment of a population using reproductive material taken from a different geographic region) and/or the genetic processes that have been most important in molding genetic variation since their establishment.

Such inference is usually model-based. Models yield predictions about the genetic structures at marker loci that are then compared to the observed genetic structures in the populations. The required characteristics of the genetic markers vary between models. Models of processes such as colonization, gene flow, or drift require the selective neutrality of the marker genotypes (or rather, phenotypes). In order to model adaptation, markers are needed that are either themselves adaptive or are associated with loci that control adaptive traits. For the analysis of mating systems, both neutral markers for parentage analysis and markers that are linked to loci that determine mating behavior may be needed. Models that strive to estimate the time since separation of two populations require selectively neutral markers in addition to the very questionable assumption that the mutation rate at the marker locus remains constant over time ("molecular clock" hypothesis).

Nevertheless, even without such assumptions, the observation of differences between the genetic

structures of populations can be indicators of historical processes. A *qualitative* difference between the genetic structures of two populations signifies the presence of one or more genetic types at appreciable frequencies in the one population but their complete absence in the other. This is a strong indication that they have been separated for many generations, either reproductively or due to their existence under widely deviating environmental conditions. A large *quantitative* difference, measured as genetic distance, between genetic structures is also an indication of a long separation of two populations.

Examples for conclusions that can be drawn from the observation of qualitative or quantitative differences among populations are the following: Clinal variation in the genetic structure (*i.e.*, a gradual change in the genetic structure of populations along a geographical transect) can be evidence for the migration of the populations along this pathway during their establishment (selectively neutral marker) or for the adaptive differentiation of populations in response to a gradual change in environmental conditions (marker locus under selection or associated with selected locus). The observation that a particular genetic structure, or especially a genetic type, is typical for all populations within one region but is not found elsewhere can be used for provenance identification (neutral marker). Genetic structures that are typical for certain environments can identify ecotypes (selective marker). The introgression of one population into another can be inferred using either type of marker, provided the introgressing population has a genetic type that does not occur in the introgressed population.

In this section of the Compendium, markers are described and used for historical inference that are presumably selectively neutral, meaning that differences among the genetic structures of populations are subject only to (rare) mutation and drift. One contribution discusses the properties of the highly variable nuclear microsatellites that determine their suitability for assessing genetic variation within and among populations (Chapter 8). In a second contribution, the geographic distribution of genetic diversity at a paternally inherited chloroplast microsatellite locus is assessed in a number of conifer species; analysis of the resulting distribution patterns provides evidence for historical processes, such as the location of European Ice-Age refugia and post-glacial migration pathways as well as for introgression between species (Chapter 9). Finally, the utility of a maternally inherited insertion/deletion polymorphism in the mitochondria of Norway spruce is demonstrated for provenance identification; in particular, evidence is found for the introduction of forest reproductive material into Sweden (Chapter 10).

The three contributions that treat purposes of this type are:

Chapter 8: Microsatellite markers as a tool for the detection of intra- and interpopulation genetic structure - I. Scotti, G. Paglia, F. Magni, M. Morgante

Chapter 9: Chloroplast microsatellites for analysis of the geographic distribution of diversity in conifer species - M. Anzidei, A. Madaghiele, C. Sperisen, B. Ziegenhagen, G.G. Vendramin

Chapter 10: Mitochondrial DNA variation provides a tool for identifying introduced provenances: A case study in Norway spruce - C. Sperisen, U. Büchler, G. Mátyás, L. Ackzell

Phylogeny reconstruction

The objective of phylogeny reconstruction among species is to infer their evolutionary relationships on the basis of observed similarities and differences in the characteristics of the individual species. For this purpose, traits are identified that show variation of trait state among the species but (ideally) fixation within species. The types of trait that are used for phylogenetic analysis generally fall into three categories:

Category 1: A number of traits considered simultaneously, the states of which are fixed within each species - It is assumed that all of the species that possess the same trait state will have originated from a common ancestor that also possessed this trait state, that is, these species form a "monophyletic group". In this case, arrangement of the monophyletic groups that are defined by each trait may yield an unambiguous phylogeny. In order to avoid the problems of defining homology of traits between species, traits are often chosen to be characteristics that show only one of the two states "presence" and "absence" in each species.

Category 2: A single trait that fulfills two criteria: The state of this trait is fixed within each species, and a measure of "phenotypic distance" between trait states is given that is assumed to reflect their evolutionary distance - For example, the number of nucleotide substitutions in homologous DNA sequences is often considered to be proportional to the time that has elapsed since the speciation event ("molecular clock" hypothesis based on an assumed constant rate of mutation).

Category 3: A single trait, the states of which may be possessed by more than one species but in varying relative frequencies - An applied measure of genetic distance between the frequency distributions of the trait states within the single species is assumed to reflect the evolutionary distances between the species. Phylogenetic reconstruction based on comparison of the frequency distributions of the states of a trait among species. In this case, the trait state varies within species, and more than one species may possess the same trait state. However, due to the difficulties of defining the same trait in different species, this approach is more commonly applied to reconstruct phylogenies between populations of the same species than between species.
Not only the types of trait that are used for phylogenetic reconstruction vary. The methods of reconstruction that have been developed for each of these types of trait are also manifold.

One contribution describes the limitations that can be imposed on using nucleotide sequences as traits for phylogeny reconstruction (Chapter 11). A second contribution compares the suitability of genetic marker types AFLP and nuclear microsatellites as traits for phylogenetic reconstruction (Chapter 12).

Chapter 11: Limitations to the phylogenetic use of ITS sequences in closely related species and populations - a case study in *Quercus petraea* (Matt.) Liebl. - G. Muir, C. Schlötterer

Chapter 12: Amplified Fragment Length Polymorphisms and Microsatellites: A phylogenetic perspective - J.P. Robinson, S.A. Harris

Identification of expressed genes

Expressed genes that are transcribed into messenger RNA (mRNA) and translated into a gene product (usually a protein or enzyme) can interact with the environment to define an individual's phenotype. Gene expression may be constitutive ("household genes") or regulated. Some regulated genes are "turned on" only in particular ontogenetic stages or cell types (*e.g.* some genes are active in seeds but not in adult trees and *vice versa*). The expression of other regulated genes is induced (or repressed) only by environmental stimuli (*e.g.* drought or fungal infestation).

Polymorphism in the form of DNA sequence variation within the transcribed regions of a gene (*i.e.*, allelic variation) can result in phenotypic differences between individuals. These differences may be selective in some environments, *i.e.*, certain phenotypes may have an advantage under natural selection (viability, fecundity) or artificial selection (*i.e.*, preferred growth form) over other phenotypes. A selective advantage in one environment may be a selective disadvantage in another.

For the purposes dealt with in the previous sections of this Compendium, it was not of primary concern whether or not a genetic marker represented an expressed gene. Some purposes even required that the genetic types show selective neutrality, and thus suitable markers were found in regions of the genome that are not transcribed (*e.g.* microsatellites). This situation is different when the object of study is the expression of polymorphic genes that are directly involved in the physiological response of individuals to their environments. The purpose of such studies can range from the search for an answer to the question of why some individuals are less sensitive to a particular external stress factor than others, over the discovery of alleles of expressed genes that identify "ecotypes" and thus can provide evidence for the introduction of reproductive material from one environment into another, to breeding for desired phenotypes. Two types of marker can be suitable for such purposes:

(1) *Expressed sequence tag (EST) markers that allow the direct discernment of allelic variants of an expressed gene:* Molecular methods allow discernment of the different types (clones) of mRNA that are present in the cells of an individual at the time of the investigation. The DNA template of each mRNA clone can be reconstructed by reverse transcription. This DNA is termed cDNA ("copy" or "complementary" DNA). Knowledge of the nucleotide sequence of the cDNA in turn allows the development of EST (expressed sequence tag) markers whose variants show a 1:1 relationship to the DNA sequence variants of the exons (*i.e.*, the coding sequences of a gene, as opposed to interspersed sequences, or introns, that are excised during RNA processing). The actual function of an expressed gene can often be inferred by comparing the nucleotide sequence of the cDNA to sequences listed in data bases. These rapidly growing data collections contain the DNA sequences of expressed genes together with their amino acid sequences, from which their function can been identified.

It is possible to detect genes whose expression is ontogenetic-stage-dependent or stress-induced by determining which mRNA clones are present in different ontogenetic stages or before and after exposure to environmental stress, respectively. The purpose of one contribution is to identify expressed genes that are involved in the physiological response to certain stress factors in two tree species (Chapter 13). In oak, these are genes that are expressed in response to osmotic stress, and in spruce, these are genes that are expressed in the presence of a fungal elicitor. For this purpose, mRNA clones were scored that were present in the cells of individuals after, but not before, exposure to the respective stressor. From these mRNA clones, the cDNA sequences were inferred. Polymorphic EST markers could be developed from some of these. The function of many of the underlying genes could be identified.

(2) *Markers that reveal genetic variation at a gene locus that is linked to the locus of an expressed gene:* The method described above can lead to knowledge of the exact DNA sequence of expressed genes. The involved work is, however, complex and time-consuming. For purposes such as breeding, it may be sufficient to develop easily scored markers, the variants of which show stochastic association to a desired phenotype. It is hoped that the marker locus will be linked to a gene, the expression of which is a main determinant of this phenotype. The marker itself need not represent an expressed gene. The final contribution to this Compendium describes the utilization of AFLP® markers in marker assisted plant breeding (Chapter 14).

Chapter 13: Isolation and sequence analysis of oak and spruce cDNA clones - M. Berenyi, S. Fluch, K. Hohl, K. Burg, R. Schubert, R. Riegel, G. Müller-Starck

Chapter 14: Application of the AFLP® technique in marker assisted breeding - J. Peleman