PATTERNS OF NUCLEOTIDE VARIATION AND GENE-ASSOCIATED SNP ANALYSIS IN A *QUERCUS* spp. FOREST AT ISOCITRATE DEHYDROGENASE GENES

Dissertation

Submitted in partial fulfillment of the requirements
for the degree of Doctor *rerum naturalium* (Dr. rer. nat.)
at the Faculty of Forest Sciences and Forest Ecology,
Georg-August University of Göttingen

by

Amaryllis Vidalis
born in Athens, Greece

Göttingen, 2010

To the memory of my father, to my mother and my sister

ACKNOWLEDGMENTS

For the completion of this thesis, I am deeply grateful to Prof. Dr. Reiner Finkeldey, whose encouragement, guidance and support from the initial to the final level enabled me to evolve scientifically and personally.

My warmest gratitude I owe to Dr. Bärbel Vornam for her valuable advice and guidance throughout the laboratory work of this dissertation, and Ass. Prof. Dr. Oliver Gailing for his accurate scientific support and for proofreading the text draft of Manuscript I and II.

I am indebted to Prof. Dr. Martin Ziehe, Dr. Ludger Leinemann, Dr. Elizabeth Gillet, Prof. Dr. Emeritus Hans Heinrich Hattemer, Prof. Dr. Hans-Rolf Gregorius, Dr. Kathleen Prinz and Prof. Dr. Konstantin Krutovsky for their contribution to this work through constructive discussions and by sharing their knowledge and opinions. My special gratitude also goes to Ass. Prof. Dr. Aristotelis C. Papageorgiou and Ass. Prof. Dr. Ioannis Tsiripidis for supporting me in many ways and believing in me.

To Alexandra Dolynska, Gerold Dinkel, Olga Artes, Thomas Seliger and Christine Radler I am sincerely thankful, for their eminent technical assistance and for their mental support. To Marita Schwahn and Regina Berkeley I am indebted for their excellent administrative work and for keeping good balance.

I am deeply grateful to Dr. Arne Weiberg for proofreading the text draft and for helping me with the German translation of the summary but mainly for our interesting discussions, his generous scientific and mental support and his devoted encouragement.

Special thanks to all my former and current colleagues in the Department of Forest Genetics for exchanging ideas, questions, answers, feelings, emotions, knowledge, culture and for living together during the last years. In particular, I am sincerely greatful to Dr. Alexandru Lucian Curtu for providing me with digital forms of morphological and structure data as well as to Dr. Hani Sitti Nuroniah, Dr. Marius R. M. Ekue and Dr. Nicolas-George H. Eliades for their willingness to support me in any way whenever needed.

The financial support of "Sophia Chlorou" endowment of National Technical University of Athens should be also acknowleged.

Lastly and most deeply I am grateful to my beloved mother and sister, who even though physically far, they supported me by all means throughout my study, and mainly through their strong, constant, and faithful love.

TABLE OF CONTENTS

1. Introduction 1
   Molecular evolution 1
   Loci-targets of selection 2
   Adaptation of forest trees 3
   Isocitrate dehydrogenase 4
      $NADP^+$ dependent isocitrate dehydrogenase 4
      NAD^+ dependent isocitrate dehydrogenase 6
   Quercus spp. – model species for adaptation 6
      Taxonomy and ecology 6
      Genetic variation and differentiation 8
   Aims of the study 10
2. Materials and methods 11
   Plant material 11
   Amplification of genomic DNA 11
   DNA cloning and sequencing 12
   SNP genotyping 12
   Genome walking 12
   Data analysis 13
      Sequence data 13
      SNP data 14
3. General results and discussion 15
   Conclusions and perspectives 20
4. Summary 21
5. Zusammenfassung 24
6. References 27
I. Patterns of nucleotide diversity and differentiation at $NADP^+$ and NAD^+ isocitrate dehydrogenases in a four-species sympatric white oak community 36
   Introduction 36
   Materials and Methods 37
      Plant material 37
      Identification of the $NADP^+$ and NAD^+ IDH sequences 38
      Amplification of genomic DNA 38
      DNA cloning and sequencing 40
      Genome walking 41
      Sequence analysis 42
   Results 43
      Identification of $NADP^+$ Isocitrate dehydrogenase 43
      Identification of NAD^+ isocitrate dehydrogenase 44
      Nucleotide diversity 45
         $NADP^+$ isocitrate dehydrogenase 45
         NAD^+ isocitrate dehydrogenase 46
      Species differentiation 49
      Statistical tests for neutrality 50
      Linkage disequilibrium 51
   Discussion 53

Nucleotide diversity and species differentiation....53
Neutrality tests....55
Linkage disequilibrium....56
Outlook....57
References....57
II. Gene-associated SNP analysis at a NADP+ specific IDH enzymegene in a four-species mixed oak forest....64
Introduction....64
Materials and Methods....66
Plant material....66
Identification of the $NADP^+$ IDH sequences....66
SNP identification and analysis....67
SNP genotyping....68
Results....70
Genetic variation within species....72
Genetic differentiation among species....73
Linkage disequilibrium and association analyses....75
Discussion....78
References....82
7. Appendix....89
Appendix 1a....89
Appendix 1b....90
Appendix 2a....92
Appendix 2b....93
Appendix 3....93
Appendix 4a....95
Appendix 4b....97
8. Curriculum vitae....98

LIST OF TABLES

Table I-1a: *Populus trichocarpa* specific primers and *Q. petraea* specific primers with corresponding T_a for the amplification of a $NADP^+$ IDH gene …….. 39

Table I-2: Gene specific primers used for the "Genome walking" technique…….. 42

Table I-3a: Summary statistics of nucleotide variation for the $NADP^+$ IDH gene……. 47

Table I-3b: Summary statistics of nucleotide variation for the NAD^+ IDH gene…… 48

Table I-4a: Among and within species AMOVA analysis from haplotypic data of the total $NADP^+$ IDH sequences…….. 49

Table I-4b: Among and within species AMOVA analysis from haplotypic data of the NAD^+ IDH sequences……. 49

Table I-5: Genetic differentiation among all species for each locus of $NADP^+$ IDH and for NAD^+ gene…….. 50

Table I-6: Tajima's *D* values for the neutrality test with $NADP^+$ and NAD^+ IDH gene sequences……. 50

Table II-1: Sequence, direction and substitution of SNP primers …….. 69

Table II-2: Position of the SNPs and amino acid - charge replacement caused by the non-synonymous SNPs …….. 71

Table II-3: Gene diversity H_s and inbreeding coefficient F_{is} per SNP and species…….. 72

Table II-4: Pairwise F_{st} values for the groups of SNPs: coding, non-coding, 73-synonymous, non-synonymous and overall loci ……... 74

Table II-5: Pairs of SNPs in linkage disequilibrium…… 77

Table II-6: Significant associations between SNPs and morphological traits……... 78

LIST OF FIGURES

Figure 1: Geographical distribution of *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* ……. 8

Figure I-1: Exon-intron organization and overlapping amplicons of the $NADP^+$ isocitrate dehydrogenase gene sequence identified in *Q. robur* ……... 44

Figure I-2: Scatterplot of the squared gametic frequency correlation against the nucleotide distance of all pairs of the parsimony informative sites for the $NADP^+$ IDH gene ……. 52

Figure I-3: Scatterplot of the squared gametic frequency correlation against the nucleotide distance of all pairs of the parsimony informative sites for the NAD^+ IDH gene ……. 52

Figure II-1: Schematic representation of the $NADP^+$ IDH gene …… 71

Figure II-2: Joint distribution of F_{st} values as a function of heterozygosity……... 75

Figure II-3: LD plot of pairwise r^2 values between all pairs of SNPs and corresponding p values ……. 76

LIST OF ABBREVIATIONS

AMOVA:	Analysis of molecular variance
BLAST:	Basic local alignment search tool
DNA:	Deoxyribonucleic acid
E.C.:	Enzyme Class (International Union of Biochemistry and Molecular Biology)
e.g.:	*exempli gratia*
EST:	Expressed sequence tag
GLM:	General linear model
IDH:	Isocitrate dehydrogenase
ITS:	Internal transcribed spacer
LD:	Linkage disequilibrium
NAD^+:	Nicotinamide adenine dinucleotide
$NADP^+$:	Nicotinamide adenine dinucleotide phosphate
PCR:	Polymerase chain reaction
QTL:	Quantitative trait loci
RNA:	Ribonucleic acid
SNP:	Single nucleotide polymorphisms
spp.:	*species pluralis*
SSR:	Simple sequence repeat

1. Introduction

Molecular evolution

Molecular genetic variation, either seen as allelic variation or as nucleotide sequence variation, allocates the fundamental characterization of genetic variation in a species. The explanation of the patterns of genetic variation in natural populations has been a considerable challenge for evolutionary biologists and geneticists. In particular, the contribution of natural selection in shaping the genetic variation of natural populations has been a matter of debate (Kreitman 1996; Nei 2005). As a result of the persisting need of understanding the process of evolution at its molecular level, as well as the relative roles of advantageous, neutral or deleterious mutations, many theories and models were developed, expressing mainly the scientific views of the so called " selectionists' " and the " neutralists' " schools of thoughts. Among those theories, the most thriving one is the neutral theory of molecular evolution. The above theory suggests that the genetic variation is primarily influenced by mutation generating it and genetic drift eliminating it (Kimura 1983). This hypothesis has been widely acting as a null model in molecular and evolutionary population genetics (Kimura 1977; Fay *et al.* 2002; Nei 2005).

Allozyme and protein polymorphisms have been in the past extensively applied, in studies for selection, aiming to address the question of neutrality concerning the molecular evolution, and the forces that maintain the large amounts of polymorphisms found in such loci (Eanes 1999; Hedrick 2005). With the argument that allozymes are mainly involved in the primary metabolic pathways, the electrophoretic investigation of their variation has been by many authors considered as a potential tool for the detection of positive selection in natural populations (Lewontin 1991; Wang *et al.* 1999; Ford 2002). The limitation of this method in obtaining the full information that causes differences in the electrophoretic mobility of the enzyme variants made the study of allozymes at the DNA sequence level more attractive and compulsory (Kreitman and Comeron 1999; Ford 2002; Nei 2005; Wheat *et al.* 2010).

DNA sequences contain the primary and most complete source of genetic information, providing the highest level of genetic resolution. DNA sequencing and genotyping methodologies based on it have introduced a new generation of research in genetics where

both quantitative and population genetic approaches are being applied to better understand the association between genotypic and phenotypic diversity, as well as the epigenetic effects that have been proven by empirical data to contribute to the phenotypic variation and fitness (Kalisz and Kramer 2008; Bonduriansky and Day 2009). The increasing interest in understanding these relationships has led the scientific attention to the detection of genes or regions of the genome, that are suspected to have been targeted by natural selection (Nielsen 2005; Stinchcombe and Hoekstra 2008). DNA sequence polymorphism data, together with the simulations of the coalescent theory can be an essential tool for inferring signatures of natural selection and identifying genes of adaptive significance (Rosenberg and Nordborg 2002; Nielsen 2005).

Loci-targets of selection

Patterns of genetic population structure have been suggested to be a useful tool in identifying loci that are under selection (Lewontin and Krakauer 1973). That is because loci involved in local adaptation should show higher levels of differentiation among populations compared to those that evolve neutrally, and have their allelic frequencies mainly determined by genetic drift alone. This idea in combination with improved statistical methodologies has been for many studies the hypothesis for identifying loci upon which positive selection acts (Akey *et al.* 2002; Scotti-Saintagne *et al.* 2004; Anderson *et al.* 2005; Stinchcombe and Hoekstra 2008; Derory *et al.* 2010). However, it has to be mentioned, that the distribution of estimates of the fixation index F_{st} being a useful way of summarizing genetic variability among populations is highly dependent on demographic history. Still, recent theoretical results argue that the approach should be generally robust to demographic effects (Beaumont and Nichols 1996; Beaumont 2005). An important application of this approach is the genetic mapping of marker loci that exhibit significant differences in the distribution of F_{st} estimates (Beaumont 2005).

Several tests of selective neutrality have been developed based on similar approaches of allelic distribution or levels of variability, either for one or for multiple loci (Nielsen 2001). One of the most common tests for nucleotide data is Tajima's *D*-test (Tajima 1989), based on the number of pairwise differences and the number of segregating sites in a sample of nucleotide sequences. Other similar tests, widely applied as tests for neutrality are *D* and *F* tests of Fu and Li (Fu and Li 1993) based on differences between the number o singletons

and the total number of mutations or the average number of nucleotide differences between pairs of sequences respectively. Or else, the test of Fay and Wu (Fay and Wu 2000) based on the average number of nucleotide differences between pairs of sequences, and the frequency of the derived variants.

Moreover, variability within and between species at multiple loci can also be informative concerning selectively neutral loci. A test widely applied using this kind information is the HKA test (Hudson *et al.* 1987), based on the idea that the expected number of segregating sites within species and the expected number of fixed differences among species are proportional to the mutation rate and their ratio should be constant among loci, given absence of selection.

Finally, another approach, that can give more direct information regarding the question of selection acting on specific loci, is the comparison of patterns of substitution among synonymous and non-synonymous sites. In this approach the advantage of analyzing sequences of genomic DNA or genetic markers such as SNPs (single nucleotide polymorphisms) derived from genomic sequences can be seen. In such markers, not only each substitution can be characterized as coding or non-coding, but also more specifically it can be determined as synonymous or non-synonymous (not causing or causing an amino-acid substitution respectively). On this kind of comparisons, several tests of selective neutrality have been developed. Among the most widely used is the McDonald-Kreitman test (Mcdonald and Kreitman 1991), where the ratio of non-synonymous to synonymous polymorphism between species is compared to the ratio of the number of non-synonymous and synonymous fixed differences between species, or else, the d_N/d_S test and its derivatives (Goldman and Yang 1994), that directly measures the rate of substitutions at silent sites to the rate of substitutions at non silent sites.

Adaptation of forest trees

The significance of identifying and analyzing genes of adaptive potential, gains a very special meaning when it comes to forest ecosystems, considering the rapidly changing environmental conditions. Forest trees are long-lived organisms with long generation times, which make their neutral molecular evolution slow (Savolainen *et al.* 2007; Kuparinen *et al.* 2010). Nevertheless, their relatively high levels of genetic diversity and gene flow foster their ability to adapt to climatic or environmental changes (Petit and Hampe 2006; Aitken *et al.*

2008), as the genetic variation is a requirement for any potential evolutionary adaptability (Finkeldey and Mátyás 2000).

Technical limitations, large genome sizes and long generation times make the use of methods such as gene manipulation, insertion mutagenesis or positional cloning for the identification of adaptive genes difficult or even impossible for forest trees. Over the past two decades, the rapid evolution of new molecular genetic technologies has placed at the disposal of forest geneticists some additional tools for the identification of adaptive variation; QTL analysis tools, "candidate gene" approaches, whole genome scan approaches less frequently for tree species, comparative, structural and functional genomics combined with population genetics of genes controlling adaptive traits in trees, provide a new aspect of forest trees being a model experimental system for studying the relationships between natural occurring genotypic and phenotypic variation (González-Martínez *et al.* 2006; Neale and Ingvarsson 2008; Gailing *et al.* 2009; Hall *et al.* 2010).

The complete genomic sequences of model plant species such as *Arabidopsis thaliana* or *Populus trichocarpa* and large databases of ESTs (expressed sequence tags) of model forest tree species such as *Picea abies*, *Pinus taeda* and *Eucalyptus* spp., as well as their constantly updated annotation status have increased the amount of the available sequence information, thus making the detection of the putatively important genes or regulatory regions easier to obtain (Gailing *et al.* 2009). Only based on database information, recent studies on forest trees have for example been based on analyzing variation in genes that are involved in wood formation in pines (*Pinus* spp.) (Pot *et al.* 2005), on estimating nucleotide diversity at the *pal*1 locus (a key enzyme of the secondary metabolism of higher plants) in *Pinus sylvestris* (Dvornyk *et al.* 2002), on assessing linkage mapping of osmotic stress induced ESTs of oaks (Porth *et al.* 2005), on analyzing patterns of nucleotide polymorphism and linkage disequilibrium of genes that are considered adaptive within and among natural populations of European aspen (*Populus tremula*) (Ingvarsson 2005) or on investigating the nucleotide diversity compared to the variation of bud burst in candidate genes of oak populations (Derory *et al.* 2010).

Isocitrate dehydrogenase

$NADP^+$ dependent isocitrate dehydrogenase

$NADP^+$ dependent isocitrate dehydrogenases (NADP-IDH, E.C. 1.1.1.42) belong to a multi-isoenzymatic family. Its members are homodimeric enzymes. They can be located

subcellularly in the cytosol (Fieuw *et al.* 1995), plastids (Gálvez *et al.* 1994), mitochondria and peroxisomes (Corpas *et al.* 1999). The cytosolic activity was found to be predominant compared to the total enzymatic activity in aerial plant parts (Chen *et al.* 1988; Hodges *et al.* 2003) and the only detectable location of activity in investigated gymnosperms (Palomo *et al.* 1998). The exact physiological role of the isozymes in plant metabolism is still obscure.

From a biochemical point of view, $NADP^+$ dependent isocitrate dehydrogenase participates in the Krebs cycle. It has been suggested, that the cytosolic class of $NADP^+$ IDH enzymes in plants could be responsible for the production of 2-oxoglutarate, necessary for glutamate synthesis and for ammonia assimilation, crucial for the production of all essential amino acids, but this hypothesis is still under discussion (Palomo *et al.* 1998; Hodges *et al.* 2003).

The amino acid sequence of $NADP^+$ dependent isocitrate dehydrogenase is highly conserved not only across plant species (Chen *et al.* 1988; Pascual *et al.* 2008b) but even across species of distinct kingdoms (eukaryotes or prokaryotes) (Fieuw *et al.* 1995; Sadka *et al.* 2000). In general, the evolution of $NADP^+$ dependent isocitrate dehydrogenase isozymes seems to have arisen by independent gene duplications in animals, fungi (Nekrutenko *et al.* 1998), and higher plants (Hodges *et al.* 2003), suggesting that they play different roles in different *phylae* (Pascual *et al.* 2008b). Examples of the significance of the $NADP^+$ IDH genes in different species are the differences in the kinetic performance of the enzyme across thermal environments in cricket (*Allonemobius socius*), (Huestis *et al.* 2009), the up-regulation of $NADP^+$ isocitrate dehydrogenase in poplar (*Populus tremula* x *P. alba*) after treatment with PPT (phosphinothricin, a common herbicide used in agriculture) in a transgenic PPT-resistance background (Pascual *et al.* 2008a), the 2-fold enhancement of the expression of the gene in ectomycorrhizal roots compared to nonmycorrhizal roots of *Eucalyptus globulus* (Boiffin *et al.* 1998) or in a recent study the association of a mutative $NADP^+$ IDH-1 gene with gliomas tumor in humans (Dang *et al.* 2009).

In terms of electrophoretic isozyme separation, known as zymogramms, the $NADP^+$ isocitrate dehydrogenase system has been widely applied among other isozymes for forest tree population genetic analyses. In some cases, it has been suggested that isocitrate dehydrogenase has a potentially adaptive role. In *Abies alba*, isozyme polymorphism in the IDH system has been argued to be a result of adaptation to high temperature regimes (Bergmann and Gregorius 1993), whereas in the case of *Fagus sylvatica* a correlation of specific alleles with the extend of beech scale insect (*Cryptococcus fagisuga*) infestation has been

recorded (Ziehe 1996a; Ziehe 1996b). Specifically, in *Quercus spp.*, IDH has been analyzed in two encoding loci, IDH-A and IDH-B. The latter, has exhibited strong differentiation patterns, between the species *Q. robur* and *Q. petraea* (Finkeldey 2001; Gömöry *et al.* 2001; Scotti-Saintagne *et al.* 2004) or *Q. robur* against *Q. frainetto* and *Q. pubescens* (Curtu *et al.* 2007a).

NAD^+ dependent isocitrate dehydrogenase

NAD^+ dependent isocitrate dehydrogenase (NAD-IDH, E.C. 1.1.1.41) is also connected to the Krebs cycle but as opposed to the $NADP^+$ dependent isozyme, it is strictly mitochondrial (Behal and Oliver 1998; Lemaitre and Hodges 2006). The NAD^+ dependent isocitrate dehydrogenase has not been widely investigated in plants, and thus, its best characterization is that from yeast (*Saccharomyces cerevisiae*) (Keys and McAlisterhenn 1990). Furthermore, in *Arabidopsis thaliana* the introns-exons organization and the deduced protein sequences of the six putative NAD^+-IDH encoding genes revealed two different types of NAD^+-IDH subunits: a catalytic and a regulatory type (Lemaitre and Hodges 2006). Recently, in *Zea mays* ssp. *mays*, an amino acid substitution was identified, located in a phylogenetically conserved region of a NAD^+-IDH isoform, which was strongly correlated to temperature-dependent enzymatic activity. Additionally, the authors argued that the LD (linkage disequilibrium) state of the same non-synonymous SNP with another SNP located on the promoter of the gene, suggests that the first SNP could be related to the IDH protein expression efficiency (Zhang *et al.* 2010). Yet, the exact physiological role of the isoforms of the NAD^+-IDH genes are still not known.

The NAD^+ dependent IDH system has not been used for electrophoretic separation in population genetic studies of oaks, due to its low enzyme signal intensities in zymogramms, resulting in non-reliable scoring of the electrophoretic phenotypes.

Quercus spp. – model species for adaptation

Taxonomy and ecology

The ecological significance of the genus *Quercus* has been widely recognized and acknowledged. Therefore, its taxonomic classification has been the centre of interest for several authors as well as an issue of controversy. One of the most accepted classifications of the genus, based on morphological traits, attributes to it four sub-genera: *Erythrobalanus* (most of the taxa distributed in northern and central America), *Sclerophyllodeys*, *Cerris* and

Lepidobalanus including 11 species, among them *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* (Schwarz 1993). The above species, were as well treated as closely related, by being classified in a separate section; section *Robur*, of the sub-genus *Lepidobalanus* in a slightly different classification (Hedi 1981). An infrageneric and sectional classification, that claims to be consistent with phylogenetic analyses, suggests only two sub-genera: *Cyclobalanopsis*, and *Quercus*, attributing to the last one, three sections: *Lobate*, *Protobalanus*, and *Quercus* in a broad sense (Nixon 1993). In the present work, the investigated species; *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* belong taxonomically to the section *Quercus* (Nixon 1993) also referred to as "white oaks".

With the latter classification agree also several phylogenetic analyses, based on molecular tools, such as the sequence comparison of the ITS1 and ITS2 regions of the 5.8S RNA encoding ribosomal DNA (Bellarosa *et al.* 2005) or/and chloroplast DNA markers (Manos *et al.* 1999). Surprisingly, a phylogenetic analysis where the same molecular tools were applied, revealed "unexpected" phylogenetic relationships between different *Quercus* taxa (Samuel *et al.* 1998). This was finally accounted for an overlooked incorporation in the analysis of the sequences of a paralogue ITS locus (Mayol and Rosselló 2001).

The ecology of the different taxa of the genus, even of the most closely related "white" oaks of the section *Quercus* (Nixon 1993) is distinct. In particular, *Q. robur* can cope better than *Q. petraea* with hard winters and late frosts (Ellenberg 1988). However, the sufficient mineral and water conditions have been proven more important for the growth of *Q. robur* than that of *Q. petraea* with the latter being even sensitive in non well drained to wet soils (Levy *et al.* 1992; Aas 2006a; Aas 2006b). On the contrary, *Q. pubescens* has less demanding preferences. It can be found in many different habitats in terms of climatic and edaphic characteristics (Bussotti 2006). Still, it is mainly found in dry warm habitats, and reaches its optimum photosynthesis at relatively high temperatures, thus can be characterized as thermophilous and xerothermic (Ellenberg 1988). As for its edaphic preferences, it can tolerate different types of soil, even the poor ones (Bussotti 2006). *Q. frainetto* is also a thermophilous species, and prefers habitats with long warm summers and mild winters. It can withstand long dry periods but it is sensitive to late frosts. In terms of soil preferences, it is less demanding (Bartha 2006). Despite their differences, the above mentioned species share at young stages the need for light (Ellenberg 1988). In terms of geographical distribution of the four species, *Q. robur* and *Q. petraea* are the most widely distributed in Europe covering northern and

central Europe, *Q. pubescens* has a more south distribution, whereas *Q. frainetto* is more limited mainly to the Balkan Peninsula (Figure 1) (Jalas and Suominen 1976; Aas 2006a; Aas 2006b; Bartha 2006; Bussotti 2006).

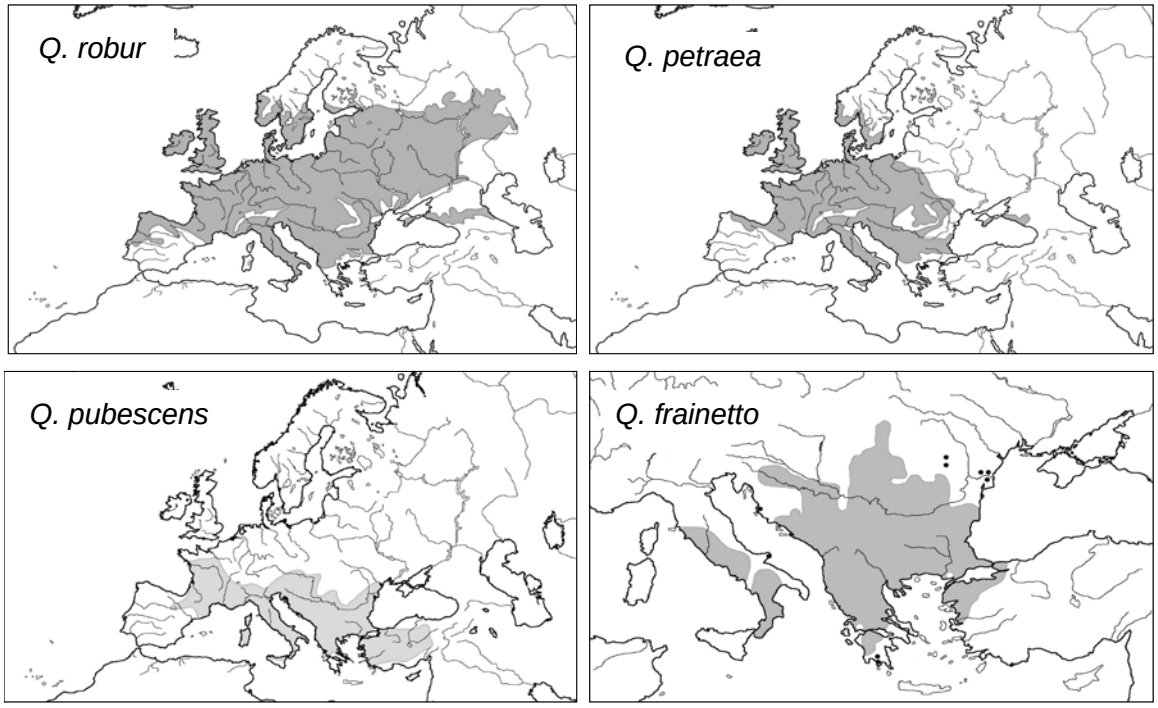

Figure 1: Geographical distribution of *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* (after Aas 2006a, Aas 2006b, Bartha 2006, Bussotti 2006, respectively)

Genetic variation and differentiation

Oaks are monoecious species, predominantly outcrossing, wind pollinated, and considered by some authors to own an incompatibility system (Ducousso *et al.* 1993) though its existence or effectiveness have been doubted by findings of selfing rates of about 2% or in some cases even above 3% (Sork *et al.* 2002; Fernandez and Sork 2005; Chybicki and Burczyk 2010; Jensen *et al.* 2009). In general, oaks are considered to be among the most diverse species of forest trees. High levels of diversity are most likely due to the maintenance of large population sizes, long distance gene flow and interfertility (Ducousso *et al.* 1993; Streiff *et al.* 1999; Petit *et al.* 2004). Hybridization in oaks is a well studied mechanism among several taxa of the genus, and it might have played a role in their postglacial re-colonization (Petit *et al.* 2004). It seems to occur predominantly among closely related species (Bacilieri *et al.* 1996; Petit *et al.* 2004; Curtu *et al.* 2009) and is strongly dependent on their relative

abundance (Salvini *et al.* 2009). However, oak species seem to remain distinct, even within areas of sympatry (Craft and Ashley 2006; Curtu *et al.* 2007a). The two more frequently investigated European oaks; *Q. robur* and *Q. pertraea* are interfertile and co-exist in most European forests despite their different ecological preferences and their succession efficiency (Petit *et al.* 2004). In chloroplast DNA studies it has been shown that they share the most frequent haplotypes (Petit *et al.* 1997; Kremer and Goenaga 2002; Petit *et al.* 2002a; Petit *et al.* 2004). Furthermore, low correlation has been found between the cpDNA and the nuclear variation between those two species (Kremer *et al.* 2002; Finkeldey and Mátyás 2003). In general, apart from a few exceptions (Finkeldey 2001; Gömöry *et al.* 2001; Scotti-Saintagne *et al.* 2004; Muir and Schlotterer 2005; Neophytou *et al.* 2010) the two species are weak differentiated. Concerning the differentiation of the other closely related species, although they have been much less investigated, *Q. frainetto* seems to be genetically more similar to *Q. pubescens* than to *Q. robur* or *Q. petraea* (Curtu *et al.* 2007a), whereas *Q. pubescens* exhibits low levels of genetic differentiation with *Q. petraea* and high levels of genetic admixture in mixed or pure stands of the two species (Salvini *et al.* 2009).

With their high levels of genetic variation and their wide distribution across Europe that often becomes sympatric, despite their different climatic and edaphic optima, oak species recently became model species to study the adaptation of trees in changing environments, and to detect signatures of natural selection. The species differentiation, even in sympatry, can be seen as population differentiation in different environmental circumstances, and can be maintained by diversifying selection acting upon them in different directions (Le Corre and Kremer 2003; Petit *et al.* 2004).

The increasing availability of genomic information about the species has made the construction of QTL maps of potentially adaptive traits feasible (Scotti-Saintagne *et al.* 2004; Porth *et al.* 2005; Casasoli *et al.* 2006; Parelle *et al.* 2007; Gailing 2008; Gailing *et al.* 2008). However, to date, very few studies reported the nucleotide diversity of genes in oaks, among which only one dealed with European species (Quang *et al.* 2008; Quang *et al.* 2009; Derory *et al.* 2010). The levels of nucleotide variation reported in the above mentioned studies are higher than that reported in *Pinus* or other conifers (Dvornyk *et al.* 2002; Pot *et al.* 2005; González-Martínez *et al.* 2006; Savolainen and Pyhajarvi 2007; Eveno *et al.* 2008) but lower than those reported for another forest angiosperm, *Populus tremula* (Ingvarsson 2005).

Aims of the study

The aims of the present study were:

- to identify the sequence of $NADP^+$ and NAD^+ isocitrate dehydrogenases for *Quercus* species
- to examine the patterns of nucleotide diversity and differentiation within and among the four sympatric, closely related *Quercus* species at these genes and to test the fit of the diversity patterns to neutrality models
- to develop SNP markers in coding and non coding regions throughout the identified $NADP^+$ IDH gene
- to estimate levels of genetic variation and differentiation of the sympatric oak species using the developed SNP markers
- to test for possible nonrandom associations between the alleles of the SNP markers as well as between the variation of the SNPs and leaf morphological traits

2. Materials and methods

Plant material

253 oak individuals (*Q. robur*, *Q. petraea*, *Q. pubescens*, *and Q. frainetto*, all considered as *sensu lato*) were investigated in the present study. The sampling was exhaustive as described in Curtu et al. (2007a) at the Bejan forest located in central-western Romania where the four species naturally co-exist.

Morphological assignment

The morphological assignment of the individual trees to each species group based on morphological traits of leaf size and pubescence was conducted and described in detail by Curtu *et al.* (2007a).

Genomic DNA

DNA was extracted from buds, using the Qiagen Dneasy96 Plant Kit (Qiagen, Hilden, Germany) and following the manufacturer's protocol as described by Curtu *et al.* (2007a).

Amplification of genomic DNA

To obtain the sequence of a cytosolic $NADP^+$ dependent IDH gene PCR reactions were performed using DNA template of a *Q. petraea* sample with oligo-primers that were designed in three overlapping parts on the basis of the public available sequences of *Populus trichocarpa* (Tuskan *et al.* 2006) (Manuscript I). Based on the derived sequence, primers were re-designed for specificity to *Quercus* spp. For sequencing purposes the PCR primers were designed to amplify the gene in seven overlapping fragment as described in Manuscript I. For the identification of the NAD^+ dependent IDH gene the technique of "genome walking" (Siebert *et al.* 1995) was applied on a sample of *Q. robur*. The method is described briefly below and in detail in Manuscript I. The amplification of the partially obtained gene was conducted in a single fragment.

For both genes primers were designed using the web based software Primer3 (Rozen and Skaletsky 2000). The quality of the primers was checked in terms of melting temperatures, self-dimerization and/or formation of hairpins with the software GeneRunner® (1994, Hastings Software, Inc.). PCR amplifications were carried out, by applying Qiagens's HotStarTaq® MasterMix. The amplification of the target DNA fragments, was checked on 1.5% agarose gel, and for further treatment, the PCR products were excised and purified

from the gel. Detailed protocol regarding the PCR concentrations and conditions can be found in Manuscript I.

DNA cloning and sequencing

DNA sequences were obtained both with direct sequencing and after cloning for different purposes. Direct sequencing was performed in order to verify the amplified target DNA. After verification of the DNA amplification by BLAST search against public databases, the PCR amplicons were cloned into the pCR®2.1-TOPO® vector, using a TA-cloning kit from Invitrogen (Carlsbad, CA). Clones were cultured overnight and the plasmid DNA was extracted using the QuickLyse Miniprep Kit (Qiagen, Hilden, Germany) and sequenced in both strands, using BigDye chemistry (Applied Biosystems, Foster City, CA) on an ABI Prism® 3100 genetic analyzer (Applied Biosystems). Details on the methods and protocols of cloning and sequencing are provided in Manuscript I.

SNP genotyping

SNP genotyping was performed applying the "minisequencing" single nucleotide primer extension method (Pastinen *et al.* 1997) using an ABI Prism® SNaPshot™ Multiplex Kit (Applied Biosystems Foster City, CA) and following the manufacturer's protocol, with modifications. The SNP primers were designed and checked for their properties with the software Primer3 (Rozen and Skaletsky 2000) and GeneRunner® (1994, Hastings Software, Inc.) respectively. The PCR fragments of the $NADP^+$ dependent IDH gene were pooled to be used as template for the primer-extension reactions. The corresponding SNP primers were pooled for multiplex primer-extensions generating maximal seven SNPs per reaction. The SNP primers had 4-8bp difference in size which was controlled by an addition of a poly$(T)_n$ tail at their 3' end. Detailed protocol and the sequences of the SNP primers are provided in Manuscript II.

Genome walking

In order to obtain the sequence of NAD^+ dependent IDH gene, we applied the PCR based method "primer walking" for identifying an unknown DNA sequence adjacent to a known sequence, (Siebert *et al.* 1995), using the Genome Walker™ Universal Kit (Clontech Laboratories, Inc.). This technique requires that the starting template DNA is of high quality, and therefore the DNA quality and quantity were tested against λ-DNA. The tested genomic

DNA was used as a template for the construction of adaptor-ligated "libraries". Each "library" was applied to a "suppression" PCR and followed by a "nested" PCR. For these PCRs gene specific primers were used, specially designed to locate in consecutive positions on the known sequence published in the INRA *Quercus* EST database accessible via the EVOLTREE organization. The gene specific primers were designed, using the web based software Primer3 (Rozen and Skaletsky 2000), and their quality was checked in terms of melting temperatures, self-dimerization and/or formation of hairpins with the software GeneRunner® (1994, Hastings Software, Inc.). The detailed procedure is described in Manuscript I.

Data analysis

The genomic DNA sequences that were used for this study were visually verified and manually edited to check for base-calling errors. Contigs were assembled using the computer software CodonCode Aligner (CodonCode Corp., Dedham, MA), and Sequencher v4.8 (Gene Codes, Ann Arbor, MI). Multiple sequence alignments were performed using the ClustalW algorithm (Thompson *et al.* 1994) and adjusted manually using BioEdit (Hall 1999). The sequences were aligned against public databases using tBLASTx or BLASTx in case of nucleotide or pBLAST in case of amino-acid query. Methods of sequence verifications and alignments are described in Manuscript I.

Sequence data

The sequences of both $NADP^+$ and NAD^+ isocitrate dehydrogenases were analysed for descriptive statistics of nucleotide polymorphism π (Nei 1987) and diversity as well as Tajima's *D* (Tajima 1989) statistic for testing neutrality. Species' differentiation was estimated from haplotypic data as F_{st} (Hudson *et al.* 1992a) and G_{st} (Nei 1973) supplemented by the nucleotide based statistics S_{nn} (the nearest-neighbor statistic) (Hudson 2000) and K_{st}*, a weighted measure of the ratio of the average pairwise differences within populations to the total average differences (Hudson *et al.* 1992b). Calculations of nucleotide diversity and differentiation were performed with the software DNAsp (Librado and Rozas 2009). The analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992) from haplotypic data was conducted with the software Arlequin v3.5 (Excoffier and Lischer 2010). Detailed description of the analysis of the sequence data is provided in Manuscript I.

SNP data

The SNP data set was analyzed in terms of gene diversity (H_s) and inbreeding coefficient F_{is} within species with the software FSTAT v2.9.3.2. (Goudet 1995). Species differentiation in terms of pairwise F_{st} between all pairs of species was analyzed with the software package Arlequin v3.5 (Excoffier and Lischer 2010). Linkage disequilibrium between all pairs of SNPs for each species separately was analyzed performing the likelihood-ratio test using an Expectation-Maximization (EM) algorithm as implemented in the software package Arlequin v3.5 (Excoffier and Lischer 2010). LD was also analyzed as the squared correlation between all possible combinations of alleles (r^2) for the total data set, with the software TASSEL v2.1 (Bradbury *et al.* 2007). Association analysis of all investigated SNPs with morphological traits that best differentiated the species (Curtu *et al.* 2007b) using F-Tests under a general linear model (GLM) was conducted also using TASSEL v2.1 (Bradbury *et al.* 2007) by incorporating in the analysis data of genetic structure to avoid biased associations. The analysis of the SNP genotyping data is described in detail in Manuscript II.

3. General results and discussion

The present study is one of the first studies to describe nucleotide diversity in European oak populations. It has been focused on the almost complete sequence of a $NADP^+$ dependent isocitrate dehydrogenase and a partial sequence of an NAD^+ dependent isocitrate dehydrogenase gene. The obtained sequences were confirmed with highly scored homologies with the corresponding genes of different organisms using Basic Local Alignment Search Tools (BLAST) against public nucleotide and protein databases.

The $NADP^+$ IDH gene was analyzed in only six overlapping amplicons (here referred to as loci) distributed throughout the identified sequence of 3481bp (see Figure I-1). One locus was excluded from the analysis because of its high variation, in order to avoid the incorporation of paralogous loci in the data set. Within the obtained sequence 1195bp were identified as putatively coding regions. Patterns of nucleotide diversity varied among the different amplicons analyzed. For the pooled data (pooled sequences of all species) the levels of nucleotide diversity were between 0.00457 and 0.01368 for locus 4 and locus 1, respectively. In particular, all species exhibited high levels of nucleotide diversity, with *Q. frainetto* showing the highest levels (π_{tot}=0.00901) among all species. This values are comparable (though slightly higher) to the reported nucleotide variation of the Asian *Q. mongolica* var. *crispula* (Quang *et al.* 2009), to that examined in the very first study that characterized nucleotide variation levels within the genus *Quercus*, *Q. crispula* (Quang *et al.* 2008) and also to the ones reported for *Q. petraea* in investigations of candidate genes for adaptive traits (Vornam *et al.* 2007) such as the timing of bud burst (Derory *et al.* 2010). Locus 4 shows the lowest levels of diversity over all species. This might be explained if the active site of the enzyme is located within this locus, as it is suggested e.g. by the active site of the highly homologous corresponding enzyme sequence of the psychrophilic bacterium, *Desulfotalea psychrophila* (Fedøy *et al.* 2007). Similar levels of nucleotide diversity were found at the NAD^+ IDH gene in general, with *Q. robur* exhibiting the lowest and *Q. frainetto* the highest values of π_{tot}. From the total variation found among the sequences, the highest percentage was within species (98% and 96% for $NADP^+$ IDH and NAD^+ IDH gene, respectively) after analysis of molecular variance (AMOVA). The genetic differentiation among species was in general low for both genes, confirming previous results that reported low levels of differentiation among closely related oak species in the majority of the markers

applied, due to homogenizing gene flow (Bodénès *et al.* 1997). On the contrary, loci that exhibit high levels of genetic differentiation among closely related oak species have been claimed to be responsible for species divergence as a consequence of diversifying selection acting upon them (Wu 2001; Petit *et al.* 2004; Scotti-Saintagne *et al.* 2004; Lexer and Widmer 2008). Locus 4 of the $NADP^+$ IDH gene, the locus with the lowest levels of genetic diversity, was the only locus - considering the two genes - that revealed a significant differentiation among species both with the nearest neighbor statistic significance test (S_{nn}) and K_{st}*, a weighted measure of the ratio of the average differences within populations to the total average differences. The high levels of differentiation in this locus combined with the low levels of genetic diversity and given the possibility that the active site of the gene might be located in this locus, makes the question of selection acting on it even more ground. However, the low sample size analysed should be taken into consideration before driving any conclusion.

The Tajima's *D* statistic calculated to test the fit of nucleotide polymorphisms observed at both genes to the neutral mutation model was found non-significant both for each single species and for the pooled sample set. However, $NADP^+$ IDH gene in almost all species and loci showed negative measurements of Tajima's *D* suggesting excess of low-frequency polymorphisms. *Q. frainetto* exhibited the lowest *D* value for the total sequence (the sequence made of all six loci). Interestingly, the patterns of Tajima's *D* species-wide were not consistent over the different loci, implying that they could not be interpreted by demography (Moeller *et al.* 2007). In particular, at locus 4 of the $NADP^+$ IDH gene, *Q. petraea* and *Q. pubescens*, exhibited very low *D* values, close to the 95% significance threshold of the computed coalescent simulations. This suggests a greater excess of low-frequency variants, despite the exclusion of singletons (single mutations) for the accuracy of the analysis. This result combined with the low gene diversity at this locus and the highly significant inter-specific differentiation points out the putative adaptive significance of the gene. Moreover, the ratio of non-synonymous to silent polymorphisms was found over all species in average 0.15 for $NADP^+$ IDH and 0.38 for NAD^+ IDH, suggesting stronger purifying selection acting on the first gene, as it would be expected for most housekeeping proteins or enzyme genes. The average estimation of the non-random associations between the pairwise polymorphisms (ZnS) showed in general high levels but different trends for the two genes among the different species. The analysis could not be conducted for each locus of the

$NADP^+$ IDH gene separately, because of the insufficient number of pairwise comparisons within each locus. Because of low sample size the estimation of LD at NAD^+ IDH gene for *Q. petraea* was not possible for the same reason. The low sample size might also explain the inconsistent patterns of the ZnS measurements for the two genes over the different species. Similar estimates of ZnS in a fragmented forest of *Q. mongolica* var. *crispula* were attributed to the possible recent bottleneck of the population due to its isolation rather than genetic admixture that also contributes to higher estimates of ZnS (Quang *et al.* 2009). For the present study the second explanation would be more suitable given the sympatry of the four species and their strong genetic admixture through hybridization (Curtu *et al.* 2009).

The SNP analysis of 13 SNPs distributed throughout the $NADP^+$ IDH gene revealed similar patterns of gene diversity for all species. In particular, *Q. frainetto* was found slightly more variable, confirming also the results of nucleotide diversity at both $NADP^+$ and NAD^+ IDH gene, but not the results of isozyme markers applied on the same individuals, that on average showed *Q, frainetto* as being the less variable species among the four (Curtu *et al.* 2007a). However, in that study, the finding of *Q. frainetto* as the less variable was only based on the mean gene diversity values, since single marker specific isozymes and/or nuclear SSR markers showed *Q. frainetto* equally or in cases even higher variable than the other species. When comparing the most widely distributed *Q. robur* and *Q. petraea* in Europe, similar to other studies of mixed or pure stands (Gömöry *et al.* 2001; Mariette *et al.* 2002; Finkeldey and Mátyás 2003) this study showed higher variability for the latter species. In total, the levels of gene diversity were found higher at the non coding SNPs than those found at the coding SNPs. Consistent with previous isozyme and nuclear SSR data (Gömöry *et al.* 2001; Finkeldey and Mátyás 2003; Curtu *et al.* 2007a; Neophytou *et al.* 2010), the estimation of the inbreeding coefficient F_{is} was found slightly negative or close to zero. An exception was the case of *Q. robur*, where F_{is} was found slightly above zero. The same pattern of F_{is} was also found for the same sample set after IDH-B isozyme analysis. In general, the slightly higher average F_{is} of *Q. robur* as compared to those observed in *Q. petraea* is in accordance with previous studies, yet in this study F_{is} for *Q. robur* did not significantly deviate from Hardy - Weinberg expectations.

Among the four species, the levels of pairwise genetic differentiation were low but significant for almost all pairs except for *Q. frainetto* and *Q. petraea*. Significant differentiation was also found for the same individuals by the use of nuclear SSR and isozyme markers

(Curtu *et al.* 2007a). In almost all groups of SNP markers (coding, non-coding, synonymous) the present results confirm previous findings that suggest *Q. robur* and *Q. frainetto* as being the most differentiated among the four species (Schwarz 1993; Curtu *et al.* 2007a). There was an interesting exception observed in the group of non-synonymous SNPs, at which the highest differentiation was found between *Q. robur* and *Q. petraea*, followed by the pair of *Q. robur* and *Q. pubescens*. The non-synonymous SNPs alone failed to differentiate significantly between *Q. frainetto* and all other species. On the contrary, the synonymous SNPs revealed the strongest differentiation levels not only for *Q. frainetto* but also for the other species. This is an effect of the synonymous SNP 6 which was found to exhibit outlier behavior compared to the null F_{st} distributions simulated under a neutral model. The above method is widely applied to detect loci under selection using data of genome scans (Beaumont and Nichols 1996; Excoffier *et al.* 2009). Hence, it is very sensitive to the number of loci used for the analysis, and shows less power in cases of low sample sizes (Eveno *et al.* 2008). Therefore care should be taken with the interpretation of the results of the F_{st} outlier analysis in the present study using 13 SNP markers within one gene. As the resulted outlier SNP 6 is a synonymous SNP, it is not likely that it is directly under the action of natural selection. However, this SNP might be linked to an adaptive polymorphic site and covariate with it through the hitchhiking effect of selective sweeps or strong linkage (Palme *et al.* 2008; Eckert *et al.* 2009).

All the possible SNP combinations were tested for pairwise non-random associations for the total sample. The analysis suggested significant non-random associations in 39.7% of all possible comparisons. This percentage is higher than that found in other studies (for instance Ingvarsson *et al.* 2008), possibly because of the location of all SNPs in one gene. However, there was no particular clustering observed among the sites in significant LD according to their physical proximity. Moreover, the levels of LD were low, as expected for a predominantly outcrossing species (Neale and Savolainen 2004; Zhang and Zhang 2005; Ingvarsson *et al.* 2008). In the LD analysis for each species separately, *Q. robur* revealed the lowest number of sites in significant LD. On the contrary, *Q. frainetto* exhibits the largest number of sites in significant LD reflecting a possible founder effect due to its marginal location regarding its geographic distribution (Bartha 2006). The smaller number if sites in significant LD for *Q. robur* might be a result of the larger effective population size of the species at that location. A larger effective population size was also the explanation that was

given for the lower LD levels of northern populations of a *Quercus* species in Japan compared to the southern ones that underwent a recent population bottleneck (Quang *et al.* 2008). Yet, in the same study the explanation of natural selection acting on the population that showed higher LD levels was not completely excluded, since slower decaying LD can make hitchhiking and background selection more effective (Charlesworth *et al.* 1993).

The association analysis of all 13 SNPs with traits of leaf morphology that successfully distinguished the four oak species in an earlier study that investigated the same individuals (Curtu *et al.* 2007a), showed seven significant associations. For two of the four morphological traits included in the analysis multiple associations were found. Two of the SNPs (SNP 1 and SNP 2) were associated with sinus width, whereas three others (SNP 4, SNP 5 and SNP7) were associated with lamina length. Among them, SNP 1, SNP 2 and SNP 4 are non-synonymous SNPs. For all significant associations, the percentage of the variation described by the corresponding SNPs was low (R^2 <4%). Low R^2 values have been argued to give associations of low power (Ingvarsson *et al.* 2008; Simko *et al.* 2009). However, in an association study of SNPs derived from candidate genes for wood property traits in *Pinus taeda*, similar R^2 percentages with those estimated in the present study were proven to be still significant after false discovery rate (FDR) corrections (González-Martínez *et al.* 2007). In general, association studies are subject to factors that influence their robustness such as population structure, sample size, number and type of markers analyzed, accurate genotyping and phenotyping methods (Hall *et al.* 2010). It has been also shown that in such types of analyses greater power can be achieved by increasing the sample size of the populations analyzed than by increasing the number of polymorphisms (markers) (Long and Langley 1999). Moreover, association studies are well known to be strongly biased by population structure (Zhao *et al.* 2007). In the present study, although the total number of samples analyzed was relatively large (253 individuals), this can not compensate the fact that the four species were strongly differentiated also by the use of unlinked neutral markers (Curtu *et al.* 2007a). However, the bias effect of the population structure in our sample was minimized by taking into account genetic structure data when the association analysis was conducted. Additionally, it is obvious that if diversifying selection maintains the differentiation among the species as it has been suggested (see Introduction), its signatures are difficult to distinguish from the pure demographic effects, and in that case by taking into

account genetic structure data in association studies some biological associations might even be lost.

Conclusions and perspectives

It has been shown that by examining the almost complete sequence of the $NADP^+$ IDH gene different patterns of nucleotide diversity and of inter-specific differentiation can be observed. Therefore the present study points out the significance of analyzing complete gene sequences in the investigation of candidate genes, if a conclusion about the signatures of selection acting upon them is to be drawn. Possible evidence for putatively adaptive significance of the $NADP^+$ IDH gene has derived from the neutrality tests and the significant inter-specific differentiation (at locus 4) observed by the sequence analysis of the gene. Additionally, the SNP markers that were designed on the basis of the $NADP^+$ IDH gene showed highly significant differentiation among the species and could be considered as an outlier locus in that context. Seven significant associations have been as well suggested between the analysed SNP markers and leaf morphological traits; although it is not likely that the SNPs have a causal relationship with the traits, the possibility that they are linked to sites that are selectively adaptive cannot be excluded.

In the present study the *Quercus* specific nucleotide sequences of an almost complete $NADP^+$ IDH and a partial NAD^+ IDH gene were identified. In that sence, it could form the basis for further analysis of these genes in oaks. Additionally, it is interesting to examine how the SNPs designed here could differentiate oak populations that are low differentiated at neutral markers. Moreover, the SNPs that showed significant associations with the leaf morphological traits are proposed to be tested by QTL mapping applications based on mapping populations and by *Quercus* pedigrees that segregate for leaf morphological traits in order to be confirmed.

4. Summary

Genomic DNA contains the primary source of genetic information. Thus, its analysis provides the highest level of genetic resolution. DNA sequencing and genotyping methodologies based on it in particular genotyping of Single Nucleotide Polymorphisms (SNPs) facilitated new research approaches in forest genetics. Both quantitative and population genetic methods are being applied to better understand the association between genotypic and phenotypic diversity, and to detect the signatures of natural selection upon different parts of the genome in forest trees and their populations.

Oak populations have been widely used as model species to study adaptation of forest trees in variable environments due to their wide geographical range and the large variation of climatic and edaphic condition that they occupy. Four oak species that coexist in the Bejan Oak Reserve (a species rich temperate oak forest in west-central Romania) were investigated in the present study: *Quercus robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto*. The four species are closely related, likely to hybridize and belong taxonomically to the section *Quercus sensu stricto* (white oaks) according to the most recent classifications. According to the geographical range of the species, the site of investigation is for the two more widely spread European species (*Q. robur* and *Q. petraea*) in the centre of their distribution, whereas for the two so called "thermophilous and xerothermic" oak species (*Q. pubescens* and *Q. frainetto*), the Bejan Oak Reserve is located at the east-northern margins of their distribution.

The nucleotide sequences of $NADP^+$ and NAD^+ dependent isocitrate dehydrogenases (IDH) were investigated in the present study. IDH genes are key metabolic enzymes participating in the Krebs cycle. Considerable evidence points towards an adaptive significance of IDH genes in many different organisms. The nucleotide sequence of an $NADP^+$ IDH locus was obtained on the basis of the published genome sequence of *Populus trichocarpa*. In total, 3481bp organized in 15 exons and 14 introns were identified representing the almost complete sequence of the gene. The sequence of the NAD^+ IDH locus was partially obtained with the "genome walking" technique. A total of 2080bp were obtained containing one complete exon and one complete intron. The obtained sequences for both genes were PCR amplified from the genomic DNA of five individuals per species, cloned and sequenced. The PCR amplification for the $NADP^+$ IDH gene was done in seven overlapping fragments.

The nucleotide diversity observed in this study was high, and similar to that found in the few earlier studies dealing with nucleotide diversity in *Quercus* spp., despite the conservative analysis of excluding the Single Sequence Repeat (SSR) motifs and the non-verified singletons from further analyses. This kind of analysis was done in order to avoid incorporation of false variation in the data. In contrast to the high within species nucleotide diversity, the species were in total weakly differentiated at both genes with the exception of significant differentiation among species at the fragment 4 of $NADP^+$ IDH gene. In order to test the neutral model of the coalescent theory Tajima's *D* statistic and the ratio between non-synonymous and silent polymorphisms were tested. Tajima's *D* was found for most loci and species negative, indicating an excess of low frequency variants. However, there were differences detected in the Tajima's *D* species-wide among the different loci, suggesting that the trends might not be subject to demographic effects but rather either an artifact of the low sample size or might reflect the different action of selection upon different fragments even within the same gene. The ratio between non-synonymous and silent polymorphisms was for the $NADP^+$ IDH gene in most fragments low, suggesting purifying selection acting, as expected for proteins and enzyme gene loci. Tajima's *D* for the NAD^+ IDH gene was found positive in three species but not significantly departing from neutral expectations.

From the $NADP^+$ IDH gene, 13 verified SNPs were chosen for further genotyping in a total of 253 white oaks, assigned as "pure" species according to morphological traits and molecular markers by a previous study. The SNPs were chosen to cover the whole gene and to locate in non-coding (SNP 9, SNP 10, SNP 11, SNP 12, and SNP 13) and coding regions. Non-synonymous (SNP 1, SNP 2, SNP 3 and SNP 4) and synonymous SNPs (SNP 5, SNP 6, SNP 7 and SNP 8) were investigated within the coding regions. SNP 3 and SNP 4 result in a charge change through the amino acid replacement caused by the respective nucleotide substitutions.

The levels of gene diversity within species were moderate and similar for all the species, comparable to those exhibited in isozyme analyses. The non-synonymous SNPs showed in general lower levels of gene diversity. Pairwise F_{st} values at all SNP markers revealed low but significant differentiation between almost all pairs of species, except *Q. petraea* and *Q. frainetto*. Non-synonymous SNPs alone failed to differentiate *Q. frainetto* from all other species. Coding SNPs differentiated the species better than non-coding SNPs. This result is mainly caused by the high F_{st} values of the synonymous SNP 6. In particular, SNP 6 was

indicated as an outlier candidate locus for selection by an F_{st} outlier analysis, which compared the observed F_{st} values as a function to the expected heterozygosities, against those expected under a neutral model.

The analysis for non-random association of alleles revealed no clear physical clustering of SNP sites in linkage disequilibrium (LD). The separate analysis of each species showed a lower number of sites in significant LD for *Q. robur* than for the other species, possibly reflecting the history of the species in the specific geographical site, and the less efficient recombination effect due to a larger effective population size of *Q. robur*. On the other hand, *Q. frainetto* and *Q. pubescens*, showed larger number of SNPs in significant LD and spanning larger physical distances, reflecting a possible founder effect due to the location at the margins of their geographical distribution, or the genetic admixture among them. An association analysis of all the SNPs with leaf morphological traits that differentiate the species (measured by Curtu *et al.* 2007b): sinus width, lamina length, basal shape of lamina and petiole length showed seven statistically significant associations under a general linear model. Three of the significant associations were found at non-synonymous SNPs. Statistical association does not necessarily mean a biological or even causal association. Genetic structure is the most serious bias in the association analyses. It is proposed that the SNPs suggested by the associated analysis in this study should be further analyzed in populations that are genetically not differentiated and structured.

5. Zusammenfassung

Genomische DNS ist der primäre Informationsträger des Erbgutes, und die Analyse von DNS-Sequenzen erlaubt die unmittelbare Beobachtung genetischer Information. Methoden der DNS-Sequenzierung und der Genotypisierung, im speziellen die Genotypisierung über *Single Nucleotide Polymorphism* (SNP), hat neue Ansätze auf dem Gebiet der Forstgenetik ermöglicht. Sowohl quantitative als auch populationsgenetische Methoden werden genutzt, um das Verständnis über das Zusammenspiel zwischen genotypischen und phänotypischen Merkmalen zu verbessern und um genomische Signaturen der natürlichen Selektion in Gehölzen und deren Populationen zu charakterisieren.

Aufgrund ihrer Bedeutung und weiten geographischen Verbreitung werden natürliche Populationen von Eichen (*Quercus* spp.) oft als Model zur Studie der Adaption von Gehölzen an verschiedenste klimatische und edaphische Bedingungen genutzt. In der hier vorgelegten Arbeit wurden vier Eichenarten betrachtet: *Quercus robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto*. Diese Arten koexistieren in einem der artenreichesten Eichenwälder Europas, dem Eichen-Reservat Bejan, gelegen im westlichen Zentral-Rumänien. Die vier Arten sind genetisch eng miteinander verwandt, und können Hybride bilden. Taxonomisch werden sie der Sektion *Quercus sensu stricto* (‚Weisseichen') zugeordnet. Bezüglich der geographischen Verbreitung liegt das Bejan Oak Reservat im Zentrum der beiden in Europa dominierenden Arten *Q. robur* und *Q. petraea*, wohingegen für die beiden zu den so genannten "thermophilen und xerophilen" zugehörigen Eichenarten *Q. pubescens* und *Q. frainetto* das Reservat am nordöstlichen Rand ihrer natürlichen Verbreitung liegt.

In der vorliegenden Arbeit wurde die genomische Sequenz der $NADP^+$ und der NAD^+ abhängigen Isocitratdehydrogenase (IDH) untersucht. IDH Gene kodieren für zentrale metabolische Enzyme des Krebs-Zyklus. Grundlegende Forschung an diesen Enzymen deutet auf eine signifikante Rolle für adaptive Ereignisse in verschiedenen Organismen hin. Die Nukleotidsequenz und die genomische Lokalisierung eines $NADP^+$ IDH Gens in *Q. robur* wurde mittels der bekannten Sequenz aus *Populus trichocarpa* entschlüsselt. Das entschlüsselte Gen umfasst 3481bp (Basenpaare) und beinhaltet 15 Exon- und 14 Intronbereiche. Des Weiteren wurde ein NAD^+ IDH Gen per "genome walking" teilweise sequenziert. Das sequenzierte Genfragment hat eine Länge von 2080bp und umfasst das erste Exon und das erste Intron des Gens. Zur populationsgenetischen Analyse der vier

Eichenarten wurden beide Gensequenzen per PCR aus genomischer DNS amplifiziert, kloniert und sequenziert, wobei je Art fünf individuelle Replikate analysiert wurden. Die Gensequenz der $NADP^+$ abhängigen IDH wurde aus sieben sequenzierten PCR Amplifikaten assembliert.

Die hohe Diversität in der DNS Sequenz der hier analysierten IDH Gene zwischen den untersuchten Eichenarten entsprach der hohen genetischen Diversität in vorangegangenen Studien, obwohl ein konservativ ausgwertet wurde; *Single Sequence Repeat* (SSR) Motive und unbestätigte ‚Singleton' SNPs wurden ausgeschlossen, um Artefakte, also ‚falsche' Variation, weitestgehend zu vermeiden. Im Gegensatz zur hohen genetischen intraspezifischen Diversität zeigten die hier untersuchten Eichenarten geringe interspezifische Variabilität mit Ausnahme des vierten Fragmentes (Locus 4) des $NADP^+$ abhängigen IDH Gens, welches eine außerordentlich hohe Differenzierung aufwies.

Das neutrale Modell der Koaleszenztheorie wurde anhand der Tajima *D* Statistik und mit der Berechnung des Verhältnisses zwischen synonymen und nicht-synonymen SNPs getestet. Der Tajima *D* Test war für den überwiegenden Teil der SNP Positionen in allen Arten negativ, was auf einen Überschuss seltener Varianten deutet. Über alle SNP Positionen innerhalb der untersuchen Arten betrachtet zeigte der Tajima *D* Test allerdings Unterschiede, was weniger auf demographische Effekte zurückzuführen sein dürfte als auf unterschiedlich starke Selektionsprozesse, die an verschiedenen Genabschnitten des gleichen Gens stattfanden. Das Verhältnis zwischen synonymen und nicht-synonymen SNPs war innerhalb des $NADP^+$ abhängigen IDH Gens klein, was auf eine gerichtete Selektion hindeutet und für essentielle Enzyme typisch ist. Der Tajima *D* Test war positiv für drei der hier untersuchten Eichenarten, allerdings nicht signifikant abweichend von der Annahme der Neutralität.

Für eine weitere Genotypisierung wurden 13 bestätigte SNP Positionen des $NADP^+$ IDH Gens ausgewählt und bei 253 Individuen untersucht, deren Artzugehörigkeit in einer vorherigen Studie anhand von morphologischen und molekularen Merkmalen bestimmt wurde. SNP Positionen wurden so ausgewählt, dass das komplette Gen und innerhalb dessen sowohl die nicht kodierenden SNPs (SNP 9, SNP 10, SNP 11, SNP 12 uns SNP 13) als auch kodierende SNPs abgedeckt waren. Innerhalb der kodierenden Sequenz wurden sowohl nicht-synonyme SNPs (SNP 1, SNP 2, SNP 3 and SNP 4) als auch synonyme SNPs (SNP 5, SNP 6, SNP 7 and SNP 8) ausgewertet. In den Positionen SNP 3 und SNP 4 kam es

aufgrund der jeweiligen Nukleotidsubstitution und dem daraus folgenden Aminosäurenaustausch zu einer Ladungsverschiebung.

Die intraspezifische Diversität was wie auch bei Isozym-Genorten moderat in allen hier untersuchten Eichenarten, wobei nicht-synonyme SNPs eine geringere Diversität aufwiesen. Paarweise F_{st} Tests mit allen SNP Markern ergaben eine geringe aber signifikante Differenzierung zwischen allen Arten, außer bei den beiden Arten *Q. petraea* und *Q. frainetto*. Nicht-synonyme SNP Marker allein konnten *Q. frainetto* nicht signifikant von einer der anderen Arten differenzieren. Des Weiteren differenzierten kodierende SNP Marker alle Arten deutlich besser als nicht kodierende SNPs, was mit dem abweichend hohen F_{st} Wert für die Position SNP 6 zu erklären ist. Dieser SNP erwies sich sogar als statistisch signifikanter ‚Ausreißer' bei der Betrachtung der Differenzierung zwischen den Arten relativ zur als ‚erwarteter Heterozygotie' gemessener Variation bei Annahme des Neutralitätsmodells.

Eine Analyse auf nicht-zufällige Assoziationen von Allelen ergab keine eindeutigen Gruppierungen der identifizierten SNP Positionen; Kopplungsungleichgewichte (Linkage Disequilibrium, LD) wurden sowohl zwischen nahe beieinanderliegenden als auch weiter voneinander entfernten SNP Positionen beobachtet. Die getrennte Analyse jeder Art ergab eine geringere Anzahl an Positionen innerhalb signifikanter LD für die Art *Q. robur*, was wahrscheinlich sowohl auf eine lange Geschichte dieser Art in der Region hinweist, als auch auf eine geringere Neigung zur Rekombination. Dies kann mit besonder großen Populationen bei *Q. robur* zusammenhängen. Auf der anderen Seite zeigten *Q. frainetto* and *Q. pubescens* eine höhere Anzahl von SNPs mit signifikantem LD auch bei Paaren mit größerer physischer Distanz auf. Dies könnte mit einem "Gründereffekt" am Rande ihrer geographischen Ausbreitung oder mit der genetischen Vermischung untereinander zusammenhängen. Eine Assoziationsanalyse aller SNPs mit blatt-morphologischen Daten (Sinus-Weite, Lamina-Länge, Grundform der Lamina, Blattstiellänge), die die untersuchten Eicharten differenzieren (Curtu *et al.* 2007b), ergaben sieben statistisch signifikante Zusammenhänge unter Verwendung eines linearen Modells. Drei der signifikanten Assoziationen wurden in nicht-synonymen SNPs identifiziert. Die Berechnung statistischer Assoziationen ist nicht automatisch gleichbedeutend mit biologischen oder sogar kausalen Assoziationen. Daher ist es ratsam, die in dieser Arbeit ausgewerteten SNP Assoziationen an weiteren genetisch nicht strukturierten und differenzierten Populationen zu bestätigen.

6. References

Aas, G. (2006a). "*Quercus robur* L." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-14. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm, Ecomed,, Landsber am Lech.

Aas, G. (2006b). "*Quercus petraea* (Mattuchka) Lieblein". Enzyklopädie der Holzgewächse: Handbuch und Atlas der Dendrologie**:** 1-16. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm., Ecomed,, Landsber am Lech.

Aitken, S. N., S. Yeaman, J. A. Holliday, T. Wang and S. Curtis-McLane (2008). "Adaptation, migration or extirpation: climate change outcomes for tree populations." Evolutionary Applications **1**(1): 95-111.

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver (2002). "Interrogating a high-density SNP map for signatures of natural selection." Genome Research **12**(12): 1805-1814.

Anderson, T. J. C., S. Nair, D. Sudimack, J. T. Williams, M. Mayxay, P. N. Newton, J. P. Guthmann, F. M. Smithuis, T. T. Hien, I. V. F. van den Broek, N. J. White and F. Nosten (2005). "Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites." Molecular Biology and Evolution **22**(12): 2362-2374.

Bacilieri, R., A. Ducousso, R. J. Petit and A. Kremer (1996). "Mating system and asymmetric hybridization in a mixed stand of European oaks." Evolution **50**(2): 900-908.

Bartha, D. (2006). "*Quercus frainetto* Ten." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-8. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm, Ecomed,, Landsber am Lech.

Beaumont, M. A. (2005). "Adaptation and speciation: what can F_{st} tell us?" Trends in Ecology & Evolution **20**(8): 435-440.

Beaumont, M. A. and R. A. Nichols (1996). "Evaluating Loci for Use in the Genetic Analysis of Population Structure." Proceedings of the Royal Society of London. Series B: Biological Sciences **263**(1377): 1619-1626.

Behal, R. H. and D. J. Oliver (1998). "NAD(+)-dependent isocitrate dehydrogenase from *Arabidopsis thaliana*. Characterization of two closely related subunits." Plant Molecular Biology **36**(5): 691-698.

Bellarosa, R., M. C. Simeone, A. Papini and B. Schirone (2005). "Utility of ITS sequence data for phylogenetic reconstruction of Italian *Quercus* spp." Molecular Phylogenetics and Evolution **34**(2): 355-370.

Bergmann, F. and H. R. Gregorius (1993). "Ecogeographical Distribution and Thermostability of Isocitrate Dehydrogenase (Idh) Alloenzymes in European Silver Fir (*Abies alba*)." Biochemical Systematics and Ecology **21**(5): 597-605.

Bodénès, C., S. Joandet, F. Laigret and A. Kremer (1997). "Detection of genomic regions differentiating two closely related oak species *Quercus petraea* (Matt) Liebl and Quercus robur L." Heredity **78**: 433-444.

Boiffin, V., M. Hodges, S. Galvez, R. Balestrini, P. Bonfante, P. Gadal and F. Martin (1998). "Eucalypt NADP-dependent isocitrate dehydrogenase - cDNA cloning and expression in ectomycorrhizae." Plant Physiology **117**(3): 939-948.

Bonduriansky, R. and T. Day (2009). "Nongenetic Inheritance and Its Evolutionary Implications." Annual Review of Ecology Evolution and Systematics **40**: 103-125.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss and E. S. Buckler (2007). "TASSEL: software for association mapping of complex traits in diverse samples." Bioinformatics **23**(19): 2633-2635.

Bussotti, F. (2006). "*Quercus pubescens* Willd." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-10. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm., Ecomed,, Landsber am Lech.

Casasoli, M., J. Derory, C. Morera-Dutrey, O. Brendel, I. Porth, J. M. Guehl, F. Villani and A. Kremer (2006). "Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map." Genetics **172**(1): 533-546.

Charlesworth, B., M. T. Morgan and D. Charlesworth (1993). "The Effect of Deleterious Mutations on Neutral Molecular Variation." Genetics **134**(4): 1289-1303.

Chen, R., P. Maréchal, J. Vidal, J.-P. Jacquot and P. Gadal (1988). "Purification and comparative properties of the cytosolic isocitrate dehydrogenases (NADP) from pea (*Pisum sativum*) roots and green leaves." European Journal of Biochemistry **175**(3): 565-572.

Chybicki, I. J. and J. Burczyk (2010). "Realized gene flow within mixed stands of *Quercus robur* L. and *Q. petraea* (Matt.) L. revealed at the stage of naturally established seedling." Molecular Ecology **19**(10): 2137-2151.

Corpas, F. J., J. B. Barroso, L. M. Sandalio, J. M. Palma, J. A. Lupianez and L. A. del Rio (1999). "Peroxisomal NADP-dependent isocitrate dehydrogenase. Characterization and activity regulation during natural senescence." Plant Physiology **121**(3): 921-928.

Craft, K. J. and M. V. Ashley (2006). "Population differentiation among three species of white oak in northeastern Illinois." Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **36**(1): 206-215.

Curtu, A. L., O. Gailing and R. Finkeldey (2007b). "Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community." BMC Evolutionary Biology **7**.

Curtu, A. L., O. Gailing and R. Finkeldey (2009). "Patterns of contemporary hybridization inferred from paternity analysis in a four-oak-species forest." BMC Evolutionary Biology **9**.

Curtu, A. L., O. Gailing, L. Leinemann and R. Finkeldey (2007a). "Genetic variation and differentiation within a natural community of five oak species (*Quercus* spp.)." Plant Biology **9**(1): 116-126.

Dang, L., D. W. White, S. Gross, B. D. Bennett, M. A. Bittinger, E. M. Driggers, V. R. Fantin, H. G. Jang, S. Jin, M. C. Keenan, K. M. Marks, R. M. Prins, P. S. Ward, K. E. Yen, L. M. Liau, J. D. Rabinowitz, L. C. Cantley, C. B. Thompson, M. G. Vander Heiden and S. M. Su (2009). "Cancer-associated IDH1 mutations produce 2-hydroxyglutarate." Nature **462**(7274): 739-744.

Derory, J., C. Scotti-Saintagne, E. Bertocchi, L. Le Dantec, N. Graignic, A. Jauffres, M. Casasoli, E. Chancerel, C. Bodenes, F. Alberto and A. Kremer (2010). "Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks." Heredity **104**(5): 438-448.

Ducousso, A., H. Michaud and R. Lumaret (1993). "Reproduction and gene flow in the genus *Quercus* L." Ann. For. Sci. **50**(Supplement): 91s-106s.

Dvornyk, V., A. Sirvio, M. Mikkonen and O. Savolainen (2002). "Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*." Molecular Biology and Evolution **19**(2): 179-188.

Eanes, W. F. (1999). "Analysis of selection on enzyme polymorphisms." Annual Review of Ecology and Systematics **30**: 301-326.

Eckert, A. J., J. L. Wegrzyn, B. Pande, K. D. Jermstad, J. M. Lee, J. D. Liechty, B. R. Tearse, K. V. Krutovsky and D. B. Neale (2009). "Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in Coastal Douglas Fir (*Pseudotsuga menziesii* var. *menziesii*)." Genetics **183**(1): 289-298.

Ellenberg, H. (1988)."Vegetation Ecology of Central Europe". 147,172 Cambridge, Cambrige University Press.

Eveno, E., C. Collada, M. A. Guevara, V. Leger, A. Soto, L. Diaz, P. Leger, S. C. González-Martínez, M. T. Cervera, C. Plomion and P. H. Garnier-Géré (2008). "Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses." Molecular Biology and Evolution **25**(2): 417-437.

Excoffier, L., T. Hofer and M. Foll (2009). "Detecting loci under selection in a hierarchically structured population." Heredity **103**(4): 285-298.

Excoffier, L. and H. E. L. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." Molecular Ecology Resources **10**(3): 564-567.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). "Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data." Genetics **131**(2): 479-491.

Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**(3): 1405-1413.

Fay, J. C., G. J. Wyckoff and C. I. Wu (2002). "Testing the neutral theory of molecular evolution with genomic data from *Drosophila*." Nature **415**(6875): 1024-1026.

Fedøy, A.-E., N. Yang, A. Martinez, H.-K. S. Leiros and I. H. Steen (2007). "Structural and Functional Properties of Isocitrate Dehydrogenase from the Psychrophilic Bacterium *Desulfotalea psychrophila* Reveal a Cold-active Enzyme with an Unusual High Thermal Stability." Journal of Molecular Biology **372**(1): 130-149.

Fernandez, J. F. and V. L. Sork (2005). "Mating patterns of a subdivided population of the Andean oak (*Quercus humboldtii* Bonpl., *Fagaceae*)." Journal of Heredity **96**(6): 635-643.

Fieuw, S., B. Mullerrober, S. Galvez and L. Willmitzer (1995). "Cloning and Expression Analysis of the Cytosolic Nadp(+)-Dependent Isocitrate Dehydrogenase from Potato - Implications for Nitrogen-Metabolism." Plant Physiology **107**(3): 905-913.

Finkeldey, R. (2001). Genetic variation of oaks (*Quercus* spp.) in Switzerland - 2. Genetic structures in "pure" and "mixed" forests of pedunculate oak (*Q. robur* L.) and sessile oak (*Q. petraea* (Matt.) Liebl.). Silvae genetica. **50:** 22-30.

Finkeldey, R. and G. Mátyás (2000). "Assessment of population history and adaptive potential by means of gene markers." Forest Genetics and Sustainability**:** 49-58. Mátyás, C., Kluwer Academic Publishers, Netherlands.

Finkeldey, R. and G. Mátyás (2003). "Genetic variation of oaks (*Quercus* spp.) in Switzerland. 3. Lack of impact of postglacial recolonization history on nuclear gene loci." Theoretical and Applied Genetics **106**(2): 346-352.

Ford, M. J. (2002). "Applications of selective neutrality tests to molecular ecology." Molecular Ecology **11**(8): 1245-1262.

Fu, Y. X. and W. H. Li (1993). "Statistical Tests of Neutrality of Mutations." Genetics **133**(3): 693-709.

Gailing, O. (2008). "QTL analysis of leaf morphological characters in a Quercus robur full-sib family (*Q. robur* x *Q. robur* ssp. *slavonica*)." Plant Biology **10**(5): 624-634.

Gailing, O., R. Langenfeld-Heyser, A. Polle and R. Finkeldey (2008). "Quantitative trait loci affecting stomatal density and growth in a *Quercus robur* progeny: implications for the adaptation to changing environments." Global Change Biology **14**(8): 1934-1946.

Gailing, O., B. Vornam, L. Leinemann and R. Finkeldey (2009). "Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model." Physiologia Plantarum **137**(4): 509-519.

Gálvez, S., E. Bismuth, C. Sarda and P. Gadal (1994). "Purification and Characterization of Chloroplastic NADP-Isocitrate Dehydrogenase from Mixotrophic Tobacco Cells - Comparison with the Cytosolic Isoenzyme." Plant Physiology **105**(2): 593-600.

Goldman, N. and Z. H. Yang (1994). "Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences." Molecular Biology and Evolution **11**(5): 725-736.

Gömöry, D., I. Yakovlev, P. Zhelev, J. Jedinakova and L. Paule (2001). "Genetic differentiation of oak populations within the *Quercus robur/Quercus petraea* complex in Central and Eastern Europe." Heredity **86**(5): 557-563.

González-Martínez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale (2006). "DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L." Genetics **172**(3): 1915-1926.

González-Martínez, S. C., K. V. Krutovsky and D. B. Neale (2006). "Forest-tree population genomics and adaptive evolution." New Phytologist **170**(2): 227-238.

González-Martínez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson and D. B. Neale (2007). "Association genetics in *Pinus taeda* L. I. Wood property traits." Genetics **175**(1): 399-409.

Goudet, J. (1995). "FSTAT (Version 1.2): A computer program to calculate F-statistics." Journal of Heredity **86**(6): 485-486.

Hall, D., C. Tegstrom and P. K. Ingvarsson (2010). "Using association mapping to dissect the genetic basis of complex traits in plants." Briefings in Functional Genomics **9**(2): 157-165.

Hall, T. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series **41**: 95-98.

Hedi, G. (1981). "*Quercus* L." Illustrierte Flora von Mittel-Europa **3:** 221-222. Paul Parey, Berlin.

Hedrick, P. W. (2005)."Genetics of Populations". Sudbury, Massachusetts, Jones and Barlett Publishers.

Hodges, M., V. Flesch, S. Gálvez and E. Bismuth (2003). "Higher plant NADP+-dependent isocitrate dehydrogenases, ammonium assimilation and NADPH production." Plant Physiology and Biochemistry **41**(6-7): 577-585.

Hudson, R. R. (2000). "A new statistic for detecting genetic differentiation." Genetics **155**(4): 2011-2014.

Hudson, R. R., D. D. Boos and N. L. Kaplan (1992a). "A Statistical Test for Detecting Geographic Subdivision." Molecular Biology and Evolution **9**(1): 138-151.

Hudson, R. R., M. Kreitman and M. Aguade (1987). "A Test of Neutral Molecular Evolution Based on Nucleotide Data." Genetics **116**(1): 153-159.

Hudson, R. R., M. Slatkin and W. P. Maddison (1992b). "Estimation of Levels of Gene Flow From DNA Sequence Data." Genetics **132**(2): 583-589.

Huestis, D. L., B. Oppert and J. L. Marshall (2009). "Geographic distributions of Idh-1 alleles in a cricket are linked to differential enzyme kinetic performance across thermal environments." BMC Evolutionary Biology **9**.

Ingvarsson, P. K. (2005). "Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (*Populus tremula* L., *Salicaceae*)." Genetics **169**(2): 945-953.

Ingvarsson, P. K., M. V. Garcia, V. Luquez, D. Hall and S. Jansson (2008). "Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, *Salicaceae*)." Genetics **178**(4): 2217-2226.

Jalas, J. and J. Suominen (1976). "*Fagaceae*". Atlas *Flora Europaea* - Distribution of Vascular Plants in Europe **3:** 128. Soc. Bot. Fennica Vanamo, Helsinki.

Jensen, J., A. Larsen, L. Nielsen, R. and J. Cottrell (2009). "Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark." Ann. For. Sci. **66**(7): 706.

Kalisz, S. and E. M. Kramer (2008). "Variation and constraint in plant evolution and development." Heredity **100**(2): 171-177.

Keys, D. A. and L. McAlisterhenn (1990). "Subunit Structure, Expression, and Function of NAD(H)-Specific Isocitrate Dehydrogenase in *Saccharomyces cerevisiae*." Journal of Bacteriology **172**(8): 4280-4287.

Kimura, M. (1977). "Preponderance of Synonymous Changes as Evidence for Neutral Theory of Molecular Evolution." Nature **267**(5608): 275-276.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge, U.K., Cambridge University Press.

Kreitman, M. (1996). "The neutral theory is dead. Long live the neutral theory." BioEssays **18**(8): 678-683.

Kreitman, M. and J. M. Comeron (1999). "Coding sequence evolution." Current Opinion in Genetics & Development **9**(6): 637-641.

Kremer, A. and X. Goenaga (2002). "Special Issue: Range wide distribution of chloroplast DNA diversity and pollen deposits in European white oaks: inferences about colonisation routes and management of oak genetic resources. Preface." Forest Ecology and Management **156**(1-3): 1-3.

Kremer, A., J. Kleinschmit, J. Cottrell, E. P. Cundall, J. D. Deans, A. Ducousso, A. O. Konig, A. J. Lowe, R. C. Munro, R. J. Petit and B. R. Stephan (2002). "Is there a correlation between chloroplastic and nuclear divergence, or what are the roles of history and selection on genetic diversity in European oaks?" Forest Ecology and Management **156**(1-3): 75-87.

Kuparinen, A., O. Savolainen and F. M. Schurr (2010). "Increased mortality can promote evolutionary adaptation of forest trees to climate change." Forest Ecology and Management **259**(5): 1003-1008.

Le Corre, V. and A. Kremer (2003). "Genetic Variability at Neutral Markers, Quantitative Trait Loci and Trait in a Subdivided Population Under Selection." Genetics **164**(3): 1205-1219.

Lemaitre, T. and M. Hodges (2006). "Expression analysis of *Arabidopsis thaliana* NAD-dependent isocitrate dehydrogenase genes shows the presence of a functional subunit that is mainly expressed in the pollen and absent from vegetative organs." Plant and Cell Physiology **47**(5): 634-643.

Levy, G., M. Becker and D. Duhamel (1992). "A Comparison of the Ecology of Pedunculate and Sessile Oaks - Radial Growth in the Center and Northwest of France." Forest Ecology and Management **55**(1-4): 51-63.

Lewontin, R. C. (1991). "25 Years Ago in Genetics - Electrophoresis in the Development of Evolutionary Genetics - Milestone or Millstone." Genetics **128**(4): 657-662.

Lewontin, R. C. and J. Krakauer (1973). "Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms." Genetics **74**(1): 175-195.

Lexer, C. and A. Widmer (2008). "The genic view of plant speciation: recent progress and emerging questions." Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1506): 3023-3036.

Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-1452.

Long, A. D. and C. H. Langley (1999). "The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits." Genome Research **9**(8): 720-731.

Manos, P. S., J. J. Doyle and K. C. Nixon (1999). "Phylogeny, Biogeography, and Processes of Molecular Differentiation in *Quercus* Subgenus *Quercus* (*Fagaceae*)." Molecular Phylogenetics and Evolution **12**(3): 333-349.

Mariette, S., J. Cottrell, U. M. Csaikl, P. Goikoechea, A. Konig, A. J. Lowe, B. C. Van Dam, T. Barreneche, C. Bodenes, R. Streiff, K. Burg, K. Groppe, R. C. Munro, H. Tabbener and A. Kremer (2002). "Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Q. petraea* (MATT.) LIEBL. and *Q. robur* L. stands." Silvae Genetica **51**(2-3): 72-79.

Mayol, M. and J. A. Rosselló (2001). "Why Nuclear Ribosomal DNA Spacers (ITS) Tell Different Stories in *Quercus*." Molecular Phylogenetics and Evolution **19**(2): 167-176.

Mcdonald, J. H. and M. Kreitman (1991). "Adaptive Protein Evolution at the Adh Locus in *Drosophila*." Nature **351**(6328): 652-654.

Moeller, D. A., M. I. Tenaillon and P. Tiffin (2007). "Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*)." Genetics **176**(3): 1799-1809.

Muir, G. and C. Schlotterer (2005). "Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.)." Molecular Ecology **14**(2): 549-561.

Neale, D. B. and P. K. Ingvarsson (2008). "Population, quantitative and comparative genomics of adaptation in forest trees." Current Opinion in Plant Biology **11**(2): 149-155.

Neale, D. B. and O. Savolainen (2004). "Association genetics of complex traits in conifers." Trends in Plant Science **9**(7): 325-330.

Nei, M. (1973). "Analysis of Gene Diversity in Subdivided Populations." Proceedings of the National Academy of Sciences of the United States of America **70**(12): 3321-3323.

Nei, M. (1987)."Molecular Evolutionary Genetics". New York, Columbia Univ. Press.

Nei, M. (2005). "Selectionism and neutralism in molecular evolution." Molecular Biology and Evolution **22**(12): 2318-2342.

Nekrutenko, A., D. M. Hillis, J. C. Patton, R. D. Bradley and R. J. Baker (1998). "Cytosolic isocitrate dehydrogenase in humans, mice, and voles and phylogenetic analysis of the enzyme family." Molecular Biology and Evolution **15**(12): 1674-1684.

Neophytou, C., F. A. Aravanopoulos, S. Fink and A. Dounavi (2010). "Detecting interspecific and geographic differentiation patterns in two interfertile oak species (*Quercus petraea* (Matt.) Liebl. and *Q. robur* L.) using small sets of microsatellite markers." Forest Ecology and Management **259**(10): 2026-2035.

Nielsen, R. (2001). "Statistical tests of selective neutrality in the age of genomics." Heredity **86**: 641-647.

Nielsen, R. (2005). "Molecular Signatures of Natural Selection." Annual Review of Genetics **39**(1): 197-218.

Nixon, K. (1993). "Infrageneric classification of *Quercus* (*Fagaceae*) and typification of sectional names." Ann. For. Sci. **50**(Supplement): 25s-34s.

Palme, A. E., M. Wright and O. Savolainen (2008). "Patterns of Divergence among Conifer ESTs and Polymorphism in Pinus sylvestris Identify Putative Selective Sweeps." Molecular Biology and Evolution **25**(12): 2567-2577.

Palomo, J., F. Gallardo, M. F. Suarez and F. M. Canovas (1998). "Purification and characterization of NADP(+)-linked isocitrate dehydrogenase from Scots pine - Evidence for different physiological roles of the enzyme in primary development." Plant Physiology **118**(2): 617-626.

Parelle, J., M. Zapater, C. Scotti-Saintagne, A. Kremer, Y. Jolivet, E. Dreyer and O. Brendel (2007). "Quantitative trait loci of tolerance to waterlogging in a European oak (*Quercus robur* L.): physiological relevance and temporal effect patterns." Plant Cell and Environment **30**(4): 422-434.

Pascual, M. B., Z. P. Jing, E. G. Kirby, F. M. Canovas and F. Gallardo (2008a). "Response of transgenic poplar overexpressing cytosolic glutamine synthetase to phosphinothricin." Phytochemistry **69**(2): 382-389.

Pascual, M. B., J. J. Molina-Rueda, F. M. Canovas and F. Gallardo (2008b). "Spatial distribution of cytosolic NADP(+)-isocitrate dehydrogenase in pine embryos and seedlings." Tree Physiology **28**(12): 1773-1782.

Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen and A.-C. Syvänen (1997). "Minisequencing: A Specific Tool for DNA Analysis and Diagnostics on Oligonucleotide Arrays." Genome Research **7**(6): 606-614.

Petit, R. J., C. Bodenes, A. Ducousso, G. Roussel and A. Kremer (2004). "Hybridization as a mechanism of invasion in oaks." New Phytologist **161**(1): 151-164.

Petit, R. J., U. M. Csaikl, S. Bordacs, K. Burg, E. Coart, J. Cottrell, B. van Dam, J. D. Deans, S. Dumolin-Lapegue, S. Fineschi, R. Finkeldey, A. Gillies, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. Konig, A. J. Lowe, S. F. Madsen, G. Mátyás, R. C. Munro, M. Olalde, M. H. Pemonge, F. Popescu, D. Slade, H. Tabbener, D. Taurchini, S. G. M. de Vries, B. Ziegenhagen and A. Kremer (2002a). "Chloroplast DNA variation in European white oaks - Phylogeography and patterns of diversity based on data from over 2600 populations." Forest Ecology and Management **156**(1-3): 5-26.

Petit, R. J., E. Pineau, B. Demesure, R. Bacilieri, A. Ducousso and A. Kremer (1997). "Chloroplast DNA footprints of postglacial recolonization by oaks." Proceedings of the National Academy of Sciences of the United States of America **94**(18): 9996-10001.

Petit, R. m. J. and A. Hampe (2006). "Some Evolutionary Consequences of Being a Tree." Annual Review of Ecology, Evolution, and Systematics **37**(1): 187-214.

Porth, I., C. Scotti-Saintagne, T. Barreneche, A. Kremer and K. Burg (2005). "Linkage mapping of osmotic stress induced genes of oak." Tree Genetics & Genomes **1**(1): 31-40.

Pot, D., L. McMillan, C. Echt, G. Le Provost, P. Garnier-Géré, S. Cato and C. Plomion (2005). "Nucleotide variation in genes involved in wood formation in two pine species." New Phytologist **167**(1): 101-112.

Quang, N. D., S. Ikeda and K. Harada (2008). "Nucleotide variation in *Quercus crispula* Blume." Heredity **101**(2): 166-174.

Quang, N. D., S. Ikeda and K. Harada (2009). "Patterns of Nucleotide Diversity at the Methionine Synthase Locus in Fragmented and Continuous Populations of a Wind-Pollinated Tree, *Quercus mongolica* var. *crispula*." J Hered **100**(6): 762-770.

Rosenberg, N. A. and M. Nordborg (2002). "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms." Nat Rev Genet **3**(5): 380-390.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." **132**: 365-386.

Sadka, A., E. Dahan, E. Or and L. Cohen (2000). "NADP(+)-isocitrate dehydrogenase gene expression and isozyme activity during citrus fruit development." Plant Science **158**(1-2): 173-181.

Salvini, D., P. Bruschi, S. Fineschi, P. Grossoni, E. D. Kjaer and G. G. Vendramin (2009). "Natural hybridisation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens* Willd. within an Italian stand as revealed by microsatellite fingerprinting." Plant Biology **11**(5): 758-765.

Samuel, R., A. Bachmair, J. Jobst and F. Ehrendorfer (1998). "ITS sequences from nuclear rDNA suggest unexpected phylogenetic relationships between Euro-Mediterranean, East Asiatic and North American taxa of *Quercus* (*Fagaceae*)." Plant Systematics and Evolution **211**(1): 129-139.

Savolainen, O. and T. Pyhäjärvi (2007). "Genomic diversity in forest trees." Current Opinion in Plant Biology **10**(2): 162-167.

Savolainen, O., T. Pyhäjärvi and T. Knurr (2007). "Gene flow and local adaptation in trees." Annual Review of Ecology Evolution and Systematics **38**: 595-619.

Schwarz, O. (1993). "*Quercus*". Flora Europaea **1** Tutin, T. G., Burges, N. A., Chater, A. O., Edmondson, J. R., Heywood, V. H., Moore, D. M., Valentine, D. H., Walters, S. M., Webb, D. A., Cambridge University Press, Cambridge, UK.

Scotti-Saintagne, C., S. Mariette, I. Porth, P. G. Goicoechea, T. Barreneche, K. Bodenes, K. Burg and A. Kremer (2004). "Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q petraea* (Matt.) Liebl.]." Genetics **168**(3): 1615-1626.

Siebert, P. D., A. Chenchik, D. E. Kellogg, K. A. Lukyanov and S. A. Lukyanov (1995). "An Improved Pcr Method for Walking in Uncloned Genomic DNA." Nucleic Acids Research **23**(6): 1087-1088.

Simko, I., D. A. Pechenick, L. K. McHale, M. J. Truco, O. E. Ochoa, R. W. Michelmore and B. E. Scheffler (2009). "Association mapping and marker-assisted selection of the lettuce dieback resistance gene Tvr1." BMC Plant Biology **9**.

Sork, V. L., F. W. Davis, P. E. Smouse, V. J. Apsit, R. J. Dyer, J. F. Fernandez and B. Kuhn (2002). "Pollen movement in declining populations of California Valley oak, *Quercus lobata*: where have all the fathers gone?" Molecular Ecology **11**(9): 1657-1668.

Stinchcombe, J. R. and H. E. Hoekstra (2008). "Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits." Heredity **100**(2): 158-170.

Streiff, R., A. Ducousso, C. Lexer, H. Steinkellner, J. Gloessl and A. Kremer (1999). "Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L and *Q. petraea* (Matt.) Liebl." Molecular Ecology **8**(5): 831-841.

Tajima, F. (1989). "Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." Genetics **123**(3): 585-595.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "ClustalW - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, et al. (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)." Science **313**(5793): 1596-1604.

Vornam, B., O. Gailing, R. Finkeldey, C. Collada, M. A. Guevara, A. Soto, d. M. N., S. C. González-Martínez, D. L., A. R., I. Aranda, J. Climent, M. T. Cervera, P. G. Goicoechea, L. V., E. Eveno, J. Derory, P. Garnier-Géré, A. Kremer and C. Plomion (2007). "Naturally occuring nucleotide diversity in candidate genes for forest tree adaptation: magnitude, distribution and association with quantitative trait variation." The German Plant Genome Research Program Progress Report 2004-2007**:** 116-120. GABI, Potsdam-Golm, Germany.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley (1999). "The limits of selection during maize domestication." Nature **398**(6724): 236-239.

Wheat, C. W., C. R. Hagg, J. H. Marden, I. Hanski and M. J. Frilander (2010). "Nucleotide Polymorphism at a Gene (Pgi) under Balancing Selection in a Butterfly Metapopulation." Molecular Biology and Evolution **27**(2): 267-281.

Wu, C.-I. (2001). "The genic view of the process of speciation." Journal of Evolutionary Biology **14**(6): 851-865.

Zhang, N., A. Gur, Y. Gibon, R. Sulpice, S. Flint-Garcia, M. D. McMullen, M. Stitt and E. S. Buckler (2010). "Genetic Analysis of Central Carbon Metabolism Unveils an Amino Acid Substitution That Alters Maize NAD-Dependent Isocitrate Dehydrogenase Activity." PLoS ONE **5**(4): e9991.

Zhang, D.-q. and Z.-y. Zhang (2005). "Single nucleotide polymorphisms (SNPs) discovery and linkage disequilibrium (LD) in forest trees." Forestry Studies in China **7**(3): 1-14.

Zhao, K. Y., M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. L. Tang, C. Toomajian, H. G. Zheng, C. Dean, P. Marjoram and M. Nordborg (2007). "An *Arabidopsis* example of association mapping in structured samples." Plos Genetics **3**(1).

Ziehe, M. (1996a)."Anpassungsprozesse auf Populationsebene" Beese, F. Berichte des Forschungzentrums Waldökosysteme der Universität Göttingen, Reihe B. **51** 126-136 Göttingen.

Ziehe, M. (1996b)."Untersuchungen von Buchenbeständen zur Beurteilung von genetischen Anpassungsprozessen und Angepasstheit anhand von Markergenloci" Beese, F. Berichte des Forschungzentrums Waldökosysteme der Universität Göttingen, Reihe B. **52** 370-378 Göttingen.

I. Patterns of nucleotide diversity and differentiation at $NADP^+$ and NAD^+ isocitrate dehydrogenases in a four-species sympatric white oak community

Introduction

Genomic DNA sequences provide the highest resolution of genetic information. Nucleotide sequence data have become one of the most promising and preferred type of data that can supply an insight into the evolutionary dynamics of populations and lead to inferences, concerning the action of natural selection. Until recently most data of nuclear DNA nucleotide variation of plants described annual or short-lived selfing plants (Kawabe and Miyashita 1999; Small *et al.* 1999; Bundock and Henry 2004; Nordborg *et al.* 2005). The outcrossing long-lived tree species have been mainly represented by conifers (Dvornyk *et al.* 2002; Garcia-Gil *et al.* 2003; Semerikov and Lascoux 2003; Heuertz *et al.* 2006; Pyhäjärvi *et al.* 2007; Palme *et al.* 2008; Li *et al.* 2010). Fewer studies have been dedicated to investigate the patterns of nucleotide diversity and differentiation in angiosperms (Järvinen *et al.* 2003; Ingvarsson 2005; Garcia and Ingvarsson 2007; Hall *et al.* 2007; Ingvarsson *et al.* 2008), among which the largest number refer to poplars (*Populus* spp.) due to the fully sequenced genome of *Populus trichocarpa* (Tuskan *et al.* 2006). Indeed, very few published reports of nucleotide diversity in oaks are available, two of them describing an Asian oak species (*Quercus mongolica* var. *crispula*) (Quang *et al.* 2008; Quang *et al.* 2009) and only one focusing on a European oak species (*Quercus petraea)* (Derory *et al.* 2010). Still, no study until now has described the levels of nucleotide diversity in the most commonly distributed European oak species *Q. robur*, as well as *Q. pubescens* and *Q. frainetto*.

Quercus robur, *Q. petraea*, *Q. pubescens* and *Q. frainetto* (*Fagaceae*) are closely related species and all belong taxonomically to the section *Quercus* of the homonymous genus (also referred to as white oaks) (Nixon 1993; Manos *et al.* 1999). The geographical distribution of the first two species is wider over north-central Europe and in most cases overlapping, whereas the distribution of the latter two species is limited to southern habitats for *Q. pubescens* and mainly to the Balkan and the Italian peninsulas for *Q. frainetto* (Bartha 2006; Bussotti 2006; Aas 2006a; Aas 2006b). Despite their overlapping distributions the four species rarely coexist. since their ecological demands are different, exhibiting different

climatic and edaphic optima (Ellenberg 1988). According to many authors this fact accounts for the distinct evolution of the two most closely related and often sympatric European *Quercus* species, *Q. robur* and *Q. petraea*, due to ongoing diversifying selection (Le Corre and Kremer 2003; Scotti-Saintagne *et al.* 2004) in spite of their general low species differentiation.

In this study we obtained the DNA sequences encoding for $NADP^+$ and NAD^+ dependent isocitrate dehydrogenases (IDH) loci (E.C. 1.1.1.41 and 1.1.1.42 respectively). Both genes are key enzymes of the metabolic pathway of the citrate cycle (Krebs cycle). IDH genes have been considered by many authors studying different organisms as potentially adaptive (Bergmann and Gregorius 1993; Boiffin *et al.* 1998; Pascual *et al.* 2008b; Huestis *et al.* 2009; Zhang *et al.* 2010). In the case of *Quercus* spp. IDH isozymes have been characterized as "outlier" loci, revealing high differentiation among different oak species (Finkeldey 2001; Gömöry *et al.* 2001; Scotti-Saintagne *et al.* 2004; Curtu *et al.* 2007) and therefore being possibly under selection. For the two loci ($NADP^+$ and NAD^+ dependent isocitrate dehydrogenases) the sequences that were obtained were analyzed for their patterns of nucleotide diversity, genetic differentiation, non-random association of allelic variants and the detection of any possible signature of natural selection acting in a mixed community of *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto*.

The four species co-existed and co-evolved in the same region, despite their different ecological demands. The area of the study is located in central-eastern Europe, at the centre of the geographical range of *Q. robur* and *Q. petraea* (Aas 2006a; Aas 2006b) and close to the northern and north-eastern distribution limits of the thermophilous species *Q. frainetto* and *Q. pubescens*, respectively (Bartha 2006; Bussotti 2006).

Materials and Methods

Plant material

Five samples per species were investigated for 20 white oaks (*Q. robur*, *Q. petraea*, *Q. pubescens*, *and Q. frainetto*. all considered as *sensu lato*). These samples were selected randomly as a sub-sample of 269 exhaustively sampled oak trees, with no *a priori* selection, as described in Curtu *et al.* (2007). The four species coexist naturally in an oak community in the Bejan forest located in central-western Romania. Extraction of genomic DNA was conducted and described in detail by Curtu *et al.* (2007).

Identification of the $NADP^+$ and NAD^+ IDH sequences

To obtain the sequence of a cytosolic $NADP^+$ dependent IDH gene PCR reactions were performed as described below using DNA template of a *Q. petraea* sample with oligo-primers that were designed in three overlapping parts on the basis of the public available sequences of *Populus trichocarpa* (jgi|Poptr1_1|770098|fgenesh4_pg.C_LG_X001588) (Tuskan *et al.* 2006) - given the lack of any public available annotated genome database of *Quercus* spp. On the basis of the *Q. petraea* sequence that was obtained (here referred to as "reference sequence"), new primers were designed for specificity to *Quercus* spp. For sequencing purposes the PCR primers were designed to amplify seven overlapping DNA fragments with the overlap varying between 60-300bp.

For the identification of the NAD^+ dependent IDH gene the technique of "genome walking" (Siebert *et al.* 1995) was applied on a sample of *Q. robur*. The method is described in detail in the corresponding paragraph. Based on the sequence obtained, PCR primers were designed to amplify a single fragment of the gene for further analysis.

Amplification of genomic DNA

All primers used for the amplification of genomic DNA were designed on the web based software Primer3 (Rozen and Skaletsky 2000; Petit *et al.* 2003). The quality of the primers was determined in terms of melting points, self-dimerization and/or formation of hairpins (secondary structures) with the software GeneRunner® (1994, Hastings Software, Inc.). Primer sequences are given in Table I-1a and I-1b.

PCRs were carried out by applying Qiagens's HotStarTaq® MasterMix (Qiagen, Hilden, Germany) on a PTC-200 Peltier Thermal Cycler (MJResearch, Inc., Waltham, Massachusetts) and under the following thermal cycling conditions: an initial denaturation step at 95°C for 15 minutes, a set of 35 repeats of the next three steps: 95°C for 1 minute, T_a (annealing temperature of the corresponding primer pair) for 45 seconds and 72°C for 1 minute, and a final extension step at 72°C for 20 minutes. The T_a for each primer pair used is provided in Table I-1a and I-1b. The size, quality and quantity of the amplification products were assessed by electrophoresis on a 1.5% agarose gel.

For sequencing, the PCR products were excised from the gel and purified using the GENECLEAN® Kit (Qbiogene Inc., Carlsbad, CA), following the manufacturer's

instructions. The quality and quantity of the eluted DNA was assessed by electrophoresis on a 1.5% agarose gel.

Table I-1a: Upper part: *Populus trichocarpa* specific primers and corresponding T_a, lower part: *Q. petraea* specific primers and corresponding T_a for the amplification of a $NADP^+$ IDH gene (F or f: foreword primers, R or r: reverse primers, a and b: alternatives)

Primer	Oligo-sequence 5'-3'	T_a in °C
1F	aat ccc atc gtt gaa atg ga	52
1Ra	tcc cat ttg gac tct tcc ac	
1Rb	cct atg cag att ggc ttg gt	
2F	acc aag cca atc tgc ata gg	48
2R	ccc ctt cac ttt tga gag ca	
3Fa	ggg gtt atg tat ggg cat gt	52
3Fb	ttc aaa gga gct gga ggt gt	
3Ra	tgc cct taa gct cct cag cta	
3Rb	gaa cac ggt aat ggc gag taa	
Q1f	gatggtaattattaattcacaccgttc	55.2
Q1r	cat tgg tgg cct cac gat tag tga g	
Q2f	tcc ctt tgt gga gtt gga tat caa g	60.2
Q2r	gtt gaa ctc ctt gac acg agc ttc	
Q3f	gtg tgc aac tat aac tcc agg tac	60
Q3r	cat tat gtc cat tgg gta ctg cag	
Q4f	gtg atc agt aca ggg caa ctg	54.5
Q4r	gcc tca att gga atg acc ca	
Q5f	aac ttt acg ggt gct gga ggt gta g	61
Q5r	gca agc cca tac ata acc tcc ttc a	
Q6f	gat gat atg gtt gct tat gcc atg	58
Q6r	ggt ttc acc acc ttt ctg atg gac	
Q7f	gac tat tga agc tga agc agc cca tg	62
Q7r	taa gct cct cag cta cag cat caa tg	

Table I-1b: *Q. robur* specific primers for the NAD^+ IDH gene and corresponding T_a (F or f: foreword primers, R or r: reverse primers)

Primer	Oligo sequence 5'-3'	T_a in °C
IDHq1F	cat cct aac ctc acc cca ag	55
IDHq1R	agc cca cct ttc aga cac ac	

DNA cloning and sequencing

Direct sequencing of PCR purified DNA fragments and sequencing of cloned PCR fragments was applied for different purposes. For cloning, the PCR amplicons were ligated into the pCR®2.1-TOPO® vector, using the TA-cloning kit (Invitrogen Carlsbad, CA). Positive *E. coli* TOP10 (Invitrogen Carlsbad, CA) clones were picked from kanamycin (0.1mg/ml) selective LB agar plates, treated with X-gal in DMF (dimethyloformamide) and cultured in LB medium overnight at 37°C. The plasmid DNA was extracted using the QuickLyse Miniprep Kit (Qiagen, Hilden, Germany). The PCR fragment insertion was determined by a digestion of the plasmid DNA with the *Eco*RI restriction enzyme (Fermentas GmbH, Leon-Rot, Germany) according to Sambrook *et al.* (1989) and incubated at 37°C for 2-3 hours or overnight.

The sequencing reaction was performed for one, two or three clones of each PCR amplicon in both directions using the BigDye chemistry using the BigDye chemistry (Applied Biosystems, Foster City, CA) with the M13 forward and reverse primers in the case of plasmid DNA, whereas the corresponding primer (forward or reverse) for each PCR amplicon was used in the case of direct sequencing. The sequencing thermal cycling profile was 1 minute at 96°C followed by a set of 35 cycles of 10 seconds at 96°C, 10 seconds at 45°C and 4 minutes at 55°C, performed on a Peqlab primus 96 thermocycler (Peqlab Biotechnologie GmbH, Erlangen, Germany). Sequencing reactions were purified by ethanol precipitation (Sambrook *et al.* 1989), and were loaded on an ABI Prism® 3100 genetic analyzer (Applied Biosystems, Foster City, CA) in an HiDi™ Formamid (Applied Biosystems, Foster City, CA) elution. When necessary, the sequencing reactions were purified using the NucleoSEQ clean-up columns (Macherey-Nagel GmbH, Düren, Germany). The capillary electrophoresis was conducted with 36cm long capillaries on a POP6® polymer and buffer with EDTA (ethylenediamine tetra-acetic acid) (Applied Biosystems, Foster City, CA) under the Rapidseq36_POP6 Default Run Module (Applied Biosystems, Foster City, CA): Run temperature 55°C, run voltage 15 kV, whereas the injection time was set to 22 seconds, and the injection voltage to 1kV. Alternatively, an ABI Prism® 3130xl genetic analyzer (Applied Biosystems, Foster City, CA) was used, with 36cm long capillaries on a POP7® polymer and buffer with EDTA (Applied Biosystems, Foster City, CA) under slightly modified conditions: oven temperature 60°C, run voltage 8.5kV, injection time 18 seconds and injection voltage up to 1.2kV. The total run time was set to

2780 seconds, whereas the run time was increased to 4000 seconds in cases of long runs up to reads of 700bp. The sequencing data were base called, trimmed and displayed for further analysis by the Sequencing Analysis Software v3.7 and v5.3.1 respectively (Applied Biosystems, Foster City, CA).

Genome walking

To aim the sequence of a NAD^+ dependent IDH gene, we applied a PCR based method named "primer walking" (Siebert *et al.* 1995) for obtaining an unknown DNA sequence adjacent to a known sequence, by using the Genome Walker™ Universal Kit (Clontech Laboratories, Inc.) following the manufacturer's protocol. This technique requires DNA of high quality and quantity. For that reason the DNA quality and quantity were tested against λ-DNA (Roche Holding GmbH, Germany). The genomic DNA of a *Q. petraea* sample was used as template for the construction of adaptor-ligated "libraries" using the adaptor primers provided by the kit, and the restriction enzymes: *Dra*I, *Eco*RI, *Eco*RV, *Pvu*II and *Stu*I provided by the Genome Walker™ Universal Kit or by Fermentas (GmbH, Leon-Rot, Germany). Two different gene specific primers (GSP1 and GSP2) were specifically designed as described above to locate in consecutive positions of the known sequence published in the INRA Quercus EST database accessible via the EVOLTREE organisation (www.evoltree.org). Each adaptor-ligated "library" was applied to "suppression" PCR on a Perkin-Elmer GeneAmp® PCR System 2400 (Perkin Elmer Corp., Foster City, CA) using the GSP1 gene specific primer and the reagents provided in the Genome Walker™ Universal Kit (Clontech Laboratories, Inc.) under the following thermal conditions: 7 cycles of 94°C for 12 seconds and 72°C for 3 minutes, followed by 32 cycles of 94°C for 12 seconds and 67°C for 3 minutes, and a final step of 67°C for 7 minutes. The amplification products were visualized on a 1.5% agarose gel. "Nested" PCR was performed on a Perkin-Elmer GeneAmp® PCR System 2400 (Perkin Elmer Corp., Foster City, CA), using the GSP2 gene specific primer on the amplification reaction of the first PCR (1:50 diluted), and PCR reagents provided by the Genome Walker™ Universal Kit (Clontech Laboratories, Inc.) under the thermal cycling profile: 5 cycles of 25 seconds at 94°C and 3 minutes at 72°C, followed by 20 cycles of 25 seconds at 94°C and 3 minutes at 67°C. An additional step of 7 minutes at 67°C finalized the reaction. The "nested" PCR products were electrophoretically separated on a 1.5% agarose gel. Fragments showing the strongest signal intensities were

excised out of the gel and purified. The quality and quantity of the eluted DNA was assessed by electrophoresis, and visual control of the ethidiumbromide stained gel. The correct gene sequence was confirmed by BLASTn and tBLASTX homology analysis against the EMBL standard plant database. A second round of "primer walking" was conducted based on the downstream sequence obtained by the first genome walking. The gene specific primers for the "suppression" and "nested" PCRs in the second round of primer walking were GSPdra1 and GSPdra2 respectively. The sequences of all the gene specific primers used are given in Table I-2.

Table I-2: Gene specific primers used for the "Genome walking" technique

Gene specific primer	Oligo sequence 5'-3'
GSP1	gag ctg gat ttg tat gct tcg ttg gtc
GSP2	cct acg agg cac gaa aat ggg ata tgt g
GSPdra1	aat gca gta cca aca gtg tga tct tc
GSPdra2	cct ata gta agg gcg aat tct gca gat a

Sequence analysis

DNA sequences used in this study were manually verified and edited for base-calling errors. Contigs were assembled using the computer software CodonCode Aligner (CodonCode Corp., Dedham, MA), or Sequencher v4.8 (Gene Codes, Ann Arbor. MI). Multiple sequence alignments were performed using the ClustalW algorithm (Thompson *et al.* 1994) and adjusted manually by using BioEdit (Hall 1999). The sequences were aligned against public databases using BLASTx or tBLASTx (Altschul *et al.* 1997) in case of protein or translated nucleotide search respectively (using a translated nucleotide query) or pBLAST (Altschul *et al.* 1997; Altschul *et al.* 2005) in case of protein search (using a deduced protein query).

For both gene loci the general description of nucleotide polymorphism $\theta=4N_e\mu$ was calculated as θ_w (Watterson 1975), based on the number of segregating sites. The pairwise nucleotide diversity π was also calculated as the average number of nucleotide differences per site between two sequences (Nei 1987). Species' differentiation was estimated from haplotypic data as F_{st} (Hudson *et al.* 1992a) and G_{st} (Nei 1973) supplemented by the nucleotide based statistics S_{nn} (the nearest-neighbor statistic) (Hudson 2000) and K_{st}*, a

weighted measure of the ratio of the average pairwise differences within populations to the total average differences (Hudson *et al.* 1992b). Calculations of nucleotide diversity and differentiation were performed with the software DNAsp (Librado and Rozas 2009). The analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992) from haplotypic data was conducted with the software Arlequin v3.5 (Excoffier and Lischer 2010).

The degree of linkage disequilibrium (LD) or non random association between nucleotide variants at different polymorphic sites was estimated also with DNAsp (Librado and Rozas 2009). For this analysis, alignment gaps or polymorphic sites segregating for three or four nucleotides were excluded. The LD analysis was performed for the parsimony – informative sites (sites that segregate for only two nucleotides that are present at least twice). Moreover, the squared correlation (r^2) (Hill and Robertson 1968) of allele frequencies of all pairs of parsimony informative sites was plotted against the average number of nucleotides that separates them, in a scatter graph. The linear regression equation that best fits the data was also calculated. Moreover, the average ZnS statistic (Kelly 1997) of r^2 over all pairwise comparisons was estimated. Additionally, the neutrality test of Tajima's *D* (Tajima 1989) was performed to examine the fit of nucleotide polymorphism data to the neutral equilibrium model. The *D* test is based on the differences between the number of segregating sites and the average number of nucleotide differences. Coalescent simulations for the significance of Tajima's *D* were performed with the software DNAsp (Librado and Rozas 2009).

Results

Identification of NADP$^+$ Isocitrate dehydrogenase

In order to confirm that the reference sequence we obtained was indeed a NADP$^+$ IDH gene locus, different homology tests were applied. The full reference nucleotide sequence was tested for homology by means of tBLASTx (Altschul *et al.* 1997) against the higher plant nucleotide collection database of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and resulted in hits of *E*-value = 0.0 with NADP$^+$ specific isocitrate dehydrogenase of *Ricinus communis*, *Passiflora edulis* and *Eucaliptus globulus*. Continuously, after defining the putative exon and intron positions by aligning the reference sequence with NADP$^+$ IDH sequences of other plants, only the nucleotide transcript sequence was used for a tBLASTx homology test against the same database. This analysis resulted as well in hits of *E*-value = 0.0 with cytosolic NADP$^+$ IDH of different plants.

Finally the transcript derived from the nucleotide sequence was tested for homology against the non-redundant SwissProt protein database (Boeckmann *et al.* 2003) with pBLAST (Altschul *et al.* 1997; Altschul *et al.* 2005). This homology test resulted in an *E*-value = 0.0 and query coverage of 100% with the isocitrate dehydrogenase protein ($NADP^+$ specific) of *Populus trichocarpa* (accession number: XP_002305928.1). The ClustalW alignments of the best hits of the BLASTx test are shown in Appendix 1a.

The reference sequence of the $NADP^+$ IDH locus that was obtained represents an almost complete sequence of the gene. Coding and non coding regions were defined according to the splicing GT-AG rule (Bon *et al.* 2003): 15 exons (the first and the last only partially) and 14 introns were identified (Figure I-1). The total size of the sequence was 3481bp, whereas the total size of the coding region was found 1195bp. The organization in exons and introns is illustrated in Figure I-1. The gene was PCR amplified from genomic DNA of each sample in seven overlapping parts (here referred to as loci): Loci 1-7, as shown in Figure I-1. The reference $NADP^+$ IDH sequence is provided in Appendix 2a.

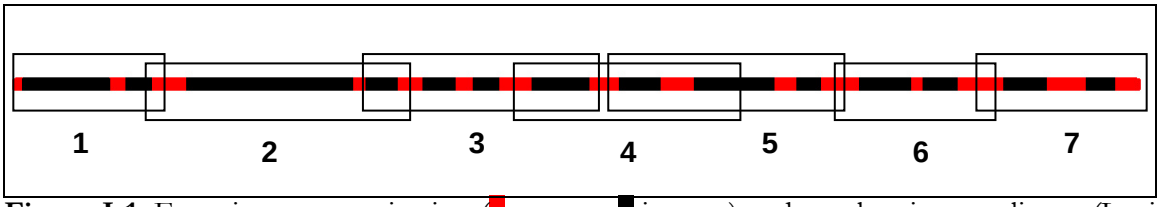

Figure I-1: Exon-intron organization (exons – introns) and overlapping amplicons (Loci 1-7 given as boxes) of the $NADP^+$ isocitrate dehydrogenase gene sequence (3481bp) identified in *Q. robur*

Five $NADP^+$ isocitrate dehydrogenase sequences per species were obtained for each of the loci. Locus 5 revealed unusually high levels of variation among the five sequences per species that were obtained and therefore was excluded from any further analysis to avoid incorporation of paralogous sequences in our analyses. Singletons and microsatellite motifs were also excluded from any further analysis to avoid incorporation of PCR or sequencing errors.

Identification of NAD^+ isocitrate dehydrogenase

To confirm that the obtained sequence by the genome walking technique was indeed an NAD^+ IDH sequence, different homology tests were conducted. Homology analysis with

tBLASTx of the genomic nucleotide reference sequence against the higher plant nucleotide collection database of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) resulted in a match with *Populus trichocarpa* NAD^+ isocitrate dehydrogenase mRNA, predicted protein (accession number: XM_002310661.1) (*E*-value= $3e^{-98}$). The obtained reference 2080bp sequence represented only a partial gene sequence of the NAD^+ IDH gene. Within this fragment a putative single intron (position 560-1949) was identified. The sequence used for further analysis was limited to a total size of 346bp covering the largest part of the first exon of the gene. The protein sequence derived from the transcript sequence of the NAD^+ IDH locus that was further analyzed was tested for homology with pBLAST against the non-redundant SwissProt sequences database (Boeckmann *et al.* 2003). This analysis resulted in a best match of *E*-value=$4e^{-39}$ with mitochondrial isocitrate dehydrogenase (NAD^+) regulatory subunit 1, of *Arabidopsis thaliana* (accession number: Q8LFC0.2). The ClustalW alignments of the best hits occurred by a BLASTx test with the reference sequence are shown in Appendix 1b.

Five NAD^+ isocitrate dehydrogenase sequences per species were obtained in one amplicon. The reference NAD^+ IDH sequence is provided in Appendix 2b.

Nucleotide diversity

$NADP^+$ isocitrate dehydrogenase

Descriptive statistics for nucleotide variation within each species but also in the pooled sample set are shown in Table I-3a, both for each amplicon separately and for the total sequence (unphased data). The 20 sequences per amplicon gave a total of 88 polymorphic sites; whereas when analyzed as a whole, a total of 83 polymorphic sites were detected (five sites are located on overlapping regions of the amplicons and therefore are analyses in both amplicons). The values of θ_w and π_{tot} were in general similar. The total nucleotide diversity (π_{tot}) in the pooled samples was estimated between 0.01295 and 0.00324 for loci 1 and 4 respectively and 0.00816 for the total sequence. The nucleotide diversity of silent sites (synonymous and non-coding sites) was estimated between 0.00583 and 0.01368 for the same pair of amplicons. The ratio of non-synonymous and silent nucleotide diversity was generally low and only in four cases larger than 0.5 (*Q. robur* at locus 1, *Q. petraea* at locus 2 and *Q. frainetto* at loci 1 and 2). Interestingly, the same ratio was notably lower when calculated for the total sequence (combination of all loci in one sequence) in all four species and in the pooled samples. The average haplotypic diversity over all loci and species was

found 0.9 corresponding on average to 4.08 haplotypes per species per locus. In particular, for most species and for the pooled sample set locus 4 revealed the lowest haplotypic diversity whereas in *Q. robur* locus 7 revealed the lowest value among all loci and species. Considering each species separately, *Q. pubescens* exhibited the lowest levels of nucleotide diversity in the total sequence and in most loci with the exception of loci 1 and 4 from one side and 7 from the other where *Q. petraea* and *Q. robur* had slightly lower values for π_{tot}, respectively.

NAD^+ isocitrate dehydrogenase

Descriptive statistics for nucleotide variation of the sequence of the NAD^+ locus are provided in Table I-3b. The 20 sequences that were analyzed gave a total of eight polymorphic sites resulting in 13 haplotypes. All measurements of nucleotide variation or diversity (θ_w or π) showed the lowest values in *Q. robur* and the highest in *Q. frainetto*. The ratio of non-synonymous to silent nucleotide diversity was found generally low for the pooled samples (0.26281). Interestingly, this ratio is close to 1 for *Q. robur* and particularly higher than the values of the other species, the lowest value occurring in *Q. pubescens* (0.15678).

The total nucleotide diversity π_{tot} that was observed in all the investigated species, over all loci for both $NADP^+$ and NAD^+ IDH genes was 0.007503, whereas the silent nucleotide diversity π_{sil} was 0.010047.

Table I-3a: Summary statistics of nucleotide variation for the NADP$^+$ IDH gene, for each locus separately and for the total sequence per species and in the pooled dataset (continued on the next page)

Species	*Locus*	*n*	*S*	*E*	*h*	H_d	S_{silent}	$S_{synonymous}$	$S_{non\ synonymous}$	θ_w	π_{tot}	π_{silent}	$\pi_{synonymous}$	$\pi_{non\ synonymous}$	$\pi_{non\ synonymous}/\pi_{silent}$
Q. robur	1	373	14	14	5	1	330.9	8.9	39.1	0.01802	0.01769	0.01874	0.04507	0.01026	0.54749
	2	738	12	12	5	1	610.5	37.5	124.5	0.0078	0.00732	0.00885	0.016	0	0
	3	733	6	6	5	1	521	60	210	0.00393	0.00437	0.00614	0	0	0
	4	637	2	3	3	0.8	434.33	56.33	198.67	0.00151	0.0022	0.00332	0	0	0
	6	489	7	7	4	0.9	321	51	168	0.00687	0.00695	0.00872	0	0.00357	0
	7	533	5	5	2	0.4	343.6	59.6	186.4	0.0045	0.00375	0.00465	0	0.00215	0.46237
	total	3027	45	46	5	1	2192.83	240.83	827.17	0.00714	0.00764	0.00894	0.00415	0.00169	0.18904
Q. petraea	1	372	7	7	4	0.9	329.83	8.83	39.17	0.00903	0.00806	0.01019	0	0	0
	2	739	29	31	5	1	610.83	37.83	127.17	0.01884	0.01827	0.0221	0.01586	0	0
	3	730	14	14	4	0.9	518	60	210	0.00921	0.00795	0.0112	0.01333	0	0
	4	640	4	4	3	0.7	437.2	56.2	198.8	0.003	0.0025	0.00275	0.00714	0.00201	0.73091
	6	490	7	7	4	0.9	321	51	168	0.00686	0.00653	0.00814	0	0.00352	0.43243
	7	534	7	8	5	1	344.6	59.6	186.4	0.00629	0.00749	0.01045	0.01679	0.00215	0.20574
	total	3035	57	59	5	1	2195.63	242.63	834.37	0.00901	0.00896	0.01175	0.00989	0.00168	0.14298
Q. pubescens	1	374	11	11	5	1	331.83	8.83	39.17	0.01412	0.01444	0.01627	0.04528	0	0
	2	725	4	4	5	1	608.17	34.17	112.83	0.00265	0.00248	0.00296	0	0	0
	3	733	2	2	3	0.8	521.07	60.07	209.93	0.00131	0.00136	0.00192	0.00999	0	0
	4	526	5	5	3	0.7	345.2	50.2	177.8	0.00456	0.0038	0.00464	0.008	0.00225	0.48491
	6	489	5	5	4	0.9	319	51	168	0.00491	0.0045	0.0069	0	0	0
	7	534	8	8	4	0.9	344.5	59.5	186.5	0.00719	0.00674	0.00929	0.01345	0.00214	0.23036
	total	2907	32	32	5	1	2099.93	230.93	792.07	0.00528	0.00516	0.00695	0.01472	0.0005	0.07194
Q. frainetto	1	391	11	11	5	1	346.23	10.23	43.77	0.0135	0.01176	0.01213	0	0.00916	0.75515
	2	646	21	21	4	0.9	538.83	32.83	105.17	0.0156	0.013	0.01559	0.01218	0	0
	3	696	11	13	5	1	498	56	196	0.00759	0.00776	0.01084	0	0	0
	4	641	6	6	3	0.7	437.93	88.93	199.07	0.00449	0.00468	0.00548	0.01071	0.00302	0.55109
	6	489	9	9	4	0.9	319	51	168	0.00883	0.00818	0.01129	0	0.00238	0.21081
	7	516	8	8	4	0.9	326.5	59.5	186.5	0.00744	0.00698	0.0098	0.01345	0.00214	0.21837
	total	3030	63	64	5	1	2189.67	242.67	837.33	0.00998	0.00901	0.01183	0.01154	0.00167	0.14117

Species	*Locus*	*n*	*S*	*E*	*h*	H_d	S_{silent}	$S_{synonymous}$	$S_{non\ synonymous}$	θ_w	π_{tot}	π_{silent}	$\pi_{synonymous}$	$\pi_{non\ synonymous}$	$\pi_{non\ synonymous}/\pi_{silent}$
Pooled	1	370	17	17	17	0.979	327.87	8.87	39.13	0.01295	0.01263	0.01368	0.02141	0.00486	0.35526
	2	627	26	26	15	0.958	530.5	29.5	93.5	0.01169	0.00999	0.01181	0.01499	0	0
	3	693	19	21	12	0.911	495.02	195.98	56.02	0.00773	0.0057	0.00797	0.00695	0	0
	4	522	6	7	6	0.842	341.17	177.83	50.17	0.00324	0.00457	0.00583	0.00789	0.00222	0.38079
	6	485	11	11	8	0.9	317.67	50.67	165.33	0.00639	0.00666	0.00893	0	0.00239	0.26764
	7	515	9	9	10	0.905	325.55	59.55	186.45	0.00493	0.00544	0.00744	0.01255	0.00203	0.27285
	total	2868	83	86	20	1	2066.93	230.93	792.07	0.00816	0.00845	0.01118	0.01174	0.00148	0.13238

Table I-3b: Summary statistics of nucleotide variation for the NAD^+ IDH gene

Species	*S*	*E*	*h*	H_d	S_{silent}	$S_{synonymous}$	$S_{non\ synonymous}$	θ_w	π_{tot}	π_{silent}	$\pi_{synonymous}$	$\pi_{non\ synonymous}$	$\pi_{non\ synonymous}/\pi_{silent}$
Q. robur	3	3	4	0.9	125.57	75.57	218.43	0.00417	0.00464	0.00478	0.00794	0.00458	0.958159
Q. petraea	6	6	5	1	125.53	75.53	218.47	0.00835	0.00754	0.01434	0.01854	0.00366	0.25523
Q. pubescens	5	5	5	1	125.43	75.43	218.57	0.00696	0.00812	0.01754	0.01591	0.00275	0.156784
Q. frainetto	7	7	4	0.9	125.63	75.63	218.37	0.00974	0.01101	0.02229	0.03174	0.00458	0.205473
Pooled	8	8	13	0.947	125.54	75.54	218.46	0.00654	0.00807	0.01522	0.01923	0.00400	0.262812

n: length of Locus in bp
S: segregating sites
E: total number of mutations
H: number of haplotypes
H_d: haplotype diversity
S_{silent}: Number of silent sites, $S_{synonymous}$: number of synonymous sites, $S_{non\ synonymous}$: number of non-synonymous sites
θ_w: total nucleotide polymorphism (Watterson, 1975)
π_{tot}: total nucleotide diversity, π_{silent}: silent nucleotide diversity, $\pi_{synonymous}$: synonymous nucleotide diversity, $\pi_{non\ synonymous}$: non-synonymous nucleotide diversity (Nei, 1987)

Species differentiation

The genetic differentiation between populations as estimated with AMOVA was very low for both genes: F_{st} = 0.01958 (not significant) and F_{st} = 0.03800 (not significant) respectively, as shown in Table I-4a for $NADP^+$ IDH and Table I-4b for NAD^+ IDH, based on haplotypic data

Table I-4a: Among and within species AMOVA analysis from haplotypic data of the total $NADP^+$ IDH sequences

	d. f.	Sum of squares	Variance components	Percentage of variation
Among species	3	56.300	0.34083Va	1.96
Within species	16	273.000	17.06250Vb	98.04
Total	19	3.293.000	1.740.333	

Table I-4b: Among and within species AMOVA analysis from haplotypic data of the NAD^+ IDH sequences

	d. f.	Sum of squares	Variance components	Percentage of variation
Among species	3	4.850	0.05333Va	3.8
Within species	16	21.600	1.35000Vb	96.2
Total	19	26.450	140.333	

Based on nucleotide sequence data information of both $NADP^+$ and NAD^+ IDH genes, only locus 4 of the first gene exhibited significant difference among species applying both the nearest neighbor statistic significance test (S_{nn}) and K_{st}*, a weighted measure of the ratio of the average differences within populations (here: species) to the total average differences, after 10000 permutations (Table I-5). Additionally, the G_{st} and F_{st} measures showed low differentiation among populations, with only locus 4 of $NADP^+$ IDH having higher values: 0.13303 and 0.19571 respectively. Detailed statistic matrices of the estimation of the pairwise genetic differentiation G_{st} (Nei 1973) and F_{st} (Hudson *et al.* 1992b) for each locus separately are provided in Appendix 3.

Table I-5: Genetic differentiation among all species for each locus of NADP$^+$ IDH and for NAD$^+$ gene (significance test after 10000 permutations: *:0.01<P<0.05; **:0.001<P<0.01)

Locus		Pool			
		G_{st}	F_{st}	K_{st}*	S_{nn}
NADP$^+$ IDH	1	0.00383	-0.06214	-0.02878	0.175
	2	-0.01695	-0.03809	-0.00744	0.02476
	3	-0.01509	0.03104	0.01721	0.285
	4	0.13303	0.19571	0.16016*	0.42143**
	6	0.00000	0.01235	0.00191	0.19167
	7	0.11111	0.11047	0.0912	0.32262
	Total	0.00000	0.03017	0.00877	0.17500
NAD$^+$ IDH		-0.00264	0.03800	0.01545	0.26662

Statistical tests for neutrality

To test the fit of nucleotide polymorphism to the neutral mutation model Tajima's *D* statistic was estimated for each species separately and for the pooled sample set, at each locus. For the NADP$^+$ IDH gene, in almost all species and all loci, *D* got a negative value, suggesting an excess of low frequency substitutions. Among all species, *Q. frainetto* exhibited the strongest negative value of *D* in the total sequence. On the contrary, for NAD$^+$ IDH the single species (with the exception of *Q. petraea*) and the pooled dataset revealed a positive *D*. However, even though close to them, no observation of Tajima's *D* was found to be exceeding the 95% confidence intervals of the computed simulations using the coalescent algorithm. It is worth taking note of the different trend of the Tajima's *D* at different loci both at the species level and in the pooled sample, as well as the different trend of *D* between the two genes. Tajima's D values for all loci and species are given in Table I-6.

Table I-6: Tajima's *D* values for the neutrality test with NADP$^+$ and NAD$^+$ IDH gene sequences

Species	NADP$^+$ IDH							NAD$^+$ IDH
	1	2	3	4	6	7	Total	
Q. robur	-0.13015	-0.45202	0.76369	-0.17475	0.08298	-1.12397	-0.36866	0.69900
Q. petraea	-0.74682	-0.69393	-0.99781	-1.09380	-0.33192	0.29358	-0.29908	-0.66823
Q. pubescens	0.16354	-0.41017	0.24314	-1.12397	-0.56199	-0.44037	-0.17549	1.12397
Q. frainetto	-0.92693	-1.23407	-0.97762	0.28638	-0.52640	-0.44037	-0.84270	0.91278
Pool	-0.09219	-0.56431	-1.27131	0.68509	0.11499	0.35676	-0.39536	0.79104

Additionally, the ratio of the non-synonymous to silent substitutions or else the ratio of the non-synonymous to the silent nucleotide diversity has been as well considered as a test against neutrality (s. Tables I-3a, I-3b).

Linkage disequilibrium

Sequences of both genes were analyzed as haploid data for linkage disequilibrium both for each species separately and for the pooled sample set, as the levels of differentiation were found low. The low sample size at species level and the short length of each locus separately resulted in many single loci in no or insufficient pairwise comparisons of the informative sites. Therefore, further analyses were based on the pooled data for the total $NADP^+$ IDH sequence and the NAD^+ sequence. Linkage disequilibrium (LD) was measured as the squared allele-frequency correlation (r^2) in 73 parsimony informative sites by 2628 pairwise comparisons for the $NADP^+$ IDH gene and in eight parsimony informative sites by conducting 28 pairwise comparisons. In the first gene, among the total number of pairwise comparisons, 274 sites were found significantly linked according to Fisher's exact test (Figure I-2) but there was no evidence of sites in linkage disequilibrium after Bonferroni's procedure, whereas 467 sites were found to be significantly linked using the χ^2 test from which 75 sites were still found to be in linkage disequilibrium after Bonferroni's procedure. In the second gene, among the three significant pairwise comparisons (tested with the χ^2 test) only one was still found significant after Bonferroni's procedure. The detailed information about the significantly linked sites after Bonferroni's adjustments their physical distance and the corresponding LD estimates *D* (Lewontin and Kojima 1960), *D′* (Lewontin 1964) and R (Hill and Robertson 1968) are provided in Appendix 4a and b, sorted in a descending order from the larger to the shorter nucleotide distances. In general, r^2 values are low. The linear regression equations that could best describe the r^2 values for $NADP^+$ IDH and NAD^+ IDH fragments plotted against the nucleotide distance were $y=0.1755-0.0531x$ and $y=0.1250-0.2969x$ respectively (x measured in kb) (Figures I-2, I-3)

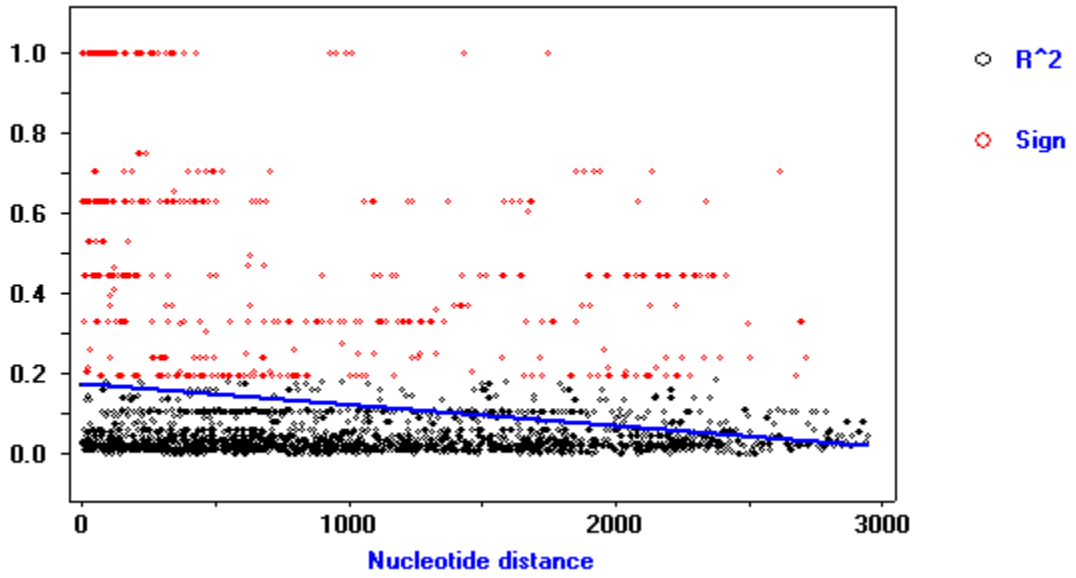

Figure I-2: Scatterplot of the squared gametic frequency correlation (r^2) (vertical axis), against the nucleotide distance of all pairs of the 73 parsimony informative sites (horizontal axis) by 2628 pairwise comparisons for the $NADP^+$ IDH gene. In red are marked the significant r^2 values (χ^2 test). The blue line represents the linear regression equation that best fits the data: y=0.1755-0.0531x

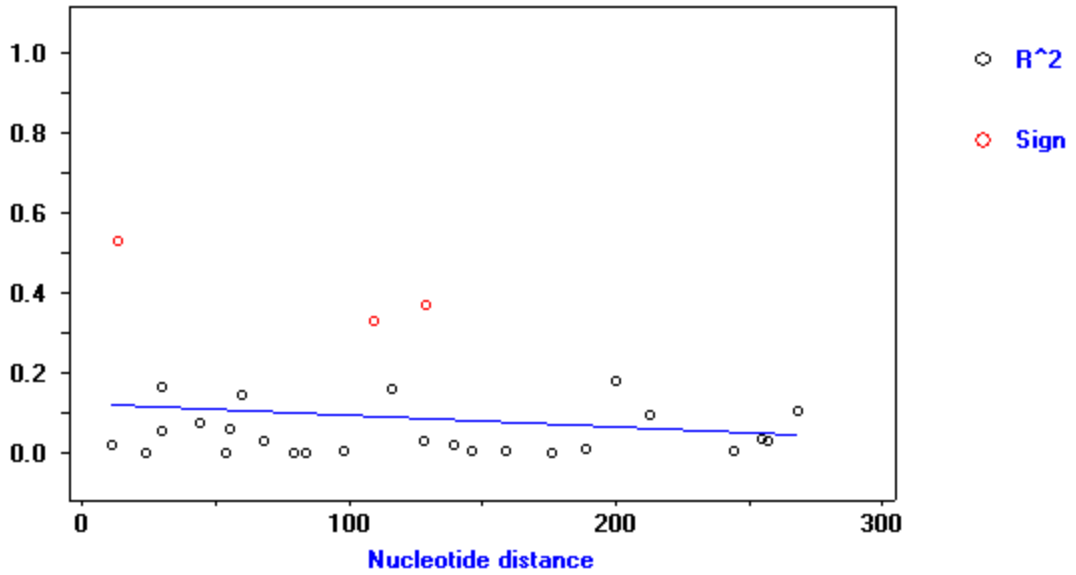

Figure I-3: Scatterplot of the squared gametic frequency correlation (r^2) (vertical axis), against the nucleotide distance of all pairs of the 8 parsimony informative sites (horizontal axis) by 28 pairwise comparisons for the NAD^+ IDH gene. In red are marked the significant r^2 values (χ^2 test). The blue line represents the linear regression equation that best fits the data: y=0.1250-0.2969x

The ZnS average of r^2 overall pairwise comparisons was found 0.4444, 0.3618, 0.3283 and 0.1937 for *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* respectively for the total sequence of $NADP^+$ IDH. For NAD^+ IDH the ZnS was estimated 0.4444 for *Q. robur*, 0.1667 for *Q. pubescens* and 0.3889 for *Q. frainetto*. In the case of *Q. petraea*, no pairwise comparisons were possible.

Discussion

Nucleotide diversity and species differentiation

The present study is an additional contribution to increase the limited existing information about nucleotide diversity in forest tree species, and one of the first studies of this kind in European oak species. In particular, no previous study has reported levels of nucleotide polymorphisms for the species: *Q. robur*, *Q. pubescens* and *Q. frainetto*.

The similarity between the values of θ_w and π_{tot} shows that the assumptions of the infinite site model were not violated. The total nucleotide diversity that was observed in the present study in all the investigated species, over all loci for both $NADP^+$ and NAD^+ IDH genes was 0.007503. This value is comparable though slightly higher with the reported nucleotide variation in the Asian *Q. mongolica* var. *crispula* (Quang *et al.* 2009), with that examined in the very first study that characterized nucleotide variation levels within the genus *Quercus*, *Q. crispula* (Quang *et al.* 2008) and also to the ones reported for *Q. petraea* in investigations of candidate genes for adaptive traits (Vornam *et al.* 2007) such as the timing of bud burst (Derory *et al.* 2010), despite our conservative analysis by excluding all singletons and microsatellite motifs from any further analysis. Comparing the present results to the results obtained by studies of different broad-leaved tree genera, our study shows levels of nucleotide variation higher than those found in the *BpMADS2* gene in *Betula pendula* (π= 0.0039-0.0045) (Järvinen *et al.* 2003), but comparable to the findings of Ingvarsson (2005) for *Populus tremula* analyzing an *Alcohol dehydrogenase* locus (ADH) (π= 0.0090). In the same study four other genes (CI-1, G3PDH, GA20ox1 and TI-3) resulted in higher levels of nucleotide diversity (π= 0.0184, 0.0176, 0.0113, 0.0216 respectively). However, the same author analyzing 76 genes distributed throughout the genome derived form EST sequences of *Populus tremula* (Ingvarsson 2007; Ingvarsson 2008) found much lower levels of diversity (π= 0.0042), an artifact attributed to the sequencing strategies. Likewise, our findings were higher than those revealed in the analysis of the *PtABI1B* gene locus of *Populus tremula* (Garcia and

Ingvarsson 2007). In general, conifers exhibit lower nucleotide diversity than angiosperms as investigated at different gene loci and different tree genera like *Pinus*, (Dvornyk *et al.* 2002; Neale and Savolainen 2004; Pot *et al.* 2005; Pyhäjärvi *et al.* 2007; Wachowiak *et al.* 2009), *Picea* (Heuertz *et al.* 2006), or *Pseudotsuga* (Krutovsky and Neale 2005; Eckert *et al.* 2009). Among all four species investigated, *Q. pubescens* seems to have the lowest and *Q. frainetto* the highest levels of nucleotide diversity for $NADP^+$ IDH and *Q. robur* and *Q. frainetto* the highest and lowest nucleotide diversity for NAD^+ IDH respectively. These results are not in accordance with the examination of the same individuals with 6 six neutral SSR markers (but with much larger sample sizes) (Curtu *et al.* 2007), showing a potential different cause shaping variation at neutral markers and in gene loci.

In terms of silent nucleotide polymorphisms our results (π_{sil}=0.01004) are similar to those of *Q. crispula* (π_{sil}=0.00901) (Quang *et al.* 2008; Quang *et al.* 2009); but slightly lower than those reported for the European species *Q. petraea* (π_{sil}=0.0112) (Derory *et al.* 2010). However, the π_{sil} level that corresponds only to *Q. petraea* considering the total $NADP^+$ IDH sequence is in accordance to that reported by Derory *et al.* (2010).

The generally high diversity observed in our sample confirms earlier findings by use of neutral molecular markers that suggest that *Quercus* spp. are highly variable (Kremer and Petit 1993; Streiff *et al.* 1999; Mariette *et al.* 2002; Petit *et al.* 2004). High levels of diversity are most likely due to the maintenance of large population sizes, long distance gene flow and interfertility (Ducousso *et al.* 1993; Streiff *et al.* 1999; Petit *et al.* 2004). Yet, it should be kept in mind that the overall results do not reflect the large among loci variation that we observed, like it was mentioned also in other multi-locus sequence variation studies in various broad-leaved or coniferous species (Ingvarsson 2005; Krutovsky and Neale 2005; González-Martínez *et al.* 2006; Quang *et al.* 2008; Quang *et al.* 2009; Derory *et al.* 2010). In particular, among all $NADP^+$ IDH loci, locus 4 shows much lower levels of diversity over all species. This might be justified if the active site of the enzyme is located within this locus, as it is suggested e.g. by the active site of the high homologous corresponding enzyme sequence of the psychrophilic bacterium, *Desulfotalea psychrophila* (Fedøy *et al.* 2007). The high inter-loci differences show the significance of analyzing the whole sequence of a gene if conclusions are to be drawn about the total action of selection on it.

In contrast to the diversity within species which was found in general high, the diversity among species was found low, in consistency to many other sequence-data based

studies in oaks (Quang *et al.* 2008; Quang *et al.* 2009; Derory *et al.* 2010). However, it has been proven applying different kind of molecular markers that despite the low species differentiation, *Quercus* species remain distinct even within areas of sympatry (Craft and Ashley 2006; Curtu *et al.* 2007). In this study, to the low levels of genetic differentiation there was a notable exception; locus 4 of the $NADP^+$ IDH gene exhibited significant S_{nn} and K_{st} estimates but also high F_{st} value for locus 4 and 7 (0.19571 and 0.11047 respectively). These values are surprisingly high considering that *Quercus* spp. are outbreeding and wind pollinated tree species. The high levels of differentiation in this locus combined with the low levels of genetic diversity and given the possibility that the active site of the gene might be located in this locus, suggest that selection might act on it. The low sample size should be yet borne in mind.

Neutrality tests

Among several statistical tests for neutrality, the ratio between non-synonymous and silent polymorphisms and Tajima's *D*, are considered to be the most appropriate ones, given the fact that they are not solely based on singleton polymorphisms which were in this study excluded for the accuracy of the analyses. The ratio of the non-synonymous nucleotide diversity to the silent nucleotide diversity at the NAD^+ IDH gene was between 0.15678 for *Q. pubescens* and 0.958159 for *Q. robur* as a result of the surprisingly low silent diversity in this species in comparison to the others. On the contrary, for the total sequence of the $NADP^+$ IDH gene the non-synonymous to silent diversity ratio in all populations was much lower, between 0.07194 for *Q. petraea* and 0.18904 for *Q. robur*, suggesting strong purifying selection acting on the codons of this gene as it would be expected for most housekeeping proteins or enzyme genes.

Tajima's *D* was negative in most of the loci of the $NADP^+$ IDH gene, suggesting an excess of low frequency variants in the sequence. At the NAD^+ IDH gene, the values of D were in general higher. For the total sequence of the $NADP^+$ IDH gene, *Q. frainetto* exhibited the lowest values of *D*. Surprisingly, species-wide, the values of *D* varyied among the loci (e.g in *Q. pubescens* -1.1239 at locus 4 of the $NADP^+$ IDH gene whereas 1.1239 for the NAD^+ IDH gene), indicating that the trends might not be subject to demographic effects as implied in natural populations of *Populus tremula* (Ingvarsson 2005) or *Q. crispula* (Quang *et al.* 2008; Quang *et al.* 2009) but they might be an artifact of the low sample size analyzed and/or the

short region analyzed for NAD^+ IDH gene, or rather they might reflect different action of selection in different populations and loci. Indeed, at locus 4 of the $NADP^+$ IDH the strongly negative Tajima's *D* observed for *Q. petraea* and *Q. pubescens* combined with the low gene diversity at this locus and the highly significant inter-specific differentiation points out the putative adaptive significance of the gene. Of course, the non significant values for Tajima's *D* do not allow for any inference about the action of natural selection. But for that reason the suppression of singletons together with the very close values of *D* to the significance thresholds indicated by coalescent simulations should be taken into account.

Linkage disequilibrium

Gametic LD was calculated for the total sequence of $NADP^+$ IDH and the NAD^+ IDH for the pooled data despite the fact that the combined multi-locus sequences for each sample do not necessarily belong to the same allele (unphased data). For that reason, for the combined sequences the analysis gives rather an estimation of non random association of the different gametic variants. LD calculations at locus level for each species separately were not possible due to the low sample size and short length of each locus (Figure I-1) that resulted in no possible pairwise comparisons. The ZnS estimation showed much stronger non random associations for *Q. robur* than that of *Q. frainetto* and *Q. petraea* for both genes. These results are contradictory to the theoretical expectations due to the estimated higher effective population size of *Q. robur*, since *Q. pubescense* and *Q. frainetto* are at the limits of their geographic distribution. The low sample size might explain this result. Alternatively this might be a result of the populations' admixture (Wilson and Goldstein 2000; Pritchard and Przeworski 2001; Varilo *et al.* 2003; Laan *et al.* 2005) due to high levels of gene flow and hybridization in mixed stands (Curtu *et al.* 2009; Salvini *et al.* 2009). The linked sites are in their majority silent sites. But selection in linked loci has proved to have a major influence on patterns of diversity, due to selective sweeps or background selection (Aguadé *et al.* 1989). Although a great effort is being made leading to the full genome sequence of *Quercus* species, still little is known about the distribution of the recombination rate throughout their genome and the corresponding levels of LD.

Outlook

The genomic resources for species of the *Fagaceae* family and specifically of oaks are updated increasingly due to the new generation sequencing technologies and the enhanced scientific interest towards this species as a model species to study adaptation of forest trees (Gailing *et al.* 2009). The detection of loci that exhibit high inter-specific differentiation especially among sympatric species is also considered as a way to identify loci that are potentially under natural selection, as selection is argued to play an important role in the genetically distinct evolution of the closely related oak species (Petit *et al.* 2004; Curtu *et al.* 2007). In the case of the isocitrate dehydrogenase genes, the present study apart from the obtained sequences provides some first insights in the direction of their analyses as candidate adaptive genes. The differences in the patterns of intra- and inter specific nucleotide diversity and differentiation drive to the conclusion that selection might be acting differently even among different sites of the same gene. It is therefore suggested that for the candidate gene approaches their complete (or almost complete) sequences should be analyzed and possibly the gene regulatory regions should be as well investigated. Moreover, the extention of the analyses to larger sample sizes and the application of a QTL mapping approach of informative polymprophic sites within the isocitrate dehydrogenase genes could give more direct answers to the association of the genes with any adaptive trait that differentiates between the species, and will promote our understanding of their role in the evolution of the species.

References

Aas, G. (2006a). "*Quercus robur* L." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-14. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm, Ecomed,, Landsber am Lech.

Aas, G. (2006b). "*Quercus petraea* (Mattuchka) Lieblein". Enzyklopädie der Holzgewächse: Handbuch und Atlas der Dendrologie**:** 1-16. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm., Ecomed, Landsber am Lech.

Aguadé, M., N. Miyashita and C. H. Langley (1989). "Reduced Variation in the yellow-achaete-scute Region in Natural Populations of *Drosophila melanogaster*." Genetics **122**(3): 607-615.

Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucl. Acids Res. **25**(17): 3389-3402.

Altschul, S. F., J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer and Y.-K. Yu (2005). "Protein database searches using compositionally adjusted substitution matrices." FEBS Journal **272**(20): 5101-5109.

Bartha, D. (2006). "*Quercus frainetto* Ten." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-8. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm, Ecomed,, Landsber am Lech.

Bergmann, F. and H. R. Gregorius (1993). "Ecogeographical Distribution and Thermostability of Isocitrate Dehydrogenase (Idh) Alloenzymes in European Silver Fir (*Abies alba*)." Biochemical Systematics and Ecology **21**(5): 597-605.

Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucl. Acids Res. **31**(1): 365-370.

Boiffin, V., M. Hodges, S. Galvez, R. Balestrini, P. Bonfante, P. Gadal and F. Martin (1998). "Eucalypt NADP-dependent isocitrate dehydrogenase - cDNA cloning and expression in ectomycorrhizae." Plant Physiology **117**(3): 939-948.

Bon, E., S. Casaregola, G. Blandin, B. Llorente, C. Neuveglise, M. Munsterkotter, U. Guldener, H. W. Mewes, J. Van Helden, B. Dujon and C. Gaillardin (2003). "Molecular evolution of eukaryotic genomes: *hemiascomycetous* yeast spliceosomal introns." Nucleic Acids Research **31**(4): 1121-1135.

Bundock, P. C. and R. J. Henry (2004). "Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley." TAG Theoretical and Applied Genetics **109**(3): 543-551.

Bussotti, F. (2006). "*Quercus pubescens* Willd." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-10. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm., Ecomed,, Landsber am Lech.

Craft, K. J. and M. V. Ashley (2006). "Population differentiation among three species of white oak in northeastern Illinois." Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **36**(1): 206-215.

Curtu, A. L., O. Gailing and R. Finkeldey (2009). "Patterns of contemporary hybridization inferred from paternity analysis in a four-oak-species forest." BMC Evolutionary Biology **9**.

Curtu, A. L., O. Gailing, L. Leinemann and R. Finkeldey (2007). "Genetic variation and differentiation within a natural community of five oak species (*Quercus* spp.)." Plant Biology **9**(1): 116-126.

Derory, J., C. Scotti-Saintagne, E. Bertocchi, L. Le Dantec, N. Graignic, A. Jauffres, M. Casasoli, E. Chancerel, C. Bodenes, F. Alberto and A. Kremer (2010). "Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks." Heredity **104**(5): 438-448.

Ducousso, A., H. Michaud and R. Lumaret (1993). "Reproduction and gene flow in the genus *Quercus* L." Ann. For. Sci. **50**(Supplement): 91s-106s.

Dvornyk, V., A. Sirvio, M. Mikkonen and O. Savolainen (2002). "Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*." Molecular Biology and Evolution **19**(2): 179-188.

Eckert, A. J., J. L. Wegrzyn, B. Pande, K. D. Jermstad, J. M. Lee, J. D. Liechty, B. R. Tearse, K. V. Krutovsky and D. B. Neale (2009). "Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in Coastal Douglas Fir (*Pseudotsuga menziesii* var. *menziesii*)." Genetics **183**(1): 289-298.

Ellenberg, H. (1988)."Vegetation Ecology of Central Europe". 147,172 Cambridge, Cambrige University Press.

Excoffier, L. and H. E. L. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." Molecular Ecology Resources **10**(3): 564-567.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). "Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data." Genetics **131**(2): 479-491.

Fedøy, A.-E., N. Yang, A. Martinez, H.-K. S. Leiros and I. H. Steen (2007). "Structural and Functional Properties of Isocitrate Dehydrogenase from the Psychrophilic Bacterium *Desulfotalea psychrophila* Reveal a Cold-active Enzyme with an Unusual High Thermal Stability." Journal of Molecular Biology **372**(1): 130-149.

Finkeldey, R. (2001). Genetic variation of oaks (*Quercus* spp.) in Switzerland - 2. Genetic structures in "pure" and "mixed" forests of pedunculate oak (*Q. robur* L.) and sessile oak (*Q. petraea* (Matt.) Liebl.). Silvae genetica. **50:** 22-30.

Gailing, O., B. Vornam, L. Leinemann and R. Finkeldey (2009). "Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model." Physiologia Plantarum **137**(4): 509-519.

Garcia, M. V. and P. K. Ingvarsson (2007). "An excess of nonsynonymous polymorphism and extensive haplotype structure at the *PtABI1B* locus in European aspen (*Populus tremula*): a case of balancing selection in an obligately outcrossing plant?" Heredity **99**(4): 381-388.

Garcia-Gil, M. R., M. Mikkonen and O. Savolainen (2003). "Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*." Molecular Ecology **12**(5): 1195-1206.

Gömöry, D., I. Yakovlev, P. Zhelev, J. Jedinakova and L. Paule (2001). "Genetic differentiation of oak populations within the *Quercus robur*/*Quercus petraea* complex in Central and Eastern Europe." Heredity **86**(5): 557-563.

González-Martínez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale (2006). "DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L." Genetics **172**(3): 1915-1926.

Hall, D., V. Luquez, V. M. Garcia, K. R. St Onge, S. Jansson and P. K. Ingvarsson (2007). "Adaptive population differentiation in phenology across a latitudinal gradient in European Aspen (*Populus tremula*, L.): A comparison of neutral markers, candidate genes and phenotypic traits." Evolution **61**(12): 2849-2860.

Hall, T. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series **41**: 95-98.

Heuertz, M., E. De Paoli, T. Kallman, H. Larsson, I. Jurman, M. Morgante, M. Lascoux and N. Gyllenstrand (2006). "Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]." Genetics **174**(4): 2095-2105.

Hill, W. G. and A. Robertson (1968). "Linkage disequilibrium in finite populations." TAG Theoretical and Applied Genetics **38**(6): 226-231.

Hudson, R. R. (2000). "A new statistic for detecting genetic differentiation." Genetics **155**(4): 2011-2014.

Hudson, R. R., D. D. Boos and N. L. Kaplan (1992a). "A Statistical Test for Detecting Geographic Subdivision." Molecular Biology and Evolution **9**(1): 138-151.

Hudson, R. R., M. Slatkin and W. P. Maddison (1992b). "Estimation of Levels of Gene Flow From DNA Sequence Data." Genetics **132**(2): 583-589.

Huestis, D. L., B. Oppert and J. L. Marshall (2009). "Geographic distributions of Idh-1 alleles in a cricket are linked to differential enzyme kinetic performance across thermal environments." BMC Evolutionary Biology **9**.

Ingvarsson, P. K. (2005). "Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (*Populus tremula* L., *Salicaceae*)." Genetics **169**(2): 945-953.

Ingvarsson, P. K. (2007). "Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*." Molecular Biology and Evolution **24**(3): 836-844.

Ingvarsson, P. K. (2008). "Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*." Genetics **180**(1): 329-340.

Ingvarsson, P. K., M. V. Garcia, V. Luquez, D. Hall and S. Jansson (2008). "Nucleotide polymorphism and phenotypic associations within and around the *phytochrome B2* locus in European aspen (*Populus tremula*, *Salicaceae*)." Genetics **178**(4): 2217-2226.

Järvinen, P., J. Lemmetyinen, O. Savolainen and T. Sopanen (2003). "DNA sequence variation in *BpMADS2* gene in two populations of *Betula pendula*." Molecular Ecology **12**(2): 369-384.

Kawabe, A. and N. T. Miyashita (1999). "DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*." Genetics **153**(3): 1445-1453.

Kelly, J. K. (1997). "A test of neutrality based on interlocus associations." Genetics **146**(3): 1197-1206.

Kremer, A. and R. Petit (1993). "Gene diversity in natural populations of oak species." Ann. For. Sci. **50**(Supplement): 186s-202s.

Krutovsky, K. V. and D. B. Neale (2005). "Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir." Genetics **171**(4): 2029-2041.

Laan, M., V. Wiebe, E. Khusnutdinova, M. Remm and S. Paabo (2005). "X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations." European Journal of Human Genetics **13**(4): 452-462.

Le Corre, V. and A. Kremer (2003). "Genetic Variability at Neutral Markers, Quantitative Trait Loci and Trait in a Subdivided Population Under Selection." Genetics **164**(3): 1205-1219.

Lewontin, R. C. (1964). "Interaction of Selection and Linkage I. General Considerations - Heterotic Models." Genetics **49**(1): 49

Lewontin, R. C. and K. Kojima (1960). "The Evolutionary Dynamics of Complex Polymorphisms." Evolution **14**(4): 458-472.

Li, Y., M. Stocks, S. Hemmila, T. Kallman, H. T. Zhu, Y. F. Zhou, J. Chen, J. Q. Liu and M. Lascoux (2010). "Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci." Molecular Biology and Evolution **27**(5): 1001-1014.

Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-1452.

Manos, P. S., J. J. Doyle and K. C. Nixon (1999). "Phylogeny, Biogeography, and Processes of Molecular Differentiation in *Quercus* Subgenus *Quercus* (*Fagaceae*)." Molecular Phylogenetics and Evolution **12**(3): 333-349.

Mariette, S., J. Cottrell, U. M. Csaikl, P. Goikoechea, A. Konig, A. J. Lowe, B. C. Van Dam, T. Barreneche, C. Bodenes, R. Streiff, K. Burg, K. Groppe, R. C. Munro, H. Tabbener and A. Kremer (2002). "Comparison of levels of genetic diversity detected

with AFLP and microsatellite markers within and among mixed *Q. petraea* (MATT.) LIEBL. and *Q. robur* L. stands." Silvae Genetica **51**(2-3): 72-79.

Neale, D. B. and O. Savolainen (2004). "Association genetics of complex traits in conifers." Trends in Plant Science **9**(7): 325-330.

Nei, M. (1973). "Analysis of Gene Diversity in Subdivided Populations." Proceedings of the National Academy of Sciences of the United States of America **70**(12): 3321-3323.

Nei, M. (1987)."Molecular Evolutionary Genetics". New York, Columbia Univ. Press.

Nixon, K. (1993). "Infrageneric classification of *Quercus* (*Fagaceae*) and typification of sectional names." Ann. For. Sci. **50**(Supplement): 25s-34s.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. G. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N. A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Y. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman and J. Bergelson (2005). "The pattern of polymorphism in *Arabidopsis thaliana*." Plos Biology **3**(7): 1289-1299.

Palme, A. E., M. Wright and O. Savolainen (2008). "Patterns of Divergence among Conifer ESTs and Polymorphism in *Pinus sylvestris* Identify Putative Selective Sweeps." Molecular Biology and Evolution **25**(12): 2567-2577.

Pascual, M. B., J. J. Molina-Rueda, F. M. Canovas and F. Gallardo (2008b). "Spatial distribution of cytosolic NADP(+)-isocitrate dehydrogenase in pine embryos and seedlings." Tree Physiology **28**(12): 1773-1782.

Petit, R. J., C. Bodenes, A. Ducousso, G. Roussel and A. Kremer (2004). "Hybridization as a mechanism of invasion in oaks." New Phytologist **161**(1): 151-164.

Petit, R. J., U. M. Csaikl, S. Bordacs, K. Burg, E. Coart, J. Cottrell, B. van Dam, J. D. Deans, S. Dumolin-Lapegue, S. Fineschi, R. Finkeldey, A. Gillies, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. Konig, A. J. Lowe, S. F. Madsen, G. Matyas, R. C. Munro, M. Olalde, M. H. Pemonge, F. Popescu, D. Slade, H. Tabbener, D. Taurchini, S. G. M. de Vries, B. Ziegenhagen and A. Kremer (2003). "Chloroplast DNA variation in European white oaks phylogeography and patterns of diversity based on data from over 2600 populations (vol 156, pg 5, 2002)." Forest Ecology and Management **176**(1-3): 595-599.

Pot, D., L. McMillan, C. Echt, G. Le Provost, P. Garnier-Géré, S. Cato and C. Plomion (2005). "Nucleotide variation in genes involved in wood formation in two pine species." New Phytologist **167**(1): 101-112.

Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: Models and data." American Journal of Human Genetics **69**(1): 1-14.

Pyhäjärvi, T., M. R. Garcia-Gil, T. Knurr, M. Mikkonen, W. Wachowiak and O. Savolainen (2007). "Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations." Genetics **177**(3): 1713-1724.

Quang, N. D., S. Ikeda and K. Harada (2008). "Nucleotide variation in *Quercus crispula* Blume." Heredity **101**(2): 166-174.

Quang, N. D., S. Ikeda and K. Harada (2009). "Patterns of Nucleotide Diversity at the Methionine Synthase Locus in Fragmented and Continuous Populations of a Wind-Pollinated Tree, *Quercus mongolica* var. *crispula*." J Hered **100**(6): 762-770.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." **132**: 365-386.

Salvini, D., P. Bruschi, S. Fineschi, P. Grossoni, E. D. Kjaer and G. G. Vendramin (2009). "Natural hybridisation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens*

Willd. within an Italian stand as revealed by microsatellite fingerprinting." Plant Biology **11**(5): 758-765.

Sambrook, J., F. E. F. and M. T. (1989)."Molecular Cloning: A Laboratory Manual". **2** 11-34 New York, USA, Cold Spring Harbor Laboratory Press.

Scotti-Saintagne, C., S. Mariette, I. Porth, P. G. Goicoechea, T. Barreneche, K. Bodenes, K. Burg and A. Kremer (2004). "Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]." Genetics **168**(3): 1615-1626.

Semerikov, V. L. and M. Lascoux (2003). "Nuclear and cytoplasmic variation within and between Eurasian *Larix* (*Pinaceae*) species." American Journal of Botany **90**(8): 1113-1123.

Siebert, P. D., A. Chenchik, D. E. Kellogg, K. A. Lukyanov and S. A. Lukyanov (1995). "An Improved Pcr Method for Walking in Uncloned Genomic DNA." Nucleic Acids Research **23**(6): 1087-1088.

Small, R. L., J. A. Ryburn and J. F. Wendel (1999). "Low levels of nucleotide diversity at homoeologous Adh loci in allotetraploid cotton (*Gossypium* L.)." Molecular Biology and Evolution **16**(4): 491-501.

Streiff, R., A. Ducousso, C. Lexer, H. Steinkellner, J. Gloessl and A. Kremer (1999). "Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L and *Q. petraea* (Matt.) Liebl." Molecular Ecology **8**(5): 831-841.

Tajima, F. (1989). "Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." Genetics **123**(3): 585-595.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "ClustalW - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer and D. Rokhsar (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)." Science **313**(5793): 1596-1604.

Varilo, T., T. Paunio, A. Parker, M. Perola, J. Meyer, J. D. Terwilliger and L. Peltonen (2003). "The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories." Human Molecular Genetics **12**(1): 51-59.

Vornam, B., O. Gailing, R. Finkeldey, C. Collada, M. A. Guevara, A. Soto, d. M. N., S. C. González-Martínez, D. L., A. R., I. Aranda, J. Climent, M. T. Cervera, P. G. Goicoechea, L. V., E. Eveno, J. Derory, P. Garnier-Géré, A. Kremer and C. Plomion (2007). "Naturally occuring nucleotide diversity in candidate genes for forest tree adaptation: magnitude, distribution and association with quantitative trait variation." The German Plant Genome Research Program Progress Report 2004-2007**:** 116-120. GABI, Potsdam-Golm, Germany.

Wachowiak, W., P. A. Balk and O. Savolainen (2009). "Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.)." Tree Genetics & Genomes **5**(1): 117-132.

Watterson, G. A. (1975). "Number of Segregating Sites in Genetic Models without Recombination." Theoretical Population Biology **7**(2): 256-276.

Wilson, J. F. and D. B. Goldstein (2000). "Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population." American Journal of Human Genetics **67**(4): 926-935.

Zhang, N., A. Gur, Y. Gibon, R. Sulpice, S. Flint-Garcia, M. D. McMullen, M. Stitt and E. S. Buckler (2010). "Genetic Analysis of Central Carbon Metabolism Unveils an Amino Acid Substitution That Alters Maize NAD-Dependent Isocitrate Dehydrogenase Activity." PLoS ONE **5**(4): e9991.

II. Gene-associated SNP analysis at a NADP+ specific IDH enzymegene in a four-species mixed oak forest

Introduction

Over the past two decades, new molecular techniques have had an important impact on the different fields of ecology, genetics and evolution. The use of single nucleotide polymorphisms (SNPs) as a marker tool for association studies but also to address common questions in population genetics becomes increasingly popular, either as application for genome scans (Akey *et al.* 2002; Montesinos *et al.* 2009; Slate *et al.* 2009), or in candidate gene approaches targeting specific sequences, the latter being more applicable to the large genomes of most forest trees (Scotti-Saintagne *et al.* 2004; Porth *et al.* 2005; González-Martínez *et al.* 2007; Eveno *et al.* 2008; Ingvarsson *et al.* 2008). SNPs are in their majority bi-allelic markers, abundant and widespread in the coding and non-coding regions of the genome. SNP markers are therefore the markers of choice for the identification of gene loci that might be responsible for phenotypic trait variation, or involved in more complex physiological functions and for studying the dynamics of these genes in natural populations (Morin *et al.* 2004). Combined with other types of informative markers such as genomic sequences (González-Martínez *et al.* 2006) or/and amplified fragment length polymorphisms (AFLPs), isozymes or nuclear simple sequence repeats (SSRs) (Scotti-Saintagne *et al.* 2004) many authors have been dealing with exploring the null distributions of their F_{st} differentiation conditional on heterozygosity at the genome level and in specific candidate genes of forest trees and aiming to identify outlier patterns, and loci potentially under the action of selection (Akey *et al.* 2002; Scotti-Saintagne *et al.* 2004; Krutovsky 2006; Eveno *et al.* 2008).

Quercus species have turned out to be excellent model species to study adaptation of forest trees to variable environments due to their wide geographical range and the large variation of climatic and edaphic conditions that they occupy (Gailing *et al.* 2009). Regarding *Q. robur* and *Q. petraea* and despite their ecological and morphological differences the two

most widely distributed European oak species show in general very low differentiation both at nuclear and at chloroplast DNA markers (Coart *et al.* 2002; Mariette *et al.* 2002; Petit *et al.* 2002a; Petit *et al.* 2002b). However, a few "outlier loci" show higher inter-specific differentiation and are potentially involved in the different local genetic adaptation that maintains the species genetically distinct (Finkeldey 2001; Gömöry *et al.* 2001; Scotti-Saintagne *et al.* 2004; Muir and Schlotterer 2005; Curtu *et al.* 2007a; Neophytou *et al.* 2010). Concerning the differentiation of the other closely related species, although they have been much less investigated, *Q. frainetto* seems to be genetically more similar to *Q. pubescens* than *Q. robur* or *Q. petraea* (Curtu *et al.* 2007a), whereas *Q. pubescens* exhibits low genetic differentiation with *Q. petraea* and high levels of genetic admixture with the latter species in mixed or pure stands (Curtu *et al.* 2009; Salvini *et al.* 2009). It has been suggested by many authors that diversifying selection has been the source of speciation among closely related *Quercus* species. Hence, the species differentiation, even in sympatry, can be seen as population differentiation under different environmental circumstances, and can be maintained by selection acting upon genes in different directions either directly or through hitchhiking effects (Le Corre and Kremer 2003; Petit *et al.* 2004; Scotti-Saintagne *et al.* 2004; Alberto *et al.* 2010)

Isocitrate dehydrogenases (IDH) have been reported in many different studies and organisms as potentially adaptive genes. $NADP^+$ IDH is a key enzyme of the citrate cycle which has been suggested to be involved in the essencial amino acid production (Palomo *et al.* 1998; Hodges *et al.* 2003). In terms of isozyme electrophoretic separation $NADP^+$ IDH has been reported as an outlier locus significantly differentiating among *Quercus* species (Finkeldey 2001; Gömöry *et al.* 2001; Scotti-Saintagne *et al.* 2004; Curtu *et al.* 2007a). Additionally, a correlation of specific alleles with the extend of beech scale insect (*Cryptococcus fagisuga*) infestation has been found for *Fagus sylvatica* (Ziehe 1996a; Ziehe 1996b). Other examples of the adaptive significance of the $NADP^+$ IDH genes in different species are the differences in the kinetic performance of the enzyme across thermal environments in cricket (*Allonemobius socius*) (Huestis *et al.* 2009), the up-regulation of $NADP^+$ isocitrate dehydrogenase in poplar (*Populus tremula* x *P. alba*) after treatment with PPT (phosphinothricin, a common herbicide used in agriculture) in a transgenic PPT-resistance background (Pascual *et al.* 2008), the 2-fold enhancement of the expression of the gene in ectomycorrhizal roots compared to non-mycorrhizal roots of *Eucalyptus globulus* (Boiffin *et al.*

1998), or in a recent study the association of a mutative $NADP^+$ IDH-1 gene with gliomas tumor in humans (Dang *et al.* 2009).

In the present study the $NADP^+$ IDH sequence of *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. frainetto* was identified. Potential SNP positions in coding and non-coding regions of the gene were detected and genotyped. Patterns of genetic variation and differentiation among the four species were determined for the non-synonymous, synonymous and non-coding SNPs. Linkage disequilibrium between the SNP loci was also estimated for each of the four *Quercus* species separately. Association analysis of all SNPs genotyped was done, aiming to identify possible co-variation between SNPs and morphological traits that differentiate the species, despite their sympatric co-existence in a four-species oak reserve in east-central Europe.

Materials and Methods

Plant material

253 oak individuals *Q. robur*, *Q. petraea*, *Q. pubescens*, *and Q. frainetto* were investigated in this study (65, 65, 73 and 50 individuals, respectively). The sampling was exhaustive and was conducted at the Bejan forest, a mixed oak reserve located in central-western Romania by Curtu (Curtu *et al.* 2007a). The four species co-exist in this location naturally. Extraction of genomic DNA was done and described by Curtu *et al.* (2007a).

Identification of the $NADP^+$ IDH sequences

To detect the informative SNPs for the analysis the sequence of a cytosolic $NADP^+$ dependent IDH gene should be first obtained. PCR reaction was performed as described in Manuscript I, using DNA template of a *Q. petraea* sample, with oligo-primers that were designed in three overlapping parts on the basis of the publicly available annotated sequences of *Populus trichocarpa* (jgi|Poptr1_1|770098|fgenesh4_pg.C_LG_X001588), (Tuskan *et al.* 2006) -given the lack of any public available annotated genome database of *Quercus* spp. On the basis of the *Q. petraea* sequence that was obtained, new primers were designed, for specificity to *Quercus* spp. For sequencing purposes the PCR primers were designed to amplify seven overlapping DNA fragments with the overlap varying between 60-300bp (see Manuscript I).

SNP identification and analysis

The sequences used in order to identify the possible SNP positions (obtained as described in Manuscript I) were manually verified and edited for the case of base-calling errors. Contigs were assembled using the computer software CodonCode Aligner (CodonCode Corp., Dedham, MA), or Sequencher v4.8 (Gene Codes, Ann Arbor. MI). Deriving from multiple alignments of the corresponding sequences using the Clustal W algorithm (Thompson et al., 1994) as applied by the BioEdit software (Hall, 1999), the potential SNP sites were identified and confirmed by SNP genotyping. The verified and clear for scoring non-synonymous SNPs were chosen for further genotyping for the total sample size (253 individuals). Similar number of synonymous SNPs and SNPs in the non-coding regions has been verified by genotyping and selected for genotyping on the total sample size.

The SNP data set was analyzed in terms of gene diversity (H_s) and inbreeding coefficient F_{is} within species with the software FSTAT v2.9.3.2. (Goudet 1995). The significance of the deviation of the observed F_{is} values from that expected in Hardy-Weinberg (HW) equilibrium was tested with 52000 randomizations. Species differentiation in terms of pairwise F_{st} between all pairs of species was analyzed with the software package Arlequin v3.5 (Excoffier and Lischer 2010). The significance of the F_{st} statistics was tested by 10000 permutations of the individuals over the populations (species in this study). An F_{st} outlier approach (FDIST approach, Beaumont and Nichols 1996) was also conducted. The null F_{st} distributions and the observed F_{st} values for each SNP as a function of their heterozygosity were plotted after performing 10000 coalescence simulations under a finite island model with 100 demes (Beaumont and Nichols 1996; Excoffier *et al.* 2009).

Linkage disequilibrium between all pairs of SNPs for each species separately was analyzed performing the likelihood-ratio test whose empirical distribution was obtained by 16000 permutations (Slatkin and Excoffier 1996), as implemented in the software package Arlequin v3.5 (Excoffier and Lischer 2010), using an Expectation-Maximization (EM) algorithm to estimate haplotype frequencies in the case of genotypic data with unknown gametic phase (Dempster *et al.* 1977; Excoffier and Slatkin 1998). For the total dataset, LD was analyzed as the squared correlation between all possible combinations of alleles (r^2) in the case of two alleles being present. In the case of multiple alleles a weighted average of r^2 was calculated according to the allele's frequency (Farnir *et al.* 2000). The corresponding *p*-values were determined by a two-sided Fisher's exact test. For the total sample set, the

analysis and the plotting were conducted with the software TASSEL v2.1 (Bradbury *et al.* 2007).

The possible association of each investigated SNP with morphological traits that best differentiated the species (Curtu *et al.* 2007b) was analyzed using F-Tests after 100000 permutations under a general linear model (GLM) taking into account the genetic structure of the samples, as the main factor of false positive associations (Marchini *et al.* 2004; Hirschhorn and Daly 2005). Data of genetic structure for the total sample set used for the analysis as covariates, were the structure data obtained by the use of six nuclear SSR markers published in a previous work (Curtu *et al.* 2007b) when analyzing the same individuals. The same authors, when analyzing the samples for their assignment to species based on morphological measurements found that the petiole ratio (PR) and the basal shape of the lamina (BS) where the quantitative characters that could best differentiate the species. PR is a transformed character derived from the measured characters: length of the lamina and the petiole length (LL and PL respectively) (Kremer *et al.* 2002). For the association analysis, we examined the possible associations between all the SNP markers with four morphological characteristics: LL, PL, SW (sinus width) and BS, as measured by Curtu *et al.* (2007b), using the statistical model: $y = marker + Q + e$, (where the *marker* component represents the SNP marker effect, Q represents the genetic structure effects, and e the phenotypic observations) as implemented in TASSEL v2.1 (Bradbury *et al.* 2007).

SNP genotyping

SNP genotyping was performed applying the "minisequencing" single nucleotide primer extension method (Pastinen *et al.* 1997) using an ABI Prism® SNaPshot™ Multiplex Kit Protocol (Applied Biosystems Foster City, CA) and following the manufacturer's protocol, with modifications. The PCR fragments of the $NADP^+$ dependent IDH gene that were obtained with the above mentioned procedure were pooled in order to be used as template for the primer extension reactions. Eventual primer residues were cleaved with 5U SAP (Shrimp Alkaline Phosphatase) enzyme and 1U Exonuclease I enzyme (USB, Europe GmbH, Staufen, Germany), by incubation at 37°C for 1 hour, followed by 15 minutes at 75°C for the deactivation of the enzymes. All SNP primers used were designed using the

web based software Primer3 (Rozen and Skaletsky 2000). The quality of the primers was determined with the software GeneRunner® (1994, Hastings Software, Inc.). The sequences of the SNP primers, their direction and the nucleotide substitution that they genotype, are provided in Table II-1.

Table II-1: Sequence, direction and substitution of SNP primers, >: forward, <: reverse. Non-synonymous: located in coding regions and cause an amino acid substitution, synonymous: located in coding regions and cause no amino acid substitution, non-coding: located in non-coding regions

	primer	direction	substitution	oligo sequence 5'-3'
non-synonymous	1-325	>	a/g	t(19)cag gag atg aaa tga ctc ga
	4-268b	>	a/g	t(26)gga ata tct gca gta ccc a
	4-490b	<	c/t	t(32)gaa gcc tca gca aaa gca
	7-325c	>	g/c	t(40)aat tgg aag cag cct gtg tt
synonymous	3-352s	>	a/g	t(18)ttc atc ccc cgt ctt gtc cc
	2-70	<	a/g	t(21)ctt gag cac ttt caa ttg taa c
	7-97s	>	t/a	t(35)aca aac agc ata gca tcc at
	7-330s	<	c/g	t(42)agt cat ctt tcc tga ttc cac
non-coding	1-51nc	>	a/g	t(16)gtt cat cat cta tta caa tca tgt ttt
	2-419nc	<	g/c	t(27)att gct aca ttg tac cta gaa agg
	3-696nc	>	a/g	t(36)tcc atc aat gcc ttc ata c
	6-171nc	<	g/t	t(39)aga gca agc acc ttt cca tt
	7-392nc	>	t/c	t(43)tca tgg gcc caa gta att tc

The SNP primers were pooled for multiplex primer-extension reactions ("minisequencing"), generating maximal seven SNPs per reaction. For the multiplexing, the SNP primers had four to eight bp difference in size, for separation of the different SNP genotypes, which was accomplished by an addition of a poly $(T)_n$ tail at their 5' end. The primer extension reaction was performed using the SNaPshot Multiplex Ready Reaction Mix chemistry ABI PRISM® SNaPshot™ Multiplex Kit Protocol (Applied Biosystems Foster City, CA), on a PTC-200 Peltier Thermal Cycler (MJResearch, Inc., Waltham, Massachusetts). The thermal cycling conditions were: 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 30seconds. The SNP reaction was post-extension treated, by adding 1U of SAP enzyme (USB, Europe GmbH, Staufen, Germany) and incubating for 1 hour at 37°C, and 15 minutes at 75°C to deactivate the enzyme.

For their electrophoretic separation, the purified primer extended fragments, were loaded in a HiDi™ Formamid (Applied Biosystems, Foster City, CA) elution, and run on a ABI Prism® 3100x*l* genetic analyzer (Applied Biosystems, Foster City, CA). The capillary electrophoresis was performed on a POP7® polymer and buffer with EDTA (Applied Biosystems, Foster City, CA) with 36cm long capillaries. The corresponding run module included oven temperature 60°C, run voltage 13.4kV, injection time 18 seconds and injection voltage 1.2kV. For fragment sizing the size standard GeneScan™ 120 LIZ™ (Applied Biosystems, Foster City, CA) was used. Data analysis was performed with GeneMapper® Software v4.0 (Applied Biosystems, Foster City, CA).

Results

In order to identify the SNPs of interest, the genomic nucleotide sequence coding for the cytosolic $NADP^+$ dependent IDH gene was obtained from five individuals per species based on their morphological classification by Curtu *et al.* (2007a). From a total of 20 sequences, 3481bp in size, 224 potential SNP positions were identified, excluding the simple sequence repeat motifs and the insertions or/and deletions. 38 of the potentially polymorphic sites were within the coding region of the gene. After testing one by one all the potential SNPs of the non-synonymous sites, four non-synonymous SNPs were finally chosen for further analyses, among others that either turned out to be false due to possible amplification errors, or led to spurious scoring due to possible double amplifications or primer mismatches. The same number of synonymous SNPs was chosen to be included in the analysis together with five additional SNPs from the non-coding regions; all SNPs were distributed throughout the total obtained sequence of the gene, covering almost the complete gene sequence (Figure II-1).

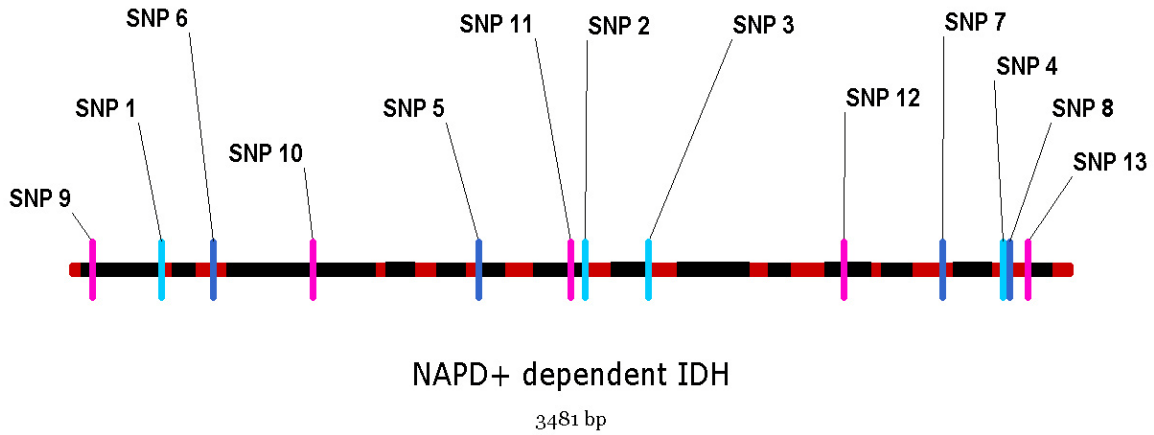

Figure II-1: Schematic representation of the $NADP^+$ IDH gene (red color: exons, black color: introns) and the position of the SNPs analyzed (non-synonymous SNPs, synonymous SNPs, non-coding SNPs)

The location of each SNP analyzed on the gene and the amino-acid replacement that is being caused is given in Table II-2. Accordingly, among the four non-synonymous SNPs that were analyzed two cause no charge change through their amino acid replacement (SNP 1 and SNP 2) whereas two other SNPs cause amino acid replacement with different charge (SNP 3 and SNP 4).

Table II-2: Position of the SNPs and amino acid - charge replacement caused by the non-synonymous SNPs

SNP		position on the gene	amino acid replacement	charge change
non-synonymous	1	306 - 2nd exon	isoleucine - valine	neutral-neutral
	2	1788 - 8th exon	asparagine - serine	neutral-neutral
	3	2009 - 9th exon	histidine - arginine	neutral-positive
	4	3267 - 14th exon	glycine - arginine	neutral-positive
synonymous	5	1417 - 6th exon	-	-
	6	489 - 3rd exon	-	-
	7	3039 - 13th exon	-	-
	8	3272 - 14th exon	-	-
non-coding	9	67 - 1st intron	-	-
	10	836 - 3rd intron	-	-
	11	1739 - 7th intron	-	-
	12	2693 - 11th intron	-	-
	13	3334 - 14th intron	-	-

Genetic variation within species

Overall, estimates of gene diversity at the polymorphic SNPs varied in our sample from 0.014 in *Q. pubescens* at SNP 2 to 0.507 in *Q. petraea* at SNP 12. SNP 2 was found monomorphic for *Q. robur* and *Q. frainetto* but also for the other species showed low diversity. The estimates for the non-coding SNPs were in general higher than those of the SNPs in coding regions. In particular, the non-coding SNP 11 and SNP 12 showed the highest levels of diversity over all species (Table II-3). On average, among similar values of gene diversity over species *Q. robur* exhibited the lowest whereas *Q. frainetto* displayed the highest mean values. SNP 7, SNP 12 and SNP 13 revealed a third allele present only once in each case (see detailed table of the genotypes in Appendix II-1).

Table II-3: Gene diversity H_s and inbreeding coefficient F_{is} per SNP and species

SNPs		*Q. robur*		*Q. petraea*		*Q. pubescens*		*Q. frainetto*	
		H_s	F_{is}	H_s	F_{is}	H_s	F_{is}	H_s	F_{is}
non-synonymous	1	0,089	-0.041	0,2	0.364	0,231	-0.024	0,205	0.086
	2	-	-	0,09	-0.041	0,014	0.000	-	-
	3	0,161	0.111	0,158	-0.086	0,28	0.007	0,237	0.209
	4	0,045	-0.016	0,206	-0.123	0,079	-0.036	0,114	0.300
Synonymous	5	0,089	-0.041	0,015	0.000	0,055	-0.022	0,116	-0.055
	6	0,317	0.063	0,498	-0.317	0,395	-0.064	0,501	-0.588
	7	0,089	-0.041	0,118	-0.059	0,178	0.064	0,303	-0.079
	8	0,015	0.000	0,076	-0.033	0,144	0.120	0,185	-0.103
non-coding	9	0,303	-0.014	0,134	-0.069	0,348	0.043	0,262	-0.013
	10	0,478	0.163	0,397	0.041	0,5	0.029	0,398	0.046
	11	0,485	0.387	0,494	0.067	0,503	0.033	0,489	0.165
	12	0,501	-0.197	0,507	-0.306	0,496	-0.464	0,48	-0.125
	13	0,131	-0.059	0,389	-0.108	0,343	-0.012	0,394	0.068
Mean		0.208	0.067	0.252	-0.090	0.274	-0.054	0.283	-0.049

The inbreeding coefficient in most of the cases was found negative, or close to zero. However, the non-coding SNPs for *Q. robur* and the non-synonymous SNPs for *Q. frainetto* exhibited the most elevated F_{is} values (0.387 at SNP 11 and 0.300 at SNP 4 respectively). Overall loci, a heterozygote deficit was displayed by a positive value in *Q. robur*, while in the other species F_{is} was negative, still in all cases not significantly different from the HW expectations.

Genetic differentiation among species

Pairwise F_{st} values over all SNPs were found in general low but significant for all pairs of species except that of *Q. petraea* and *Q. frainetto*. The analysis of the coding and non-coding SNPs separately revealed higher differentiation within the first group of markers among all pairs with the exception of *Q. pubescens* and *Q. frainetto*.

Table II-4: Pairwise F_{st} values (lower diagonal) and their significance (*: p<0.05, **: p<0.01, ***: p<0.001, n.s.: non significant - after 10000 permutations) for the groups of SNPs: coding, non-coding, synonymous, non-synonymous and overall loci

Coding SNPs

	Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Q. robur	-	***	**	***
Q. petraea	0.067	-	***	*
Q. pubescens	0.019	0.027	-	***
Q. frainetto	0.089	0.012	0.026	-

Non-coding SNPs

	Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Q. robur	-	**	*	**
Q. petraea	0.040	-	**	n.s.
Q. pubescens	0.015	0.025	-	**
Q. frainetto	0.047	-0.003	0.027	-

Synonymous SNPs

	Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Q. robur	-	***	**	***
Q. petraea	0.096	-	***	***
Q. pubescens	0.017	0.039	-	***
Q. frainetto	0.122	0.018	0.046	-

Non-synonymous SNPs

	Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Q. robur	-	*	*	n.s.
Q. petraea	0.026	-	*	n.s.
Q. pubescens	0.022	0.014	-	n.s.
Q. frainetto	0.018	0.003	-0.008	-

Overall				
	Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Q. robur	-	***	*	***
Q. petraea	0.050	-	***	n.s.
Q. pubescens	0.016	0.026	-	**
Q. frainetto	0.063	0.003	0.026	-

The pattern of almost all groups of markers (non-coding, coding and synonymous SNPs) suggested that among all species *Q. robur* and *Q. frainetto* were most highly differentiated (F_{st}=0.122 for synonymous SNPs). Only the non-synonymous SNPs resulted in higher F_{st} values between *Q. robur* against *Q. petraea* firstly and *Q. pubescens* secondly. Particularly, all pairs of species showed lower levels of differentiation with the non-synonymous SNPs alone, apart from the species pair *Q. robur* - *Q. pubescens* that was better differentiated with the non-synonymous SNPs. Additionally, *Q. petraea* and *Q. frainetto* were found the least differentiated species regardless which SNPs were analyzed. Regarding the non-synonymous SNPs alone, *Q. frainetto* is not significantly differentiated from any other species as opposed to the synonymous and the coding SNPs, as well as the overall SNP set.

The distribution of F_{st} values across all loci as a function of heterozygosity between populations by performing 10000 coalescence simulations under a finite island model with 100 demes (Beaumont and Nichols 1996; Excoffier *et al.* 2009) has suggested a significant departure from the neutral expectations at the 95% confidence level of the observed values of SNP 6, implying that it is potentially an outlier locus, candidate for being under selection (Figure II-2). Although the repetition of the simulations not always found the observed values of SNP 6 above the 95% confidence interval line, it was in all cases found above the 90% envelope of values corresponding to neutral expectations.

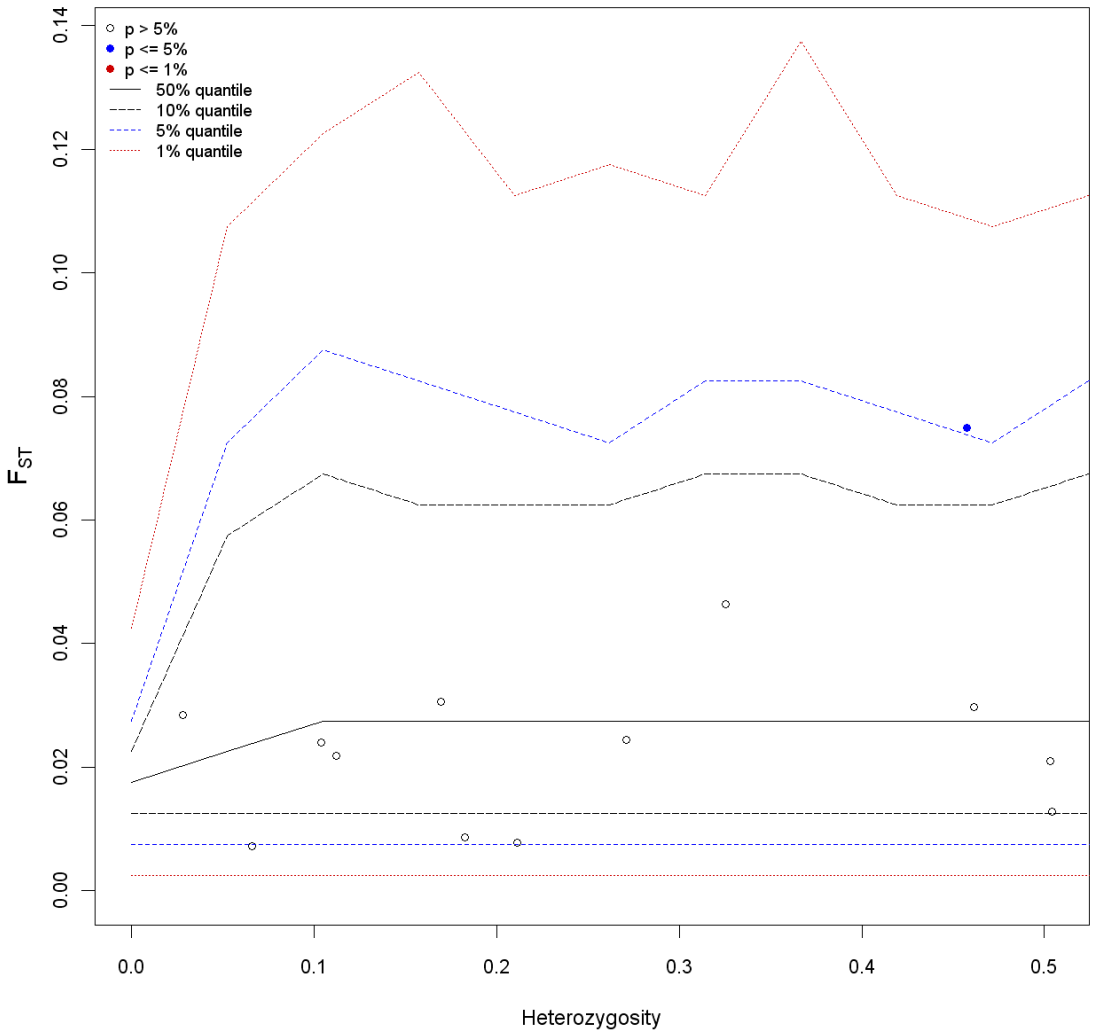

Figure II-2: Joint distribution of F_{st} values as a function of heterozygosity - The 90%, 95% and 99% envelopes of values corresponding to neutral expectations with the infinite allele model (Beaumont and Nichols 1996) are illustrated with black, blue and red dashed lines respectively. The blue dot corresponds to SNP 6.

Linkage disequilibrium and association analyses

Linkage disequilibrium (LD) analysis showed varying patterns of LD across the $NADP^+$ IDH gene over the different SNPs, but the overall effects of LD were relatively low. For the total sample set, among 78 pairs of comparisons 31 SNP pairs showed evidence for significant LD after 10000 permutations (39.7%). However, there was no clear physical clustering of sites in LD as shown in Figure II-3. In particular, several sites of close physical distance (a few nucleotides apart) showed levels of LD close to zero (e.g. SNP 4, SNP 8 and

SNP 13) whereas other sites with physical distance larger than thousand nucleotides were found to be significantly linked (e.g. SNP 1 and SNP 13).

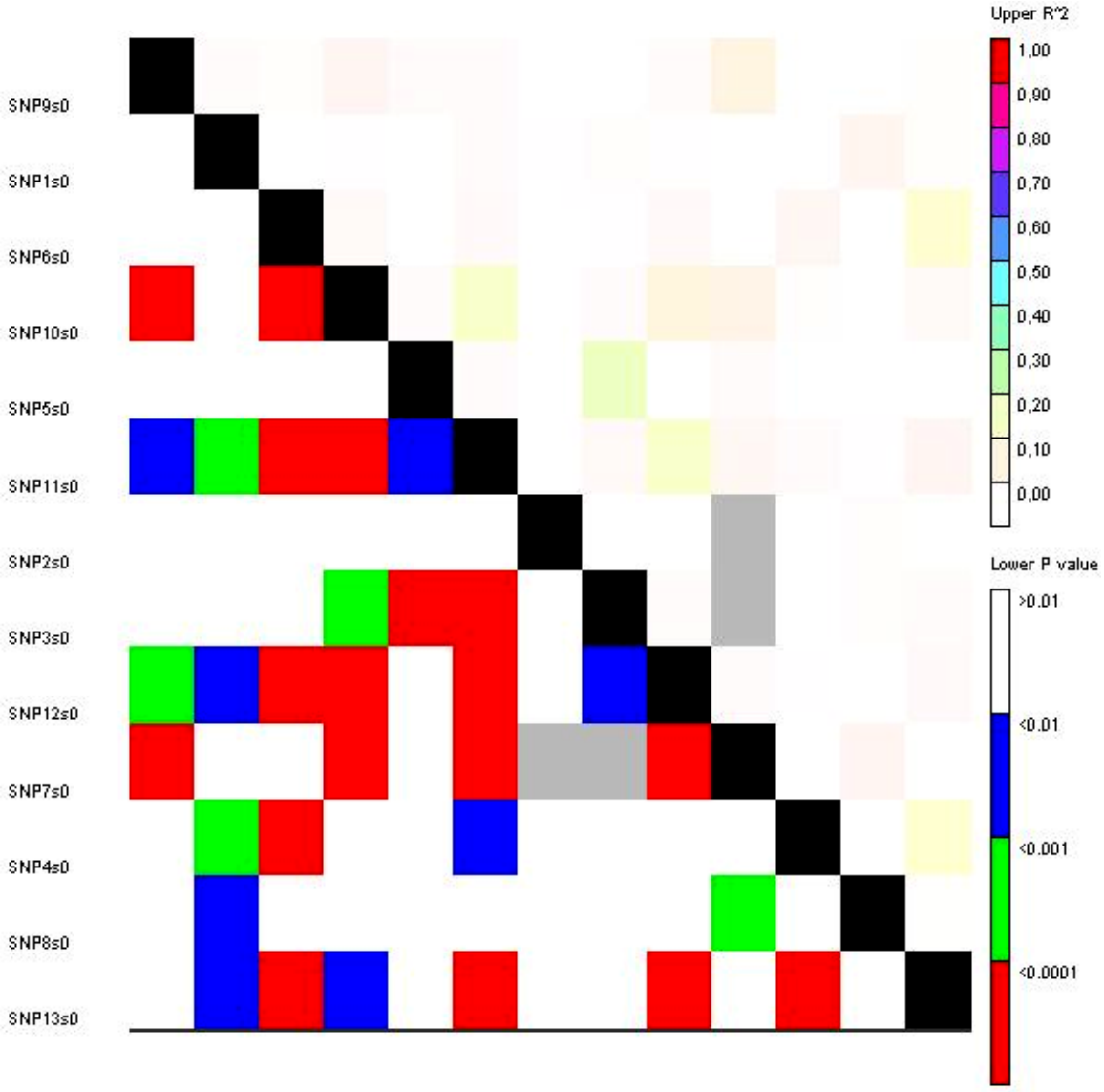

Figure II-3: LD plot of pairwise r^2 values between all pairs of SNPs (upper diagonal) and corresponding p values of the correlations after 10000 permutations (lower diagonal) according to the legend. The SNP markers are demonstrated with ascending physical distance downward (vertical axis) and from the left to the right (horizontal axis). – Grey color: not calculated.

The analysis of LD among the 13 different SNP sites for each species separately, revealed some differences among the species. In particular *Q. frainetto* had the highest numbers of significant ($p<0.05$) correlations (28 pairs), despite the fact that SNP 2 was

monomorphic for this population (Table II-5). *Q. robur* showed 15 pairs of significant squared allelic correlations (r^2). *Q. petraea* and *Q. pubescens* had 27 and 24 pairs of significant LD, respectively. In Table II-5, the summary of SNPs that are in significant LD ($p<0.05$) using the EM algorithm (Slatkin and Excoffier 1996) for estimating haplotype frequencies for unphased data is given. *Q. robur* exhibited the smallest number of SNPs in LD (14) with the other species having almost twice the number of SNPs linked significantly, beyond the neutral expectations and spanning larger physical distances.

Table II-5: Pairs of SNPs in linkage disequilibrium ($p<0.05$) (Numbers refer to the corresponding SNP)

SNP	*Q. robur*	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
1	-	9, 11, 12, 13	9	8, 9, 11
2	monomorphic	3, 6, 13	3, 5, 6	monomorphic
3	5, 11, 12	2, 6, 13	2, 5, 7, 10	5, 10, 11, 12
4	6, 13	6, 10, 11, 13	13	6, 13
5	3	-	2, 3, 6, 11	3, 10, 11
6	4, 10, 12, 13	2, 3, 4, 10, 11, 12, 13	2, 5, 10, 11, 13	4, 10, 11, 12
7	-	8, 9, 10, 11	3, 8, 9, 10, 11, 12, 13	8, 9, 10, 11, 12, 13
8	-	7	7, 13	1, 7, 13
9	10, 11, 12	1, 7	1, 7	1, 7, 10, 11
10	6, 9, 11, 12	4, 6, 7, 11, 12, 13	3, 6, 7, 11, 12	3, 5, 6, 7, 9, 11, 12
11	3, 9, 10, 12	1, 4, 6, 7, 10, 12, 13	5, 6, 7, 10, 12, 13	1, 3, 5, 6, 7, 9, 10, 12, 13
12	3, 6, 9, 10, 11	1, 6, 10, 11, 13	7, 10, 11	3, 6, 7, 10, 11, 13
13	4, 6	1, 2, 3, 4, 6, 10, 11, 12	4, 6, 7, 8, 11	4, 7, 8, 11, 12
Total pairs	14	27	24	27

The possible association of each investigated SNP with morphological traits was examined for the measured morphological traits that best differentiated the species in their original form (without any transformation) (Curtu *et al.* 2007b) after 100000 permutations under a general linear model (GLM) and taking into account the genetic structure of the samples, as being the most serious systematic bias creating false positive associations. The analysis showed for seven SNPs significant ($p<0.05$) associations with one of the four traits. However the R^2 percentage of the total variation explained by each mutation (SNP) was in general low (below 4%).

Table II-6: Significant ($p<0.05$) associations between SNPs and morphological traits (F) and the percentage of the phenotypic variation explained by the corresponding SNP (R^2)

SNP	Trait	Marker effect		
		F	p	R^2 (%)
SNP 1	sinus width	3.2605	0.04	2.1
SNP 2	sinus width	5.4461	0.0204	1.79
SNP 4	lamina length	4.4435	0.0127	3.45
SNP 5	lamina length	7.3993	0.007	2.91
SNP 7	lamina length	3.0501	0.0293	3.6
SNP 8	basal shape of lamina	3.8505	0.0226	3.05
SNP 12	petiole length	3.1907	0.0243	3.58

Discussion

In the present study a gene-associated approach for SNP analysis of a $NADP^+$ isocitrate dehydrogenase gene was followed. The total number of possible SNP positions - not taking into account the simple sequence repeat motifs and the insertions or the deletions- was much higher (224 sites in a total sequence of 3481 bp) than that reported previously for *Quercus* spp. (Vornam *et al.* 2007; Quang *et al.* 2008; Quang *et al.* 2009; Derory *et al.* 2010) but also for other plant genera (Ingvarsson 2005; Yamasaki *et al.* 2005; Zhu *et al.* 2007) but this might be due to possible PCR amplification or sequencing errors, or even due to cases of unspecific amplifications, as not all SNPs were resequenced. The SNPs that were finally analyzed and genotyped were tested for repetitive patterns of chromatographs and for clear-undoubtful scoring. Among the four non-synonymous SNPs that were analyzed only two revealed a change in the charge of the amino acid that they code for, leading potentially to structure differences and thus being potentially adaptive.

The levels of gene diversity that were measured for the total number of SNPs were in general similar to those revealed by isozyme investigations for the same sample set; SNP 2 exhibited low diversity in all species, but was specifically monomorphic for *Q. robur* and *Q. frainetto*, as it was also found at 6-Pgdh-B3 and IDH-B5 isozyme alleles for the two species respectively (Curtu *et al.* 2007a). The measurements of H_s for *Q. robur* in this study were found slightly lower than that for the other species, with the highest value found for *Q. frainetto* being in contrast with the isozyme investigation of Curtu *et al.* (2007a) that found in

Q. frainetto the lowest H_s value. This was attributed to the smaller sample size examined for this species. Indeed, bi-allelic SNP markers might be less influenced by the smaller sample size in terms of gene diversity than markers with larger number of alleles and higher mutation rate such as nuclear SSRs. However, the finding by Curtu *et al.* (2007a) of *Q. frainetto* as the less variable was only based on the mean gene diversity values, since single isozyme markers and/or nuclear SSR markers showed *Q. frainetto* equally or in cases even higher variable than the other species. In comparisons of the most widely distributed *Q. robur* and *Q. petraea* in Europe, similar to other studies of mixed or pure stands (Gömöry *et al.* 2001; Mariette *et al.* 2002; Finkeldey and Mátyás 2003) this study also showed higher variability for the latter species. In total, the levels of gene diversity were found higher at the non-coding SNPs than those found at the coding SNPs. Consistent with previous isozyme and nuclear SSR data (Gömöry *et al.* 2001; Finkeldey and Mátyás 2003; Curtu *et al.* 2007a; Neophytou *et al.* 2010), the estimation of the inbreeding coefficient F_{is} was found slightly negative or close to zero. An exception was the case of *Q. robur*, where F_{is} was found slightly above zero. The same pattern of F_{is} was also found for the same sample set after IDH-B isozyme analysis. In general, the slightly higher average F_{is} of *Q. robur* as compared to those observed in *Q. petraea* is in accordance with previous studies, yet in this study F_{is} for *Q. robur* did not significantly deviate from HW expectations.

The genetic differentiation among the four species was overall low but significant, in almost all pairs of species, with the exception of Q. *frainetto* and *Q. petraea*. The significant differentiation of the species was in consistence with the findings of a previous study based on isozymes and nuclear SSR markers, applied on the same individuals as in the present work (Curtu *et al.* 2007a). The non-synonymous SNPs alone failed to distinguish significantly *Q. frainetto* and all other species. On the other hand, the synonymous SNPs revealed the largest significance not only in differentiating between *Q. frainetto* from the other species, but also between the other species pairs. This is an effect of the contribution of SNP 6 which was found to have outlier behavior in terms of its F_{st} distributions (Figure II-2) to the pairwise F_{st} of the group of coding SNPs. Pairwise F_{st} analysis of the coding markers not including SNP 6 revealed lower differentiation measurements than the non-coding group (data not shown). In almost all groups of SNP markers the present results confirm previous findings in that among the four species, *Q. robur* and *Q. frainetto* are the most differentiated (Schwarz 1993; Curtu *et al.* 2007a). There was a surprising exception of this rule within the

group of non-synonymous SNPs (SNP 1, SNP 2, SNP 3 and SNP 4), at which the highest differentiation was found between the pairs *Q. robur* - *Q. petraea* followed by *Q. robur* - *Q. pubescens*; the F_{st} values being low but still significant at the 95% level. With the same group of markers *Q. frainetto* was found to be least differentiated from *Q. robur*, in accordance with a phylogenetic study of Italian oaks based on sequence comparison of the ITS1 and ITS2 regions of the 5.8S RNA encoding ribosomal DNA (Bellarosa *et al.* 2005) where a closer relationship between *Q. robur* and *Q. frainetto* was found as compared to *Q. pubescens* and *Q. petraea.*

The plotting of the distributions of F_{st} as a function of heterozygosity resulted in SNP 6 exceeding the simulated expectations under neutrality and thus being potentially under selection. The method of the detection of F_{st} outlier loci is yet designed for genome scan data (Beaumont and Nichols 1996; Excoffier *et al.* 2009). Hence it is quite sensitive to the number of loci that are included in the analysis and to low sample sizes (Eveno *et al.* 2008). In cases of low sample sizes or loci number and of populations that are significantly differentiated also with neutral markers as in the present study the results of this analysis should be interpreted with caution. Additionally, the 13 SNPs analyzed represent SNPs of three different categories within one gene, but in any case do not represent the whole genome. Moreover, the specific SNP marker that was found above the neutral threshold was a SNP of the coding region of the gene, but silent. This rejects the hypothesis of its direct adaptive significance and the action of selection upon it. But the hypothesis that this SNP might be linked to an adaptive SNP and covariate with it through a selective sweep or/and high levels of linkage disequilibrium cannot be rejected (Palme *et al.* 2008; Eckert *et al.* 2009; Alberto *et al.* 2010), if the high F_{st} levels do not just reflect the high differentiation among species of this stand observed also by neutral markers (Curtu *et al.* 2007a). Another explanation of course might be the incorporation of any systematic errors due to possible unspecific or double amplifications or even cases of null alleles in the SNP genotyping, but the likelihood of this kind of errors has been reduced during the design and testing of the SNP primers before multiplexing them. Yet, although the SNP extension primers were designed after sequencing and genotyped after repeated testing, errors might still have remained undetectable. In general, null alleles in SNP genotyping are frequent but hardly detectable. They can be caused by deletions spanning a polymorphic site (McCarroll *et al.* 2006), secondary polymorphisms interfering with genotyping at the primary polymorphic

target and even unexpected alleles at the primary polymorphic sites (such as triallelic sites) (Carlson *et al.* 2004). All of these are important potential sources of reproducible, but inaccurate, genotypes. Similar challenges are posed by polymorphisms within segmental duplications, where the number of copies of the surrounding sequence itself can be variable (Sharp *et al.* 2005). This type of error meets the case of SNP genotyping belonging to gene families like isocitrate dehydrogenases or duplicated genes. Although null alleles could be a potentially important sources of genotyping errors, these alleles are difficult to detect because null allele heterozygotes are indistinguishable from the expected homozygotes or/and paralogous homozygotes are difficult to tell from orthologous heterozygotes on most genotyping platforms (Carlson *et al.* 2006).

The estimates for nonrandom linkage of different alleles among pairs of SNPs in the total sample set, revealed evidence for significant LD in 39.7% of all possible comparisons. This percentage is in general high. For instance, 12.8% of all possible comparisons showed significant LD in an analysis within and surrounding the *phyB2* gene in *Populus tremula* (Ingvarsson *et al.* 2008). The high percentage observed can be explained by the location of all SNPs within the same gene, despite the relatively large size of the gene. However, the significant associations were not necessarily found between the SNPs of closer proximity. Still, as illustrated in Figure II-2, the estimates of r^2 are low with very few comparisons exceeding the levels of 0.2-0.3, as it would be expected for a predominantly outcrossing species (Neale and Savolainen 2004; Morrell *et al.* 2005; Zhang and Zhang 2005; Ingvarsson *et al.* 2008) In all such cases, the physical distance of the SNPs was larger than 0.5kb (s. Figure II-3 and Table II-2). In the LD analysis of each species separately, the low number of linked loci in the case of *Q. robur* might reflect the highest effectiveness of recombination, due to a larger effective population's size. *Q. frainetto* is possible to reflect a founder effect on the higher number of significant associations spanning larger physical distances, since the stand of this study is at the edge of its distribution (Bartha 2006). The same could be the case for *Q. pubescens* since Romania is considered close to the north-eastern part of the species' distribution (Bussotti 2006), but the genetic admixture with the other sympatric oak species could explain better the high number of linked polymorphic sites in larger physical distances. Indeed, higher hybridization and admixture between *Q. petraea* and *Q. pubescens* was reported by conducting paternity analyses by Curtu *et al.* (2009) in the same stand.

The association study of all SNPs with the leaf morphological traits that more successfully differentiated the species (except *Q. petraea* and *Q. pubescens* that needed pubescence information to be distinguished, for details refer to Curtu *et al.* 2007b) revealed seven significant associations. In some cases multiple associations were found for one trait. In particular, SNP 1 and SNP 2 were both significantly associated with the sinus width in the sample. Both SNPs seem statistically to be independently associated with the trait, since they are not linked with each other according to the LD analysis. The same holds true for the tree SNP markers that showed significant association with lamina length but were not found in significant LD (SNP 4, SNP 5 and SNP 7). Among the seven statistically associated loci, SNP 1, SNP 2, and SNP4 are interestingly non-synonymous mutations. However, all associations exhibited a low R^2 percentage, indicating low robustness of the statistical associations according to several authors (Ingvarsson *et al.* 2008; Simko *et al.* 2009). On the contrary, in an association study of SNPs derived from candidate genes for wood property traits in *Pinus taeda*, similar R^2 percentages were proven to be still significant after false discovery rate (FDR) corrections (González-Martínez *et al.* 2007).

In conclusion, the present study has shown that the designed and analyzed SNP markers of the $NADP^+$ IDH gene can differentiate the closely related oak species significantly. SNP 6 was shown to behave as an outlier candidate for being under selection. Among the SNPs there was evidence of several being in significant LD. The analysis of linkage disequilibrium of the SNPs in each population separately suggested a possible founder effect for the population of *Q. frainetto* and high levels of admixture among the sympatric oak species. The obtained significant associations of seven SNPs with four different traits of leaf morphology can lead to the proposal of their further examination by QTL mapping applications and by *Quercus* pedigrees that segregate for leaf morphological traits.

References

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver (2002). "Interrogating a high-density SNP map for signatures of natural selection." Genome Research **12**(12): 1805-1814.

Alberto, F., J. Niort, J. Derory, O. Lepais, R. Vitalis, D. Galop and A. Kremer (2010). "Population differentiation of sessile oak at the altitudinal front of migration in the French Pyrenees." Molecular Ecology **19**(13): 2626-2639.

Bartha, D. (2006). "*Quercus frainetto* Ten." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-8. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm, Ecomed,, Landsber am Lech.

Beaumont, M. A. and R. A. Nichols (1996). "Evaluating Loci for Use in the Genetic Analysis of Population Structure." Proceedings of the Royal Society of London. Series B: Biological Sciences **263**(1377): 1619-1626.

Bellarosa, R., M. C. Simeone, A. Papini and B. Schirone (2005). "Utility of ITS sequence data for phylogenetic reconstruction of Italian *Quercus* spp." Molecular Phylogenetics and Evolution **34**(2): 355-370.

Boiffin, V., M. Hodges, S. Galvez, R. Balestrini, P. Bonfante, P. Gadal and F. Martin (1998). "Eucalypt NADP-dependent isocitrate dehydrogenase - cDNA cloning and expression in ectomycorrhizae." Plant Physiology **117**(3): 939-948.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss and E. S. Buckler (2007). "TASSEL: software for association mapping of complex traits in diverse samples." Bioinformatics **23**(19): 2633-2635.

Bussotti, F. (2006). "*Quercus pubescens* Willd." Enzyklopädie der Holzsgewächse: Handbuch und Atlas der Dendrologie**:** 1-10. Schütt, P., H. Weisgerber, U. Lang, A. Roloff, B. Stimm., Ecomed,, Landsber am Lech.

Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak and D. A. Nickerson (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." American Journal of Human Genetics **74**(1): 106-120.

Carlson, C. S., J. D. Smith, I. B. Stanaway, M. J. Rieder and D. A. Nickerson (2006). "Direct detection of null alleles in SNP genotyping data." Hum. Mol. Genet. **15**(12): 1931-1937.

Coart, E., V. Lamote, M. De Loose, E. Van Bockstaele, P. Lootens and I. Roldan-Ruiz (2002). "AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl] in Flemish populations." Theoretical and Applied Genetics **105**(2-3): 431-439.

Curtu, A. L., O. Gailing and R. Finkeldey (2007b). "Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community." BMC Evolutionary Biology **7**.

Curtu, A. L., O. Gailing and R. Finkeldey (2009). "Patterns of contemporary hybridization inferred from paternity analysis in a four-oak-species forest." BMC Evolutionary Biology **9**.

Curtu, A. L., O. Gailing, L. Leinemann and R. Finkeldey (2007a). "Genetic variation and differentiation within a natural community of five oak species (*Quercus* spp.)." Plant Biology **9**(1): 116-126.

Dang, L., D. W. White, S. Gross, B. D. Bennett, M. A. Bittinger, E. M. Driggers, V. R. Fantin, H. G. Jang, S. Jin, M. C. Keenan, K. M. Marks, R. M. Prins, P. S. Ward, K. E. Yen, L. M. Liau, J. D. Rabinowitz, L. C. Cantley, C. B. Thompson, M. G. Vander Heiden and S. M. Su (2009). "Cancer-associated IDH1 mutations produce 2-hydroxyglutarate." Nature **462**(7274): 739-744.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data Via EM Algorithm." Journal of the Royal Statistical Society Series B-Methodological **39**(1): 1-38.

Derory, J., C. Scotti-Saintagne, E. Bertocchi, L. Le Dantec, N. Graignic, A. Jauffres, M. Casasoli, E. Chancerel, C. Bodenes, F. Alberto and A. Kremer (2010). "Contrasting

relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks." Heredity **104**(5): 438-448.

Eckert, A. J., J. L. Wegrzyn, B. Pande, K. D. Jermstad, J. M. Lee, J. D. Liechty, B. R. Tearse, K. V. Krutovsky and D. B. Neale (2009). "Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in Coastal Douglas Fir (*Pseudotsuga menziesii* var. *menziesii*)." Genetics **183**(1): 289-298.

Eveno, E., C. Collada, M. A. Guevara, V. Léger, A. Soto, L. Díaz, P. Léger, S. C. González-Martínez, M. T. Cervera, C. Plomion and P. H. Garnier-Géré (2008). "Contrasting Patterns of Selection at Pinus pinaster Ait. Drought Stress Candidate Genes as Revealed by Genetic Differentiation Analyses." Mol Biol Evol **25**(2): 417-437.

Excoffier, L., T. Hofer and M. Foll (2009). "Detecting loci under selection in a hierarchically structured population." Heredity **103**(4): 285-298.

Excoffier, L. and H. E. L. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." Molecular Ecology Resources **10**(3): 564-567.

Excoffier, L. and M. Slatkin (1998). "Incorporating genotypes of relatives into a test of linkage disequilibrium." American Journal of Human Genetics **62**(1): 171-180.

Farnir, F., W. Coppieters, J. J. Arranz, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, M. Mni, C. Nezer, P. Simon, P. Vanmanshoven, D. Wagenaar and M. Georges (2000). "Extensive genome-wide linkage disequilibrium in cattle." Genome Research **10**(2): 220-227.

Finkeldey, R. (2001). Genetic variation of oaks (*Quercus* spp.) in Switzerland - 2. Genetic structures in "pure" and "mixed" forests of pedunculate oak (*Q. robur* L.) and sessile oak (*Q. petraea* (Matt.) Liebl.). Silvae genetica. **50:** 22-30.

Finkeldey, R. and G. Mátyás (2003). "Genetic variation of oaks (*Quercus* spp.) in Switzerland. 3. Lack of impact of postglacial recolonization history on nuclear gene loci." Theoretical and Applied Genetics **106**(2): 346-352.

Gailing, O., B. Vornam, L. Leinemann and R. Finkeldey (2009). "Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model." Physiologia Plantarum **137**(4): 509-519.

Gömöry, D., I. Yakovlev, P. Zhelev, J. Jedinakova and L. Paule (2001). "Genetic differentiation of oak populations within the *Quercus robur*/*Quercus petraea* complex in Central and Eastern Europe." Heredity **86**(5): 557-563.

González-Martínez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale (2006). "DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L." Genetics **172**(3): 1915-1926.

González-Martínez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson and D. B. Neale (2007). "Association genetics in *Pinus taeda* L. I. Wood property traits." Genetics **175**(1): 399-409.

Goudet, J. (1995). "FSTAT (Version 1.2): A computer program to calculate F-statistics." Journal of Heredity **86**(6): 485-486.

Hall, T. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series **41**: 95-98.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nature Reviews Genetics **6**(2): 95-108.

Hodges, M., V. Flesch, S. Gálvez and E. Bismuth (2003). "Higher plant $NADP^+$-dependent isocitrate dehydrogenases, ammonium assimilation and NADPH production." Plant Physiology and Biochemistry **41**(6-7): 577-585.

Huestis, D. L., B. Oppert and J. L. Marshall (2009). "Geographic distributions of Idh-1 alleles in a cricket are linked to differential enzyme kinetic performance across thermal environments." BMC Evolutionary Biology **9**.

Ingvarsson, P. K. (2005). "Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (*Populus tremula* L., *Salicaceae*)." Genetics **169**(2): 945-953.

Ingvarsson, P. K., M. V. Garcia, V. Luquez, D. Hall and S. Jansson (2008). "Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, *Salicaceae*)." Genetics **178**(4): 2217-2226.

Kremer, A., J. L. Dupouey, J. D. Deans, J. Cottrell, U. Csaikl, R. Finkeldey, S. Espinel, J. Jensen, J. Kleinschmit, B. Van Dam, A. Ducousso, I. Forrest, U. L. de Heredia, A. J. Lowe, M. Tutkova, R. C. Munro, S. Steinhoff and V. Badeau (2002). "Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands." Annals of Forest Science **59**(7): 777-787.

Krutovsky, K. V. (2006). "From population genetics to population genomics of forest trees: Integrated population genomics approach." Russian Journal of Genetics **42**(10): 1088-1100.

Le Corre, V. and A. Kremer (2003). "Genetic Variability at Neutral Markers, Quantitative Trait Loci and Trait in a Subdivided Population Under Selection." Genetics **164**(3): 1205-1219.

Marchini, J., L. R. Cardon, M. S. Phillips and P. Donnelly (2004). "The effects of human population structure on large genetic association studies." Nature Genetics **36**(5): 512-517.

Mariette, S., J. Cottrell, U. M. Csaikl, P. Goikoechea, A. Konig, A. J. Lowe, B. C. Van Dam, T. Barreneche, C. Bodenes, R. Streiff, K. Burg, K. Groppe, R. C. Munro, H. Tabbener and A. Kremer (2002). "Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Q. petraea* (MATT.) LIEBL. and *Q. robur* L. stands." Silvae Genetica **51**(2-3): 72-79.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly and D. M. Altshuler (2006). "Common deletion polymorphisms in the human genome." Nature Genetics **38**(1): 86-92.

Montesinos, A., S. J. Tonsor, C. Alonso-Blanco and F. X. Pico (2009). "Demographic and Genetic Patterns of Variation among Populations of *Arabidopsis thaliana* from Contrasting Native Environments." Plos One **4**(9).

Morin, P. A., G. Luikart, R. K. Wayne and S. W. Grp (2004). "SNPs in ecology, evolution and conservation." Trends in Ecology & Evolution **19**(4): 208-216.

Morrell, P. L., D. M. Toleno, K. E. Lundy and M. T. Clegg (2005). "Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2442-2447.

Muir, G. and C. Schlotterer (2005). "Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.)." Molecular Ecology **14**(2): 549-561.

Neale, D. B. and O. Savolainen (2004). "Association genetics of complex traits in conifers." Trends in Plant Science **9**(7): 325-330.

Neophytou, C., F. A. Aravanopoulos, S. Fink and A. Dounavi (2010). "Detecting interspecific and geographic differentiation patterns in two interfertile oak species (*Quercus petraea* (Matt.) Liebl. and *Q. robur* L.) using small sets of microsatellite markers." Forest Ecology and Management **259**(10): 2026-2035.

Palme, A. E., M. Wright and O. Savolainen (2008). "Patterns of Divergence among Conifer ESTs and Polymorphism in *Pinus sylvestris* Identify Putative Selective Sweeps." Molecular Biology and Evolution **25**(12): 2567-2577.

Palomo, J., F. Gallardo, M. F. Suarez and F. M. Canovas (1998). "Purification and characterization of NADP(+)-linked isocitrate dehydrogenase from Scots pine - Evidence for different physiological roles of the enzyme in primary development." Plant Physiology **118**(2): 617-626.

Pascual, M. B., Z. P. Jing, E. G. Kirby, F. M. Canovas and F. Gallardo (2008). "Response of transgenic poplar overexpressing cytosolic glutamine synthetase to phosphinothricin." Phytochemistry **69**(2): 382-389.

Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen and A. C. Syvänen (1997). "Minisequencing: A Specific Tool for DNA Analysis and Diagnostics on Oligonucleotide Arrays." Genome Research **7**(6): 606-614.

Petit, R. J., C. Bodenes, A. Ducousso, G. Roussel and A. Kremer (2004). "Hybridization as a mechanism of invasion in oaks." New Phytologist **161**(1): 151-164.

Petit, R. J., S. Brewer, S. Bordacs, K. Burg, R. Cheddadi, E. Coart, J. Cottrell, U. M. Csaikl, B. van Dam, J. D. Deans, S. Espinel, S. Fineschi, R. Finkeldey, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. Konig, A. J. Lowe, S. F. Madsen, G. Matyas, R. C. Munro, F. Popescu, D. Slade, H. Tabbener, S. G. M. de Vries, B. Ziegenhagen, J. L. de Beaulieu and A. Kremer (2002b). "Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence." Forest Ecology and Management **156**(1-3): 49-74.

Petit, R. J., U. M. Csaikl, S. Bordacs, K. Burg, E. Coart, J. Cottrell, B. van Dam, J. D. Deans, S. Dumolin-Lapegue, S. Fineschi, R. Finkeldey, A. Gillies, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. Konig, A. J. Lowe, S. F. Madsen, G. Matyas, R. C. Munro, M. Olalde, M. H. Pemonge, F. Popescu, D. Slade, H. Tabbener, D. Taurchini, S. G. M. de Vries, B. Ziegenhagen and A. Kremer (2002a). "Chloroplast DNA variation in European white oaks - Phylogeography and patterns of diversity based on data from over 2600 populations." Forest Ecology and Management **156**(1-3): 5-26.

Porth, I., C. Scotti-Saintagne, T. Barreneche, A. Kremer and K. Burg (2005). "Linkage mapping of osmotic stress induced genes of oak." Tree Genetics & Genomes **1**(1): 31-40.

Quang, N. D., S. Ikeda and K. Harada (2008). "Nucleotide variation in *Quercus crispula* Blume." Heredity **101**(2): 166-174.

Quang, N. D., S. Ikeda and K. Harada (2009). "Patterns of Nucleotide Diversity at the Methionine Synthase Locus in Fragmented and Continuous Populations of a Wind-Pollinated Tree, *Quercus mongolica* var. *crispula*." J Hered **100**(6): 762-770.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." **132**: 365-386.

Salvini, D., P. Bruschi, S. Fineschi, P. Grossoni, E. D. Kjaer and G. G. Vendramin (2009). "Natural hybridisation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens*

Willd. within an Italian stand as revealed by microsatellite fingerprinting." Plant Biology **11**(5): 758-765.

Schwarz, O. (1993). "*Quercus*". Flora Europaea **1** Tutin, T. G., Burges, N. A., Chater, A. O., Edmondson, J. R., Heywood, V. H., Moore, D. M., Valentine, D. H., Walters, S. M., Webb, D. A., Cambridge University Press, Cambridge, UK.

Scotti-Saintagne, C., S. Mariette, I. Porth, P. G. Goicoechea, T. Barreneche, K. Bodenes, K. Burg and A. Kremer (2004). "Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]." Genetics **168**(3): 1615-1626.

Sharp, A. J., D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel and E. E. Eichler (2005). "Segmental duplications and copy-number variation in the human genome." American Journal of Human Genetics **77**(1): 78-88.

Simko, I., D. A. Pechenick, L. K. McHale, M. J. Truco, O. E. Ochoa, R. W. Michelmore and B. E. Scheffler (2009). "Association mapping and marker-assisted selection of the lettuce dieback resistance gene Tvr1." BMC Plant Biology **9**

Slate, J., J. Gratten, D. Beraldi, J. Stapley, M. Hale and J. M. Pemberton (2009). "Gene mapping in the wild with SNPs: guidelines and future directions." Genetica **136**(1): 97-107.

Slatkin, M. and L. Excoffier (1996). "Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm." Heredity **76**: 377-383.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, et al. (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)." Science **313**(5793): 1596-1604.

Vornam, B., O. Gailing, R. Finkeldey, C. Collada, M. A. Guevara, A. Soto, d. M. N., S. C. González-Martínez, D. L., A. R., I. Aranda, J. Climent, M. T. Cervera, P. G. Goicoechea, L. V., E. Eveno, J. Derory, P. Garnier-Géré, A. Kremer and C. Plomion (2007). "Naturally occuring nucleotide diversity in candidate genes for forest tree adaptation: magnitude, distribution and association with quantitative trait variation." The German Plant Genome Research Program Progress Report 2004-2007**:** 116-120. GABI, Potsdam-Golm, Germany.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley, B. S. Gaut and M. D. McMullen (2005). "A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement." Plant Cell **17**(11): 2859-2872.

Zhang, D.-q. and Z.-y. Zhang (2005). "Single nucleotide polymorphisms (SNPs) discovery and linkage disequilibrium (LD) in forest trees." Forestry Studies in China **7**(3): 1-14.

Zhu, Q. H., X. M. Zheng, J. C. Luo, B. S. Gaut and S. Ge (2007). "Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of rice." Molecular Biology and Evolution **24**(3): 875-888.

Ziehe, M. (1996a)."Anpassungsprozesse auf Populationsebene" Beese, F. Berichte des Forschungzentrums Waldökosysteme der Universität Göttingen, Reihe B. **51** 126-136 Göttingen.

Ziehe, M. (1996b)."Untersuchungen von Buchenbeständen zur Beurteilung von genetischen Anpassungsprozessen und Angepasstheit anhand von Markergenloci" Beese, F. Berichte des Forschungzentrums Waldökosysteme der Universität Göttingen, Reihe B. **52** 370-378 Göttingen.

7. Appendix

Appendix 1a

BLASTx hits with query the reference NADP$^+$ IDH (partial) transcript sequence and ClustalW alignment of the corresponding proteins

```
>protein NADP-IDH from reference sequence
PIVEMDGDEMTRIFWKSIKDKLIFPFVELDIKYFDLGLTNREATNDNVTIESAQATLRYNVAIKCATITPD
EARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKNIPRLVPGWTKPICIGRHAFGDQYRATDIVIQESG
KLKLVFVPNGHNEKKEFEVFNFTGAGGVALSMYNTDESIHAFAEASMNTAYQKKWPLYLSTKNTILKKYDG
RFKDIFQEVYETQWKSKFEAAGIWYEHRLIDDMVAYAMKSEGGYVWACKNYDGDVQSDFLAQGFGSLGMMT
SVLVCPDGKTIEAEAAHGAVTRHYRVHQKGGETSTNSIASIFAWTRGLAHRAKLDSNARLLDFTEKLEAAC
VGTVESGKMTRDLALLIHGPKVTRSRYLNTEEFIDAVAEELKG

>gi|224079726|ref|XP_002305928.1|     predicted     protein     [Populus
trichocarpa]*
MAYEKIKVANPIVEMDGDEMTRIFWQSIKDKLIFPFVELDIKYFDLGLPHRDATDDKVTVESAEAALKYN
VAIKCATITPDEARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKNIPRLVPGWTKPICIGRHAFGDQ
YRATDAVIKGAGKLKLVFVPEGQDEKTELEVYNFTGAGGVALAMYNTDESIRAFAEASMNTAYQKKWPLY
LSTKNTILKKYDGRFKDIFQEVYEANWKSKYEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQS
DFLAQGFGSLGLMTSVLVCPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWSRGLAHRAKLDD
NARLLDFTEKLEAACIGAVESGKMTKDLALLIHGSKVSRDHYLNTEEFIDAVAEELKARLSIKA

>gi|255575226|ref|XP_002528517.1|        NADP-specific        isocitrate
dehydrogenase, putative [Ricinus communis]*
MAFEKIKVANPIVEMDGDEMTRIFWKSIKDKLIFPFLELDIKYFDLGLPHRDATDDKVTIESAEATLKYN
VAIKCATITPDEARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKNVPRLVPGWTKPICIGRHAFGDQ
YRATDAVIKGAGKLKLVFVPEGQDEKTELEVYNFTGAGGVALSMYNTDESIRAFADASMNTAYQKKWPLY
LSTKNTILKKYDGRFKDIFQEVYEASWKSKFEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQS
DFLAQGFGSLGLMTSVLVCPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWSRGLAHRAKLDD
NARLLDFTEKLEAACIGVVESGKMTKDLALLIHGSKVTRDQYLNTEEFIDAVAADLAERLSKA

>gi|213493066|dbj|BAG84436.1|  NADP-dependent  isocitrate  dehydrogenase
[Passiflora edulis]*
MAFEKIKVANPIVEMDGDEMTRVFWKSIKDKLIFPFVELDIKYFDLGLPHRDATDDKVTIESAEATLKYN
VAIKCATITPDEARVEEFSLKQMWRSPNGTIRNILNGTVFREPILCKNIPRLVPGWTKPICIGRHAFGDQ
YRATDTVIKGAGKLKLVFVPEGQDEKTELEVFNFTGEGGVALSMYNTDESIRAFADASMNTAYQKKWPLY
LSTKNTILKKYDGRFKDIFQEVYEASWKSKFEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQS
DFLAQGFGSLGLMTSVLVCPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWTRGLAHRAKLDD
NARLLDFTEKLEAACVGAVESGKMTKDLALIIHGSKLSRDKYLNTEEFIDAVADELKARLSIKA
```

*: accession numbers and annotations from the NCBI databank, online publicly available: http://www.ncbi.nlm.nih.gov/

Alignment of NADP$^+$ dependent IDH

```
Populus trichocarpa      MAYEKIKVANPIVEMDGDEMTRIFWQSIKDK--LIFPFVELDIKYFDLGLPHRDATDDKV 58
Ricinus communis         MAFEKIKVANPIVEMDGDEMTRIFWKSIKDK--LIFPFLELDIKYFDLGLPHRDATDDKV 58
Passiflora edulis        MAFEKIKVANPIVEMDGDEMTRVFWKSIKDK--LIFPFVELDIKYFDLGLPHRDATDDKV 58
Quercus robur            ----------PIVEMDGDEMTRIFWKSIKDKVLLIFPFVELDIKYFDLGLTNREATNDNV 50
                                    ************:**:*****  *****:***********.:*:**:*:*
```

```
Populus trichocarpa      TVESAEAALKYNVAIKCATITPDEARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKN 118
Ricinus communis         TIESAEATLKYNVAIKCATITPDEARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKN 118
Passiflora edulis        TIESAEATLKYNVAIKCATITPDEARVEEFSLKQMWRSPNGTIRNILNGTVFREPILCKN 118
Quercus robur            TIESAQATLRYNVAIKCATITPDEARVKEFNLKQMWKSPNGTIRNILNGTVFREPIICKN 110
                         *:***:*:*:*****************:**.*****:******************:***

Populus trichocarpa      IPRLVPGWTKPICIGRHAFGDQYRATDAVIKGAGKLKLVFVPEGQDEKTELEVYNFTGAG 178
Ricinus communis         VPRLVPGWTKPICIGRHAFGDQYRATDAVIKGAGKLKLVFVPEGQDEKTELEVYNFTGAG 178
Passiflora edulis        IPRLVPGWTKPICIGRHAFGDQYRATDTVIKGAGKLKLVFVPEGQDEKTELEVFNFTGEG 178
Quercus robur            IPRLVPGWTKPICIGRHAFGDQYRATDIVIQESGKLKLVFVPNGHNEKKEFEVFNFTGAG 170
                         :************************* **: :*********:*::**.*:**:**** *

Populus trichocarpa      GVALAMYNTDESIRAFAEASMNTAYQKKWPLYLSTKNTILKKYDGRFKDIFQEVYEANWK 238
Ricinus communis         GVALSMYNTDESIRAFADASMNTAYQKKWPLYLSTKNTILKKYDGRFKDIFQEVYEASWK 238
Passiflora edulis        GVALSMYNTDESIRAFADASMNTAYQKKWPLYLSTKNTILKKYDGRFKDIFQEVYEASWK 238
Quercus robur            GVALSMYNTDESIHAFAEASMNTAYQKKWPLYLSTKNTILKKYDGRFKDIFQEVYETQWK 230
                         ****:********:***:************************************:.**

Populus trichocarpa      SKYEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQSDFLAQGFGSLGLMTSVLV 298
Ricinus communis         SKFEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQSDFLAQGFGSLGLMTSVLV 298
Passiflora edulis        SKFEAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQSDFLAQGFGSLGLMTSVLV 298
Quercus robur            SKFEAAGIWYEHRLIDDMVAYAMKSEGGYVWACKNYDGDVQSDFLSQGFGSLGMMTSVLV 290
                         **:******************:*********************:*******:******

Populus trichocarpa      CPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWSRGLAHRAKLDDNARLLDFT 358
Ricinus communis         CPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWSRGLAHRAKLDDNARLLDFT 358
Passiflora edulis        CPDGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWTRGLAHRAKLDDNARLLDFT 358
Quercus robur            CPDGKTIEAEAAHGAVTRHYRVHQKGGETSTNSIASIFAWTRGLAHRAKLDSNARLLDFT 350
                         **************:************************:**********.********

Populus trichocarpa      EKLEAACIGAVESGKMTKDLALLIHGSKVSRDHYLNTEEFIDAVAEELKARLSIKA 414
Ricinus communis         EKLEAACIGVVESGKMTKDLALLIHGSKVTRDQYLNTEEFIDAVAADLAERLSKA- 413
Passiflora edulis        EKLEAACVGAVESGKMTKDLALIIHGSKLSRDKYLNTEEFIDAVADELKARLSIKA 414
Quercus robur            EKLEAACVGTVESGKMTRDLALLIHGPKVTRSRYLNTEEFIDAVAEELKG------ 400
                         *******:*.*******:****:***.*::*.:************ :*
```

Appendix 1b

BLASTx hits with query the reference NAD^+ IDH (partial) transcript sequence and ClustalW alignment of the corresponding proteins

```
>protein NAD-IDH from reference sequence
MARRHIIPLLNQLTSSTRLPLTRSVTYMPRPGDGAPRAVTLIPGDGVGPLVTNAVEQVMEAMHAPVFFEKY
DVHGDLSKVPQEVIESIKKNKVCLKGG

>gi|224096682|ref|XP_002310697.1|  predicted  protein  [Populus
trichocarpa]*
MARRSIPVLKHLLTSSSTPTLRRSVTYMPRPGDGAPRPVTLIPGDGIGPLVTNAVEQVMEAMHAPVYFEK
YDIHGDMMRVPSEVIESIKKNKVCLKGGLATPMGGGVSSLNVQLRKELDLYASLVNCFNLQGLPTRHENV
DIVVIRENTEGEYAGLEHEVVPGVVESLKVITKFCSERIAKYAFEYAYLNNRKKVTAVHKANIMKLADGL
FLESCREVATKYPGIKYNEIIVDNCCMQLVSKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMPGGNVGA
DHAIFEQGASAGNVGNDKLLEQKTANPVALLLSSAMMLRHLQFPSFADRLETAVKRVISESHYRTKDLGG
TSTTQEVVDAVIGALD

>gi|255575724|ref|XP_002528761.1|  isocitrate  dehydrogenase,  putative
[Ricinus communis]*
MARRSIPILKKLLSSSNNESTCSRLVSRRSVTYMPRPGDGAPRGVTLIPGDGIGPLVTGAVEQVMEAMHA
PVYFERYEVHGDMKKVPAEVIESIKKNKVCLKGGLATPMGGGVSSLNVQLRKELDLYASLVNCFNLPGLP
TRHENVDIVVIRENTEGEYSGLEHEVVPGVVESLKVITKFCSERIAKYAFEYAYLNNRKKVTAVHKANIM
KLADGLFLESCREVATKYPGIKYNEIIVDNCCMQLVSKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMP
GGNVGADHAIFEQGASAGNVGNEKIVEQKKANPVALLLSSAMMLRHLQFPSFADRLETAVERVISEGKYR
TKDLGGDSSTQEVVDAVIAALD
```

>gi|21537157|gb|AAM61498.1| NAD+ dependent isocitrate dehydrogenase subunit 1 [Arabidopsis thaliana]*
MSRRSLTLLKNLARNANGSGIQTRSVTYMPRPGDGAPRAVTLIPGDGIGPLVTNAVEQVMEAMHAPIFFE
KYDVHGEMSRVPPEVMESIRKNKVCLKGGLKTPVGGGVSSLNVQLRKELDLFASLVNCFNLPGLPTRHEN
VDIVVIRENTEGEYAGLEHEVVPGVVESLKVITNFCSERIAKYAFEYAYLNNRKKVTAVHKANIMKLADG
LFLESCREVAKKYPSITYNEIIVDNCCMQLVAKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMPGGNVG
ADHAVFEQGASAGNVGKDKIVLENKANPVALLLSSAMMLRHLQFPSFADRLETAVKKVIAEGKFRTKDLG
GTSTTQEVVDAVIAKLD

*: accession numbers and annotations from the NCBI databank, online publicly available: http://www.ncbi.nlm.nih.gov/

Alignment of NAD^+ dependent IDH

```
Ricinus communis         MARRSIPILKKLLSSSNNESTCSRLVSRRSVTYMPRPGDGAPRGVTLIPGDGIGPLVTGA 60
Arabidopsis thaliana     MSRRSLTLLKNLARNANGSG-----IQTRSVTYMPRPGDGAPRAVTLIPGDGIGPLVTNA 55
Populus trichocarpa      MARRSIPVLKHLLTSSSTPT------LRRSVTYMPRPGDGAPRPVTLIPGDGIGPLVTNA 54
Quercus robur            MARRHIIPLLNQLTSSTRLP------LTRSVTYMPRPGDGAPRAVTLIPGDGVGPLVTNA 54
                         *:** :  * :   .:.          ************** ********:*****.*

Ricinus communis         VEQVMEAMHAPVYFERYEVHGDMKKVPAEVIESIKKNKVCLKGGLATPMGGGVSSLNVQL 120
Arabidopsis thaliana     VEQVMEAMHAPIFFEKYDVHGEMSRVPPEVMESIRKNKVCLKGGLKTPVGGGVSSLNVQL 115
Populus trichocarpa      VEQVMEAMHAPVYFEKYDIHGDMMRVPSEVIESIKKNKVCLKGGLATPMGGGVSSLNVQL 114
Quercus robur            VEQVMEAMHAPVFFEKYDVHGDLSKVPQEVIESIKKNKVCLKGG---------------- 98
                         ***********::**:*::**:: :** **:***:*********

Ricinus communis         RKELDLYASLVNCFNLPGLPTRHENVDIVVIRENTEGEYSGLEHEVVPGVVESLKVITKF 180
Arabidopsis thaliana     RKELDLFASLVNCFNLPGLPTRHENVDIVVIRENTEGEYAGLEHEVVPGVVESLKVITNF 175
Populus trichocarpa      RKELDLYASLVNCFNLQGLPTRHENVDIVVIRENTEGEYAGLEHEVVPGVVESLKVITKF 174
Quercus robur            ------------------------------------------------------------

Ricinus communis         CSERIAKYAFEYAYLNNRKKVTAVHKANIMKLADGLFLESCREVATKYPGIKYNEIIVDN 240
Arabidopsis thaliana     CSERIAKYAFEYAYLNNRKKVTAVHKANIMKLADGLFLESCREVAKKYPSITYNEIIVDN 235
Populus trichocarpa      CSERIAKYAFEYAYLNNRKKVTAVHKANIMKLADGLFLESCREVATKYPGIKYNEIIVDN 234
Quercus robur            ------------------------------------------------------------

Ricinus communis        CCMQLVSKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMPGGNVGADHAIFEQGASAGNV 300
Arabidopsis thaliana    CCMQLVAKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMPGGNVGADHAVFEQGASAGNV 295
Populus trichocarpa     CCMQLVSKPEQFDVMVTPNLYGNLVANTAAGIAGGTGVMPGGNVGADHAIFEQGASAGNV 294
Quercus robur           ------------------------------------------------------------

Ricinus communis        GNEKIVEQKKANPVALLLSSAMMLRHLQFPSFADRLETAVERVISEGKYRTKDLGGDSST 360
Arabidopsis thaliana    GKDKIVLENKANPVALLLSSAMMLRHLQFPSFADRLETAVKKVIAEGKFRTKDLGGTSTT 355
Populus trichocarpa     GNDKLLEQKTANPVALLLSSAMMLRHLQFPSFADRLETAVKRVISESHYRTKDLGGTSTT 354
Quercus robur           ------------------------------------------------------------

Ricinus communis        QEVVDAVIAALD 372
Arabidopsis thaliana    QEVVDAVIAKLD 367
Populus trichocarpa     QEVVDAVIGALD 366
Quercus robur           ------------
```

Key for the alignments:
"*" means that the residues or nucleotides in that column are identical in all sequences in the alignment
":" means that conserved substitutions are observed
"." means that semi-conserved substitutions are observed

Appendix 2a

> Reference sequence of $NADP^+$ IDH gene

TCCCATCGTTGAAATGGATGGTAATTATTAATTCACACCGTTCATCATCTATTACAATCATGTTTTAATTA
TTACTACATATATAATCATTCAATTTTTACAAAAACCCAGAAACAAAAACATGTTGCAAATCATGTTCTAA
TTGTTCATGTGTGTTGGTTTGATATCAAAATGAACCATGAAAAGCACCACTACATGAATTATGATAACAGG
ACTACTTTGACTCATCGAGTATGTAATTTTCATCAGCTATTTTTTTATATTCAAAACATTGCTTTACTGAT
GCAGGAGATGAAATGACTCGAATCTTTTGGAAATCAATCAAAGACAAGGTCCTTTTACATTCGTAATTCAT
AATTCATCTCCTTTTTTTTTTTTTTTCTTGGGTATTGTGTAAATATTGTATTGATTCTCCACAGCTTATAT
TTCCCTTTGTGGAGTTGGATATCAAGTACTTTGATCTTGGTCTCACTAATCGTGAGGCCACCAATGATAAC
GTTACAATTGAAAGTGCTCAAGCCACTCTCAGGTATATATATGTCTTTATCTTTTTTCTCTCATAGTTCAT
ACCTTCTATTTTCTTAACATCATCATCATAATCATCATTTTTTTTTTTTTTTTTAAGATTGATAAATTATAC
ACGCACCCAGTGAGATTTGAAACCATGATCTAACCCCACCGTTCGTTTCCAGAGGAGGAACTTAAGTTTTA
TTTGAGTTAGTGCTTAATTAACATTATCAATTATCATCATTTTCTGCAAAATATGCCTCATTGACATTTCA
ATTAGTGTTGGGTTTCCTTTTGAAGAATGCATTGCTATTTAGCATTGTCTCAAAGGAAGAGATTGTGATTG
CTTGTCTCTCTCTCTCTTTTTTTTTTTTTTTTTTTTTTGGTTCGAATAGCTTGCTTGTATAGTTGTTAGAACT
TAGAAGTGTTCTCAATTCCTGTGGAGGCTACTTACTCATATGTACTGTCTTTAACAAAATTTTGCTACCCT
TTAGTGCCATTTATTTTTGTTGTTTAATTCATCCTATAATTTGGCCTTTCTAGGTACAATGTAGCAATAAA
GTGTGCAACTATAACTCCAGGTACTCATTCCTAGCCTCAATTGGAATGACCCAAATTTAAAAGATGAAATT
TTGGCCCAGTAACTGGCTTATGTGTAAAATTGGTTTTGGTTTTTGATAGATGAAGCTCGTGTCAAGGAGTT
CAACTTGAAACAAATGTGGAAGAGTCCAAATGGGACAATTCGGAATATTTTAAATGGTCAATGAACTTCCC
TGATGGTTTTTTTTGTGTCCTTTTATTTGCTCCAAATTTTATTTTTATTAAATTAAGGGGGACATTTTGTT
TTCCATTTACCACAGGTACTGTTTTCAGAGAGCCTATTATCTGCAAAAACATCCCCCGTCTTGTCCCAGGT
TTGTAATTTTTTTCTTTTTCAGTGCCGGTTTCCTTTTCATTATGAAGCCAAGCACTAGGTTGACTTGGTTT
TCTCTTCAGGTTGGACAAAGCCAATATGCATTGGAAGGCATGCTTTTGGTGATCAGTACAGGGCAACTGAT
ATTGTTATACAAGAATCTGGAAAGCTTAAATTGGTATTTGGTAAAATCCTACTTTCAGTCATGAGCTTCTC
AATACTCTTTTTAGTTTCACTTTTTATGGAGACTTGTTAAAGTCTGCACAAAAGCTGAATTCAGCTCAGTT
TTTCTGTCCTATATTTCCATCAATGCCTTCATACGTAATAATAATCATCTTATCAAGGGGAATTGGAATAT
CTGCAGTACCCAATGGACATAATGAGAAGAAAGAGTTTGAGGTTTTCAACTTTACGGGTGCTGGAGGTGTA
GCCTTGTCCATGTACAACACTGATGAGGTTCAGATTTCCATAAATGCTTGTTCTTTTCCCTGCTTCCTTTA
ATACTTGCACCGACCCTGAAATTGACTTGCCAGAAGAGTTTTTTGATGTTGTTTTTATTTTTCATTTTTTT
GCCTTGATAACAGTCTATTCATGCTTTTGCTGAGGCTTCAATGAATACTGCTTACCAGAAAAAGTGGCCAC
TTTATCTTAGCACTAAAAATACCATTCTTAAGAAATATGATGGAAGGTATGCTGTTACTTGTTGGACTTTC
TTTTCTCTTCTCATTCTGCTCAATTTTTGTTGCAAACTAAATTATCAAAGCCCCTCTTGTACCTTTGTGTG
TGAAAGAGAAAGAGACAGAGAGAGAAATGGGGTGTTACTGTATGATTTTCTGTGATTTTTTTGGGGTCATT
TTTTGTCAATGTGACTTGATTATCTGTCTTATGCTCTTACTGATTATGTATTGTCATTTTATTATTACGTG
CTTTTCATGAAGATTCAAGGACATATTTCAGGAAGTTTATGAAACTCAATGGAAATCTAAGTTTGAAGCTG
CAGGGATATGGTGAGATATCAAACTATCATGCTAATTAGTTTGTACAGTTTGAGAAGCTTGTCATCAACTT
ACAGCTGAATTGGTAGGTATGAACACCGTCTCATAGATGATATGGTTGCTTATGCCATGAAAAGTGAAGGA
GGTTATGTATGGGCTTGCAAAAACTATGATGGGGATGTGCAGAGTGATTTCTTAGCCCAAGGTTTGCATCA
ATATGGAAATAATCTGTGATTGACATATCATTCTTTGTAACTTCTGCTACTTTCGTTGTAAGTAAATGGAA
AGGTGCTTGCTCTGTGGGTGTCCTAGCCAAGGTATTCAATTATTAATTTAGCTTATGATAAAAAAATTTAT
CATTGATACAGGATTTGGATCTCTTGGGATGATGACATCGGTGCTGGTATATCTTGCTGTTCTCCTTTTTG
TCTTCCCCAATTATTTTTGCATCTCATTTTGGTCCATTTTTCACTTCTAAACCTAATCATAAGAAGTTGGA
GTGGGATTGCAGGTGTGCCCAGATGGAAAGACTATTGAAGCTGAAGCAGCCCATGGGGCAGTGACCCGCCA
CTACCGGGTCCATCAGAAAGGTGGTGAAACCAGCACAAACAGCATAGCATCCATTTTTGCTTGGACTCGAG
GTCTTGCACACAGGTATCATATGTATTTATTCTTTTATATGGATGAATTCTTGTACAATTTTTGTTCTTGA
AAGTTAAAAATTTTAGGAATGGGATTCTTTTCTTATTAGATCATCAAATAAAAATTGAGTAAATCAATGTT
ATCAGGGCAAAGTTGGACAGCAATGCCAGACTTTTGGATTTTACTGAGAAATTGGAAGCAGCCTGTGTTGG
AACGGTGGAATCAGGAAAGATGACTAGGGATCTTGCACTTCTTATTCATGGGCCCAAGTAATTTCTGTTAA
AAGGATATTTCTGTAGAAAGTTCTGCTGTTGTGCTTAATTTTTTTGTATAATCAAAACTTTCATGTTTGGT
ATAGGGTTACTAGGTCTCGGTATCTAAATACCGAAGAGTTCATTGATGCTGTAGCTGAGGAGCTTAAGGGC
A

Appendix 2b

> Reference sequence of NAD^+ IDH gene

```
CATCCTAACCTCACCCCAAGAGGCACGCACAAAACCCCAAAATAAACCCACATGGCTCGTCGACACATCAT
CCCTCTCCTCAACCACCTCACCTCCTCCACGCGCCTCCCTCTCACGCGCTCCGTCACCTACATGCCACGAC
CGGGCGACGGCGCCCCACGCGCCGTGACCTTGATCCCGGGTGACGGGGTGGGTCCACTGGTGACAAACGCG
GTCGAACAGGTGATGGAGGCCATGCACGCGCCGGTGTTCTTCGAGAAGTACGATGTACACGGGGACCTGAG
CAAGGTCCCGCAGGAGGTGATCGAGTCGATTAAGAAAAACAAAGTGTGTCTGAAAGGTGGGCTTAGGACAC
CGGTGGGTGGTGGAGTCAGCTCGCTCAATGTTCAGTTGAGGAAAGAGCTGGATTTGTATGCTTCGTTGGTC
AACTGTTTCAACCTTCCTGGGCTTCCTACGAGGCACGAAAATGGGATATTGTTGTGATTCGAGAGAATACT
GAGGGTGAGTACTCGGGTCTCGAGCATGAAGTTGTTCCTGGCGTTGTTGAAAGCCTCAAGGTACATTCTTT
TTGTTTTCATTTTCAACTTTTCAATTTCCAATTCTATGTTGTACAATCAAATTTTACTCTTTATGATTAGA
GCATTGTTTTCATGTGGACCCCATCATTGATGTTACTTTTTTTATTACACAACAATTAAAAAAATGCTACA
CTAGCAGTCTAGCACTAATCGATAATGTTGTAAAGAGTATCATCTTTATATAAAAATACAATGGGTTATTT
GTTGGAGTATCATTTGAATTTTGTTGTCATTAGTTAATTTAGCACTTTCTATGGAAAGAGATGGCTTTTAG
GGTCTAATCCGTTGAGCTCTAGCTCAAATTTTACCGAGGGCGAGGTTGTGGGATCAAAACCTGTTGGATGC
ACAAAACCTTGGGCTGCCTAGTTAAATTATAATATTCAATGATGTTGCTATATTGGAATGCAGTACCAACA
GTGTGATCTTCACTACCAACAATTAAATTTAATCGCGTGGCTTATTGATAATGTAACCACATGGGGAAATT
AAATGGGGAACCCAAGGTAGAACTACTAAGAAACTGTACCAAAGTTTTGCCCTTGATTAGAACCCCCTTAT
GGTTTTGGAATAATAAAAACAAATAGTTTGTATGTTTTTCATTCCAATTGATGAATTTTCCAACTTTACGT
AATTGTATATTGCTTTCCACACAACATTTTAACTTTAGGGTATGTTTGGAGTGGAGGAGGGAGATGGAGAA
AGAGTATAGGGGAAAGAGGTTAAAATATATAGAAATTTGTACATTCCATTCACCTTCCTTTCTCCTCCCTT
CTCCTCTCCCACTTTCTCCCTGCACCCAAACATAGAGTTATGTTAGGCCTTAAATGGGGTTTAAAATTTAA
AGTTTCTAAGCTTCATAGAAGATGTAGTACAATAATCATGTTCTTGGATTTGAAATCATAATGAGGTAGTG
AGTTTTGGCTCAAGTGCCACTTTCTTTTATCATAAGAATGAGTGGAGGGTGAGGTTGTTGGTTCAAACCAC
TGGATGCGTGTGAAACGTATGAGAAAAAAAAACGATAACAAAGCCTTAGTCCCCAAACTATGGGGTTGGCT
ATGGACTTTAACTAGACTAATTAGGGTTGATTACATGTATTCTTTCATGTTATATAATGTTATCCGCAATT
GTACTCCGTCACCTCCATAAAATTAGAAATAAAACCTTGTACCTTATTCAGAAGGTTAAATTTTGTATCTT
ATGCAACATGTGCATGTCACAAGGAAGAATTTGTCAATTGGAACAAACCGATTCATTTTAGCCCAAGGCAT
GGGGAGTTTTAGTGTTGGTTCCTAAAATTTCTAATTATCTATCTATCTATCTATAATGTATTATTATTATA
AAGGGATTGTTTACAATTTGGGATTTGTTAGGTGATAACAAAGTTCTGCTCAGAACGTATTGCAAAATATG
CCTTTGAGTATGCTTATCTGAACAACAGAAAAACGGTGACAGACCAGCCCGGGCCGTCGACCACGCGTGCC
CTATAGTGAGTCGTATTAC
```

Appendix 3

Pairwise differentiation measurements, G_{st} (lower diagonal) and F_{st} (upper diagonal) for each locus of $NADP^+$ IDH and for NAD^+ IDH gene (continued on the next page)

		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#1	*Q. robur*	*	-0,01695	-0,07914	-0,10236
	Q. petraea	-0,01604	*	0	-0,04396
	Q. pubescens	0	0,02564	*	-0,00607
	Q. frainetto	0	0,02564	-0,02041	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#2	*Q. robur*	*	-0,07735	0,07738	-0,07383
	Q. petraea	-0,02041	*	0,09218	-0,14486
	Q. pubescens	0,02041	0	*	-0,02
	Q. frainetto	-0,01604	0,02564	-0,03825	*

		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#3	*Q. robur*	*	0,03846	0,27083	-0,00467
	Q. petraea	0,00524	*	0,04494	-0,14754
	Q. pubescens	0,03226	-0,0303	*	0,10112
	Q. frainetto	-0,02041	-0,03825	-0,01124	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#4	*Q. robur*	*	0,25	0,33594	0,3125
	Q. petraea	0,07975	*	0,02174	0,04167
	Q. pubescens	0,14286	0,09091	*	0,14384
	Q. frainetto	0,1018	0,09091	0,0411	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#6	*Q. robur*	*	-0,03125	0,24731	-0,08824
	Q. petraea	0,01099	*	0,02174	-0,08434
	Q. pubescens	0,03226	-0,01124	*	-0,01974
	Q. frainetto	-0,05882	0,03226	-0,01124	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
#7	*Q. robur*	*	-0,05556	0,375	0,01639
	Q. petraea	0,06667	*	0,22414	-0,01974
	Q. pubescens	0,21212	0,00524	*	0,01235
	Q. frainetto	0,21212	0,02564	-0,03448	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
Total $NADP^+$ IDH	*Q. robur*	*	0,00446	0,16833	-0,0186
	Q. petraea	0	*	0,10946	-0,06938
	Q. pubescens	0	0	*	0,01303
	Q. frainetto	0	0	0	*
		Q. robur	*Q. petraea*	*Q. pubescens*	*Q. frainetto*
NAD^+ IDH	*Q. robur*	*	-0,1413	0,01786	0,12338
	Q. petraea	-0,01604	*	-0,05469	0,01235
	Q. pubescens	-0,03825	-0,02041	*	0,14948
	Q. frainetto	0,03226	0,00524	0,02564	*

Appendix 4a

Significant LD for the parsimony informative sites of the pooled data set for the $NADP^+$ IDH and the corresponding distances (significant comparisons with χ^2 test with Bonferroni's adjustment) LD estimates: *D* (Lewontin and Kojima 1960), *D'* (Lewontin 1964) and R (Hill and Robertson 1968) (continued on the next page)

Site 1	Site 2	Distance	D	D'	R	Fisher	chi-square
133	1901	1746	0.128	1.000	1.000	0.001***	20.000***B
133	1587	1432	0.128	1.000	1.000	0.001***	20.000***B
816	1847	1010	0.090	1.000	1.000	0.005**	20.000***B
816	1828	991	0.090	1.000	1.000	0.005**	20.000***B
880	1847	947	0.090	1.000	1.000	0.005**	20.000***B
880	1828	928	0.090	1.000	1.000	0.005**	20.000***B
1504	1950	426	0.188	1.000	1.000	0.000***	20.000***B
572	950	377	0.090	1.000	1.000	0.005**	20.000***B
1504	1870	346	0.188	1.000	1.000	0.000***	20.000***B
617	950	332	0.090	1.000	1.000	0.005**	20.000***B
625	950	324	0.090	1.000	1.000	0.005**	20.000***B
572	887	314	0.090	1.000	1.000	0.005**	20.000***B
1587	1901	314	0.128	1.000	1.000	0.001***	20.000***B
664	950	285	0.090	1.000	1.000	0.005**	20.000***B
617	887	269	0.090	1.000	1.000	0.005**	20.000***B
625	887	261	0.090	1.000	1.000	0.005**	20.000***B
572	824	252	0.090	1.000	1.000	0.005**	20.000***B
725	950	224	0.090	1.000	1.000	0.005**	20.000***B
726	950	223	0.090	1.000	1.000	0.005**	20.000***B
664	887	222	0.090	1.000	1.000	0.005**	20.000***B
732	950	217	0.090	1.000	1.000	0.005**	20.000***B
733	950	216	0.090	1.000	1.000	0.005**	20.000***B
617	824	207	0.090	1.000	1.000	0.005**	20.000***B
625	824	199	0.090	1.000	1.000	0.005**	20.000***B
572	733	161	0.090	1.000	1.000	0.005**	20.000***B
725	887	161	0.090	1.000	1.000	0.005**	20.000***B
572	732	160	0.090	1.000	1.000	0.005**	20.000***B
664	824	160	0.090	1.000	1.000	0.005**	20.000***B
726	887	160	0.090	1.000	1.000	0.005**	20.000***B
572	726	154	0.090	1.000	1.000	0.005**	20.000***B
732	887	154	0.090	1.000	1.000	0.005**	20.000***B
572	725	153	0.090	1.000	1.000	0.005**	20.000***B
733	887	153	0.090	1.000	1.000	0.005**	20.000***B
824	950	125	0.090	1.000	1.000	0.005**	20.000***B
617	733	116	0.090	1.000	1.000	0.005**	20.000***B
617	732	115	0.090	1.000	1.000	0.005**	20.000***B
617	726	109	0.090	1.000	1.000	0.005**	20.000***B

Site 1	Site 2	Distance	D	D'	R	Fisher	chi-square
617	725	108	0.090	1.000	1.000	0.005**	20.000***B
625	733	108	0.090	1.000	1.000	0.005**	20.000***B
625	732	107	0.090	1.000	1.000	0.005**	20.000***B
625	726	101	0.090	1.000	1.000	0.005**	20.000***B
625	725	100	0.090	1.000	1.000	0.005**	20.000***B
725	824	99	0.090	1.000	1.000	0.005**	20.000***B
726	824	98	0.090	1.000	1.000	0.005**	20.000***B
572	664	92	0.090	1.000	1.000	0.005**	20.000***B
732	824	92	0.090	1.000	1.000	0.005**	20.000***B
733	824	91	0.090	1.000	1.000	0.005**	20.000***B
1870	1950	80	0.188	1.000	1.000	0.000***	20.000***B
664	733	69	0.090	1.000	1.000	0.005**	20.000***B
664	732	68	0.090	1.000	1.000	0.005**	20.000***B
816	880	63	0.090	1.000	1.000	0.005**	20.000***B
887	950	63	0.090	1.000	1.000	0.005**	20.000***B
664	726	62	0.090	1.000	1.000	0.005**	20.000***B
824	887	62	0.090	1.000	1.000	0.005**	20.000***B
2945	3007	62	0.160	1.000	1.000	0.000***	20.000***B
664	725	61	0.090	1.000	1.000	0.005**	20.000***B
572	625	53	0.090	1.000	1.000	0.005**	20.000***B
617	664	47	0.090	1.000	1.000	0.005**	20.000***B
572	617	45	0.090	1.000	1.000	0.005**	20.000***B
625	664	39	0.090	1.000	1.000	0.005**	20.000***B
2294	2331	37	0.090	1.000	1.000	0.005**	20.000***B
2482	2515	33	0.128	1.000	1.000	0.001***	20.000***B
135	161	26	0.128	1.000	1.000	0.001***	20.000***B
1044	1070	26	0.128	1.000	1.000	0.001***	20.000***B
1828	1847	19	0.090	1.000	1.000	0.005**	20.000***B
617	625	8	0.090	1.000	1.000	0.005**	20.000***B
725	733	8	0.090	1.000	1.000	0.005**	20.000***B
773	781	8	0.160	1.000	1.000	0.000***	20.000***B
51	58	7	0.188	1.000	1.000	0.000***	20.000***B
725	732	7	0.090	1.000	1.000	0.005**	20.000***B
726	733	7	0.090	1.000	1.000	0.005**	20.000***B
1017	1024	7	0.090	1.000	1.000	0.005**	20.000***B
726	732	6	0.090	1.000	1.000	0.005**	20.000***B
725	726	1	0.090	1.000	1.000	0.005**	20.000***B
732	733	1	0.090	1.000	1.000	0.005**	20.000***B

Appendix 4b

Significant LD for the parsimony informative sites of the pooled data set for the NAD^+ IDH and the corresponding distances (significant comparisons with χ^2 test with Bonferroni's adjustment) LD estimates: *D* (Lewontin and Kojima 1960), *D′* (Lewontin 1964) and R (Hill and Robertson 1968)

Site 1	Site 2	Distance	D	D'	R	Fisher	chi-square
287	300	13	0.113	0.728	1.000	0.009**	10.588***B

8. Curriculum vitae

AMARYLLIS A. VIDALIS

Personal data:
Date of birth: April 7th 1981
Place of birth: Athens, Greece
Marital status: single

Education:

- **2006 - 2010** Scholarship holder: "Sophia Chlorou" endowment of National Technical University of Athens, Ph.D. candidate at the Georg-August Universität Göttingen, Faculty of Forest Sciences and Forest Ecology, Büsgen Institute, Department of Forest Genetics and Forest Tree Breeding, Germany
- **2004 (SS)** Scholarship holder (IKY- Erasmus), Institute of Forest Genetics and Forest Tree Breeding, Georg-August-University, Göttingen: Experimental work for the Diploma Thesis
- **2003** Practical work at the National Agricultural Research Foundation, Institute of Mediterranean Forest Ecosystems & Forest Products Technology, Department of Forest Genetics, Greece
- **1999-2004** Democritus University of Thrace, Faculty of Forestry, Environmental Management and Natural Resources, Greece
- Diploma Thesis: "Genetic variation of European beech (*Fagus sylvatica* L.) in greek Rodopi Mt. applying molecular markers" (in greek)

Working experience:

- **2005 - 2006** Employed at the Head Office of the Forest Department of the District of Corinth, Ministry of Agricultural Development, Greece

Publications:

- Papageorgiou, A.C., Vidalis, A., Gailing, O., Tsiripidis, I., Hatziskakis, S., Boutsios, S., Galatsidas, S., Finkeldey, R.: Genetic variation of beech (*Fagus sylvatica* L.) in Rodopi (N.E. Greece). European Journal of Forest Research 127, 81-88, 2008. http://dx.doi.org/10.1007/s10342-007-0185-3

Certified language skills:

- English (Cambridge - Proficiency in English)
- German (Goethe Institut - Mittelstufe)
- Italian (basics)
- Greek (native)