Chapter 6 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

Microsatellite analysis of anonymous seedlot samples from oak: a promising approach to monitor the number of different seed parents and pollen donors

C. Lexer[1]*, B. Heinze[2], S. Gerber[3], H. Steinkellner[1], B. Ziegenhagen[4], A. Kremer[3], J. Glössl[1]

[1] Zentrum für Angewandte Genetik, Universität für Bodenkultur Wien, Muthgasse 18, 1190 Vienna, Austria

[2] Forstliche Bundesversuchsanstalt Wien, Hauptstraße 7, 1140 Vienna, Austria

[3] INRA, Laboratoire de Génétique et d'Amélioration des Arbres Forestiers, BP 45, 33611 Gazinet Cedex, France

[4] Institut für Forstgenetik und Forstpflanzenzüchtung, Bundesanstalt für Forst- und Holzwirtschaft, Sieker Landstraße 2, 22927 Großhansdorf, Germany

*Corresponding author: Email: clexer@edv2.boku.ac.at

Introduction

The sustainable conservation of genetic resources in forestry is of great concern to the governments in Europe (Anonymous 1990 - Strasbourg resolution). Several countries have made provisions in their legislation for regulation of the marketing or use of forest reproductive material (seeds and plants). *Genetic diversity* at the within-stand level (often called "a broad genetic base" in public debate) is increasingly regarded as one of the key criteria for quality of forest seed material (Geburek and Heinze 1998).

Genetic diversity within commercial forest seed material can be measured using diversity parameters derived from population genetic theory (*e.g.* Nei 1987, Hattemer *et al.* 1993). However, in practice it may be desirable to express diversity in a simpler way: as the *number of seed parents* included in the seed harvests and the *number of pollen donors* contributing to the seedlots.

A first attempt to include the *number of different seed parents* in legal regulations about forest seed material has been made in Austria. Here, a federal forest seed law regulates the marketing of seed material for a variety of broad leaf tree species and conifer species (Anonymus 1996 a,b). According to this law, marketable seedlots have to be collected from a certain *minimum number of seed parents* per stand, depending on the particular tree species. In order to allow genetic monitoring of the seedlots, *small anonymous seedlot samples* must be shipped to the Federal Forest Research Centre in Vienna by the companies involved in the seed harvesting activities. *Samples from different seed parents are supplied separately*. The aim is to monitor *genetic relationships* within and among the seedlot samples, to detect *seed contaminations* if necessary and to infer the *number of different seed parents*.

In the specific case of European oaks (*Quercus petraea* and *Quercus robur*), seedlots in Austria

must be collected from at least *20 different trees* per stand. For any marketable acorn seedlot in Austria, at least *5 acorns* per seed parent must be shipped to the Federal Forest Research Centre for genetic analysis. In the frame of the present EU project, we have tested the potential of *microsatellite markers* to analyse the genetic composition of commercial seedlots harvested from *Q. robur* and *Q. petraea*. European oaks were chosen as model tree species because a sufficient number of microsatellite markers was available.

Microsatellites are short, repetitive DNA sequences, consisting of tandemly repeated mono-, di-, tri-, tetra- or pentanucleotide units dispersed throughout the genomes of most eucaryotic organisms (see review by Powell *et al.* 1996). The uniqueness and value of microsatellites arises mainly from their multi-allelic ("polymorphic") nature: microsatellite loci generally exhibit many different alleles, each present at low frequency. Microsatellites can be isolated from virtually any target species of interest, and the entire locus, including the repeat region and the flanking regions, can be sequenced. The polymorphism in the microsatellite repeat region can then be visualised by PCR amplification of the microsatellite repeat, using specific PCR primers complementary to the DNA sequences flanking the repeat. The polymorphism becomes visible after running the PCR products on a suitable electrophoresis system (Morgante and Olivieri 1993). Following this approach, the alleles at a microsatellite locus can be PCR-amplified in virtually any individual of a given species.

In the present contribution, microsatellites were chosen as molecular tools because they are highly polymorphic (*i.e.*, very informative) and because they are inherited in a codominant Mendelian manner (*i.e.*, they reveal either heterozygosity or homozygosity in each individual). Hence, microsatellites have a high potential to resolve genetic relatedness.

Objectives

Our objective was to test the potential of microsatellite markers to elucidate the genetic composition of *small anonymous seedlot samples* from oak. We focussed on (1) detection of *seed contaminations*, (2) inference of the *number of seed parents* directly from small anonymous seedlot samples and (3) inference of *the number of pollen donors* directly from seedlots or samples thereof. Our aim was to identify and test methods of data analysis that require *no genotype information of the parent population*.

Materials and methods

The experiments presented here are based on *maternal half-sib families* (= open pollinated families). As seedlots from oaks or other outcrossing tree species are generally harvested in a single-tree manner, such families represent the "basic elements" of commercial forest seed material. In the present pilot study we analysed a *model half-sib family* harvested from a known seed parent, 8 *simulated half-sib families*, and a limited number of *small anonymous seedlot samples*. These seedlot samples had been shipped to the Federal Forest Research Centre in Vienna by a forestry company in the course of commercial seed harvest. All of the samples as well as the computer simulations are described in more detail in Lexer et al. (1999) and Lexer et al (in press).

Nine microsatellite markers isolated from the genomes of *Q. petraea* (Steinkellner *et al.* 1997) and *Q. robur* (Kampfer *et al.* 1998) were used for genetic analysis. All of the markers have been located on a genetic linkage map of *Q. robur* (Barreneche *et al.* 1998). Three of the microsatellites are closely linked on the genetic map, with recombination frequencies ranging between 1.1 and 5.6 %. The linkage relationships as well as information about the polymorphism of the markers can be found in Table 1.

Locus	A_O	A_E	H_E	Linkage
ssrQpZAG 1/5	9	5.6	0.82	Lg 7[a]
ssrQpZAG 9	10	6.7	0.85	Lg 7[a]
ssrQpZAG 15	9	4.0	0.75	Lg 9
ssrQpZAG 3/64	18	10.0	0.90	Lg 6
ssrQpZAG 110	8	2.5	0.60	Lg 8
ssrQrZAG 112	15	7.1	0.86	Lg 12
ssrQpZAG 36	9	5.6	0.82	Lg 2[b]
ssrQpZAG 46	17	9.1	0.89	Lg 2[b]
ssrQpZAG 104	15	9.1	0.89	Lg 2[b]

Table 1: Microsatellite markers used in this study, their observed number of alleles (A_O), effective number of alleles (A_E), and expected heterozygosity (H_E) calculated from 36 seedlings originating from one population. Assignments to linkage groups refer to the genetic map of *Q. robur* (Barreneche *et al.* 1998). [a]) positioned on the same linkage group, but separated by 24 cM (male map) and 37.5 cM (female map), respectively. [b]) closely linked (see Materials and Methods).

General estimates of diversity were calculated for each locus in terms of expected heterozygosity $H_E = 1- \text{Sum}_i\, p_i^2$, where p_i is the relative frequency of the i^{th} allele, and in terms of the effective number of alleles $A_E = 1/(1-H_E)$, according to Nei (1987).

For *detection of seed contaminations* and reconstruction of the maternal genotypes, we made use of a color code that assigns a specific color to each allele within a sample. Subsequently, the correct maternal genotypes were inferred from the offspring genotypes by using the rules of codominant Mendelian inheritance, as described in more detail by Lexer *et al.* (1999).

Unrelated individuals within seedlot samples were detected using two methods, one based on private alleles, present only in one individual within a sample, the other one based on the proportion of shared alleles among pairs of seedlings within a sample. Private alleles were identified using color coding as described above. The proportion of shared alleles Ps for pairs of individuals was calculated as the number of shared alleles summed over loci divided by twice the number of loci, as in Bowcock *et al.* (1994). A genetic distance between pairs of individuals was obtained by -ln (Ps) using the computer program Microsat (Minch 1997). UPGMA cluster analysis was conducted using the PHYLIP software package (Felsenstein, 1989).

The number of different seed parents was inferred by comparing the reconstructed maternal genotypes and by calculating F_{ST} pairwise between the seedlot samples. F_{ST} was calculated according to Weir and Cockerham (1984) using the FSTAT software (Goudet 1995). Seven unlinked or loosely linked loci were used for this purpose, including ssrQpZAG 104 from linkage

group 2 (see Table 1). A principal component analysis based on pairwise F_{ST} values was conducted using the SPSS software package (SPSS Inc., Chicago).

An approach to estimating *the number of different pollen donors* contributing to seedlots or samples thereof was developed with the help of simulated half-sib families (sample size 40 progeny). First, pollen haplotypes were obtained for 8 different simulated families by subtracting the (known) maternal alleles from the progeny genotypes. Next, the genetic variability of 3 linked microsatellite loci (listed in Table 1) was combined in order to count paternal haplotypes for each family. The haplotype sorting and counting was conducted with the software EXCEL 97 (Microsoft Corporation). Next, the haplotype counts were corrected for rare recombination events between linked markers as described in more detail in Lexer *et al.* (in press). The resulting values were adjusted to give discrete numbers of paternal chromosomes. This "haplotype approach" was used to count the number of paternal chromosomes in 8 simulated half-sib families. The haplotype counts were compared to the (known) number of pollen donors in the datasets by regression statistics. Regression analysis and curve estimation were conducted using the SPSS 8.0 software package (SPSS Inc, Chicago). Finally, the regression function was used to estimate the number of pollen donors in a "real" genotyped half-sib family.

Results

Seed contaminations

An approach *to detecting seed contaminations* was identified and tested using the genotype data of 5 *anonymous* seedlot samples from a forestry company, consisting of 4 to 8 individuals each. It was shown that seedlot samples containing contaminations, *i.e.*, unrelated plants within the half-sib offspring, can be identified using a color coding that assigns a specific color to each allele within a sample (not shown; Lexer *et al.* 1999). Subsequently, the unrelated plants can be identified within such samples by calculating a genetic distance between individuals (Bowcock *et al.* 1994) and by resolving the distance matrix with a *cluster analysis*. As a confirmation, unrelated plants were also identified by counting the number of *private alleles* for each seedling, where private alleles are defined as alleles that are present in only one seedling within a sample. Seedlings that were located in a distal position on the UPGMA phenograms also displayed the largest number of private alleles. Using these approaches, 3 samples were shown to contain 1 unrelated plant each (Figure 1). After removing these 3 unrelated plants from the datasets the genotypes of the remaining seeds were consistent with the rules of codominant Mendelian inheritance, indicating that the remaining seeds were indeed half-sibs (for details see Lexer *et al.* 1999). The results suggest that microsatellites are useful (i) to identify samples with contaminations and (ii) to identify the unrelated plants within such samples.

The detection of seed contaminations may be explained by the fact that the acorns were collected from the ground, as is usually done for commercial acorn seedlots. This opens the possibility of dispersal by animals such as squirrels or jays. Moreover, acorns may be removed beyond the crown of their seed parent simply by gravitation. The experiments presented here show that such contaminations, whatever their cause, can be detected with microsatellites.

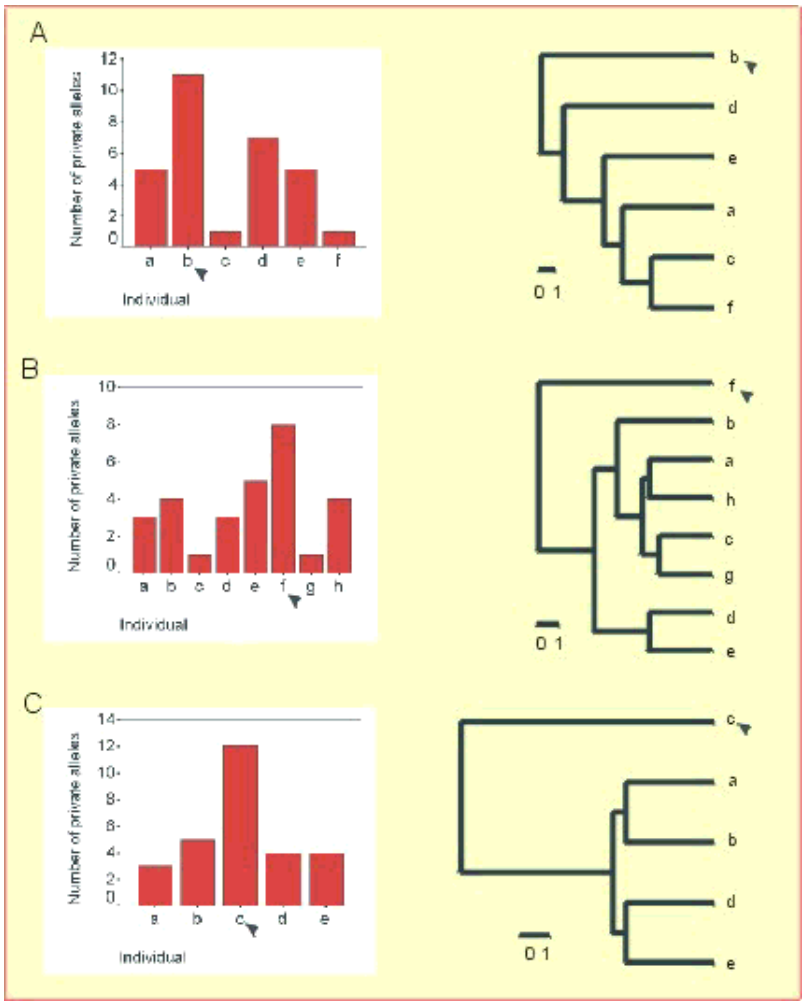

Figure 1: Two methods for detecting unrelated individuals within seedlot samples. *Left:* The number of private alleles - alleles that are present in only one seedling within a sample. *Right:* UPGMA phenograms using the proportion of shared alleles. The calculations were based on 9 microsatellite loci. A, B, C indicate 3 different seedlot samples. Unrelated individuals within the samples are indicated by *black arrowheads*.

Inferring the number of different seed parents

Two approaches were identified and tested to infer *the number of different seed parents* directly from small, anonymous seedlot samples. The first approach involves inference of the maternal alleles directly from the seedlot samples and subsequent comparison of the maternal genotypes. This approach is described in detail by Lexer *et al.* (1999). A second approach is based on F_{ST} calculated pairwise between the seedlot samples. An example is given by Lexer *et al.* (1999): in that study, F_{ST} was calculated pairwise between 10 seedlot samples. Five of these samples were known to originate from *different* seed parents, while the remaining 5 samples were known to originate from the *same* seed parent. Pairwise F_{ST} was calculated among the 10 samples using 7 unlinked or loosely linked microsatellites (see Table 1 and "Materials and methods"). Subsequently, the pairwise F_{ST} matrix was subjected to a principal component analysis (PCA). This multivariate technique allowed reduction of the dataset into 3 variables, the first 3 principal components, accounting for 56%, 24% and 11% of the total variance in the data, respectively. A plot of the first

3 principal components revealed that F_{ST} between the 5 samples harvested from the same seed parent was extremely weak, while F_{ST} between the 5 samples harvested from different seed parents appeared to be much more pronounced (Figure 2).

The seed parents' genotypes have a strong effect on F_{ST}. Identical seed parents create low pairwise F_{ST} values, whereas different seed parents cause higher pairwise F_{ST} values. Our results suggest that microsatellites may be suitable to infer the number of different seed parents directly from small anonymous seedlot samples, either by direct comparison of the reconstructed seed parents' genotypes or by calculating F_{ST} pairwise between the families.

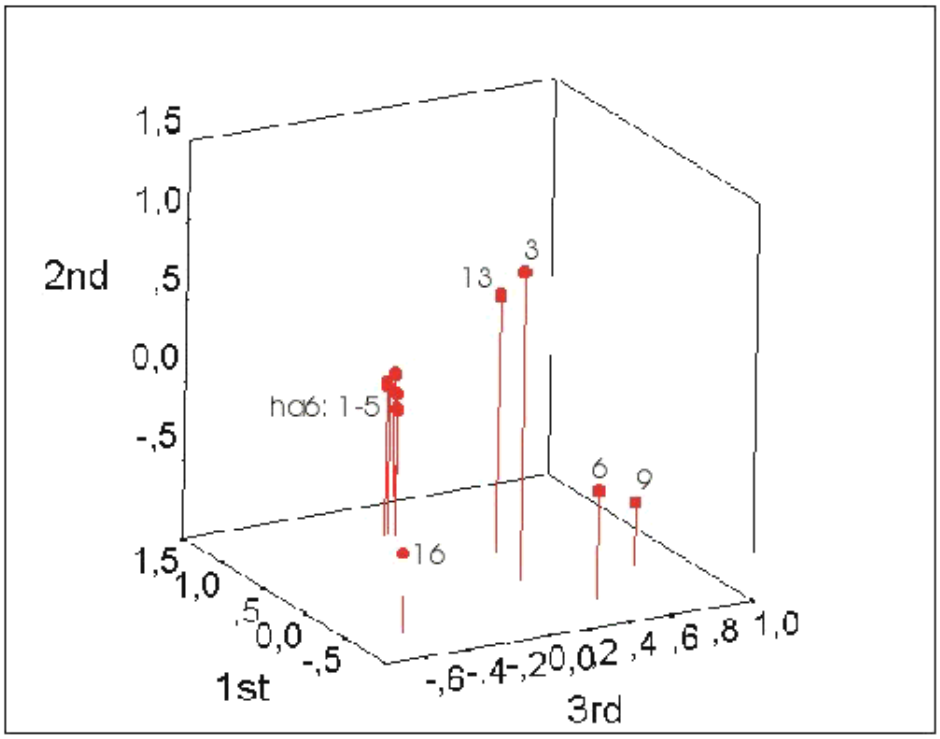

Figure 2: Plot of the first three principal components of pairwise F_{ST} between 10 samples. The calculations were based on 7 unlinked or loosely linked microsatellites. *3, 6, 9, 13, 16* indicate samples harvested from different seed parents, *ha6: 1-5* indicate samples harvested from the same seed parent. The first three principal components accounted for 56%, 24% and 11% of the total variance, respectively.

Inferring the number of different pollen donors

An approach to infer the number of different pollen donors contributing to seedlots or samples thereof was developed and tested using simulated data of *linked microsatellites* segregating in maternal half-sib families. Linked markers were used, as such markers are transmitted from the pollen donors to the progeny with limited recombination. Therefore, the genetic information of

linked markers can be combined to count paternal haplotypes, as described in "Materials and methods" and in more detail by Lexer *et al.* (in press). In that study, up to 5 linked (simulated) microsatellites were employed to count paternal haplotypes in 8 different (simulated) families, and the haplotype counts were compared to the (simulated) number of pollen donors in each family by regression statistics. The main result of this simulation study was a logarithmic *regression function* relating the *haploid chromosome count* to the *diploid number of pollen donors* in simulated datasets (Lexer et al. in press). Figure 3 of the present contribution shows the corresponding regression function when three linked microsatellites are employed to count paternal haplotypes.

Finally, the regression curve shown in Figure 3 was employed to estimate the number of pollen donors in a "real" genotyped half-sib family (sample size 43 progeny). In this family, paternal haplotypes were counted using the "haplotype approach", then the number of pollen donors was estimated using the regression function. This calculation revealed a lower 95 % confidence limit of 27, suggesting that at least 27 pollen donors had contributed to the progeny. The calculation as well as the plausibility and accuracy of the result are discussed in detail by Lexer *et al.* (in press).

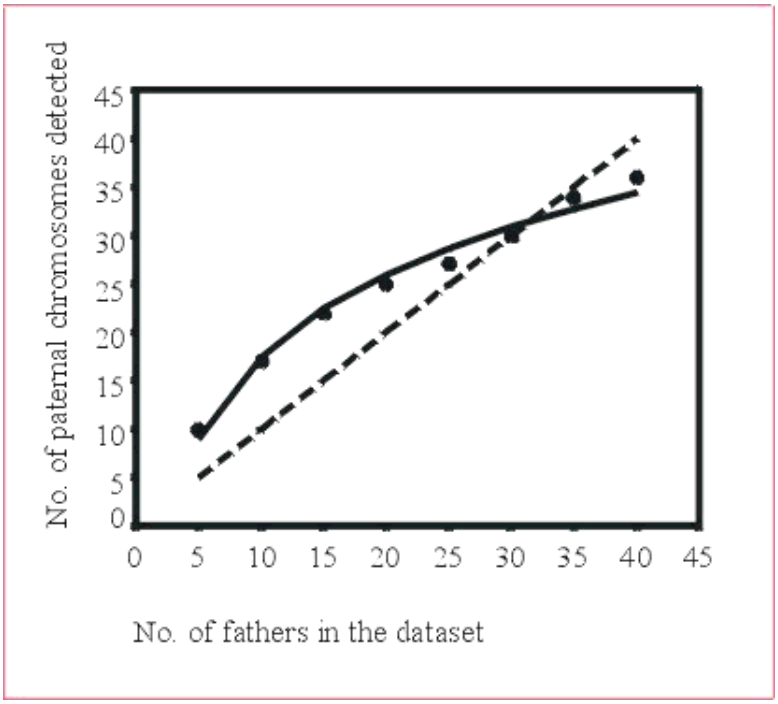

Figure 3: The relationship between the number of paternal chromosomes detected with three linked microsatellite loci and the number of pollen donors in different simulated datasets. The *datapoints* represent the results of 8 different simulated families of 40 progeny each. The *solid line* represents the logarithmic regression curve, the *dashed line* represents perfect collinearity.

Conclusions

According to the present pilot study, microsatellites are promising tools to study the genetic composition of small anonymous seedlot samples. We have identified and tested methods of data analysis to detect *seed contaminations*, to infer *the number of different seed parents* and to infer *the number of different pollen donors* from anonymous seedlot samples. The methods identified and

tested within this study have several features in common: they are suitable for small to moderate sample sizes, they do not require any genotype information of the parent population, and they rely on small to moderate numbers of PCR-based genetic markers. Hence, they are well suited to practical situations with limited financial resources and a need for quick answers. Moreover, the approaches presented here may be useful for genetic analysis of seed material from other outcrossing tree species as well. In the present project, European oaks were chosen as a model, because a sufficient number of microsatellites was available and because the markers have been mapped genetically. This provided an opportunity to choose either independent or closely linked microsatellites from the genome, as the situation demanded. However, as the seed material of many other outcrossing tree species is structured in a similar way, the approaches tested within this study may also be applicable to other tree species in the future.

The present pilot study was stimulated by recent changes in Austrian forestry legislation (see Introduction). Our results suggest that microsatellites are suitable to fulfil these requirements, *i.e.,* to detect *seed contaminations* and to infer *the number of different seed parents*. The number of different pollen donors contributing to seedlot samples has been included here as an additional feature. In principle, one could think of combining all of the above methods into one survey. Genetic relationships within small anonymous seedlot samples, the number of seed parents and the number of pollen donors could be monitored within one and the same study. Moreover, the number of pollen donors could also be studied by combining paternal "chromosome counts" of more than one family. However, to establish microsatellite analysis of commercial acorn seedlots in practice, more data are required on the effect of *sample size* (*i.e.,* the number of seeds per family) and the effect *of sampling the genome* (*i.e.,* the number of markers chosen from the genetic map).

References

Anonymous (1990) Ministerial Conference on the Protection of Forests in Europe (Strasbourg resolution). Ministère de l'Agriculture et des Forêts, Paris, France.

Anonymus (1996a) Bundesgesetzblatt für die Republik Österreich. 1996/419. Bundesgesetz über forstliches Vermehrungsgut (Forstliches Vermehrungsgutgesetz), Vienna, Austria.

Anonymus (1996b) Bundesgesetzblatt für die Republik Österreich. 1996/512. Verordnung: Forstliches Vermehrungsgut, Vienna, Austria.

Barreneche T, Bodenes C, Lexer C, Trontin JF, Fluch S, Streiff R, Plomion C, Roussel G, Steinkellner H, Burg K, Favre JM, Glössl J, Kremer A (1998) A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5s rDNA markers. *Theor. Appl. Genet.* 97: 1090-1103.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455-457.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package. *Cladistics* 5: 164-166.

Geburek TH, Heinze B (eds.) (1998) *Erhaltung genetischer Ressourcen im Wald-Normen, Programme, Maßnahmen.* Ecomed-Verlagsgesellschaft, Landsberg, Germany.

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F statistics. *J. Hered.* 86: 485-486.

Hattemer HH, Bergmann F, Ziehe M (1993) *Einführung in die Genetik für Studierende der Forstwissenschaft*. 2nd edition, J.D. Sauerländer's Verlag, Frankfurt am Main, Germany.

Kampfer S, Lexer C, Glössl J, Steinkellner H (1998) Characterization of $(GA)_n$ microsatellite loci from *Q. robur*. *Hereditas* 129: 183-186.

Lexer C, Heinze B, Steinkellner H, Kampfer S, Ziegenhagen B, Glössl J (1999) Microsatellite analysis of maternal half-sib families of *Quercus robur*, pedunculate oak: Detection of seed contaminations and inference of the seed parents from the offspring. *Theor. Appl .Genet.* 99: 185-191.

Lexer C, Heinze B, Gerber S, Macalka-Kampfer S, Steinkellner H, Kremer A, Glössl J (2000) Microsatellite analysis of maternal half-sib families of *Quercus robur*, pedunculate oak (II): Inferring the number of pollen donors from the offspring. *Theor. Appl .Genet.,* in press.

Minch E (1997) MICROSAT, Version 1.5b. Stanford University Medical Center, Stanford.

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. *The Plant Journal* 3 (1): 175-182.

Nei M (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York, U.S.A.

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* 1 (7): 215-222.

Steinkellner H, Fluch S, Turetschek E, Lexer C, Streiff R, Kremer A, Burg K, Glössl J (1997) Identification and characterization of $(GA/CT)_n$ -microsatellite loci from *Quercus petraea. Plant Mol. Biol.* 33: 1093-1096.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.