Whole-genome screening for near-diagnostic genetic markers for white oak species identification in Europe

Antoine Kremer[1], Adline Delcamp[2], Isabelle Lesur[3], Stefanie Wagner[4], Christian Rellstab[5], Erwan Guichoux[2,*], Thibault Leroy[1,6] *

[1]: UMR BIOGECO, INRAE, Université de Bordeaux, 69 Route d'Arcachon, 33612 Cestas, France

[2]: UMR BIOGECO, INRAE, Université de Bordeaux, PGTB, 69 Route d'Arcachon, 33612 Cestas, France

[3]: Helix Venture, 33700 Merignac, France

[4]: UMR CAGT, CNRS, Université Paul Sabatier, 37 Allées Jules Guesde, 31000Toulouse, France

[5]: Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland

[6]: GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France

*: These authors contribute equally to the research.

Corresponding author: Antoine Kremer

Email: antoine.kremer@inrae.fr

Phone: (33) 535385365

Orcid ID: 0000-0002-3372-3235

Declarations:

Ethics approval and consent to participate: not applicable

Consent for publication: not applicable

Availability of data: The data that support the findingd of this study will be available at the publicly accessible data repository of INRAE : The url address will be completed after the manuscript is accepted for publication.

Competing interests: there are no competing interests

Funding: This research was supported by the European Research Council through an Advanced Grant (project TREEPEACE # FP7-339728), by an ANR (Agence Nationale de la Recherche) Grant (project GENOAK 2022, #BSV6-009-02), and by the EVOLTREE Opportunity call (project OakID2).

Authors' contributions: Conception of the study: TL, AK; Sampling and collection of material: AK, CR, TL; Discovery of near-diagnostic markers in pool-sequenced resources: TL; Discovery of near-diagnostic markers in sequence captured resources: IL; Design of multiplexes and genotyping of natural populations: EG, AD; Data analysis: AK, TL, SW; Writing of the manuscript: AK, EG, TL. All authors reviewed the manuscript.

Acknowledgements**:** We thank colleagues that contributed to the collection of material made for this study: Dalibor Ballian (Bosnia and Herzegovina), María Valbuena Carabaña and Luis Gil (Spain), Giovanni Giuseppe Vendramin (Italy). We extent our appreciation to partners of the former EU supported FAIROAK and OAKFLOW projects, and of the EVOLTREE Network of Excellence, who collected material included in this study. The MassArray genotyping was performed at the PGTB (doi:10.15454/1.5572396583599417E12) with the help of Laure Dubois, Céline Lalanne and Marie Massot. We are grateful to François Ehrenmann for his contributions to the figures of the manuscript.

Key message:

Mining genome-wide DNA sequences enabled the discovery of near-diagnostic markers for species assignment in European white oaks despite their low interspecific differentiation.

Abstract:

Context: Identifying species in the European white oak complex has been a long standing concern in taxonomy, evolution, forest research and management. *Quercus petraea, Q. robur, Q. pubescens* and *Q. pyrenaica* are part of this species complex in western temperate Europe and hybridize in mixed stands, challenging species identification.

Aims: Our aim was to identify diagnostic single nucleotide polymorphisms (SNPs) for each of the four species that are suitable for routine use and rapid diagnosis in research and applied forestry.

Methods: We first scanned existing whole-genome and target-capture data sets in a reduced number of samples (training set) to identify candidate diagnostic SNPs, ie genomic positions being characterized by a reference allele in one species and by the alternative allele in all other species. Allele frequencies of the candidates SNPs were then explored in a larger, range-wide sample of populations in each species (validation step).

Results: We found a subset of 38 SNPs (ten for *Q. petraea*, seven for *Q. pubescens*, nine for *Q. pyrenaica* and twelve for *Q. robur*) that showed near-diagnostic features across their species distribution ranges with *Q. pyrenaica* and *Q. pubescens* exhibiting the highest and lowest diagnosticity, respectively.

Conclusions: We provide a new, efficient and reliable molecular tool for the identification of the species *Q. petraea, Q. robur, Q. pubescens* and *Q. pyrenaica*, which can be used as a routine tool in forest research and management. This study highlights the resolution offered by whole-genome sequencing data to design diagnostic marker sets for taxonomic assignment, even for species complexes with relatively low differentiation.

Keywords: *Quercus*, diagnosticity, genetic differentiation, pool-seq data, captured sequences

1. Introduction

Identifying species in the European white oak complex has been a long-standing concern in evolutionary biology as well as in forest research and management. According to the latest taxonomic classification, there are about fifteen oak species in Europe, which form the subsection of the Roburoids within the *Quercus* section (white oak section) (Denk *et al* 2017; Hipp *et al*, 2020). Within the continent, however, species richness varies, with higher species diversity in the Mediterranean region and in Eastern Europe compared to other areas (Camus, 1938; Le Hardy de Beaulieu and Lamant, 2006). In western temperate Europe, four white oaks species occur north of the Pyrenees and Alps (*Q. petraea, Q. robur, Q. pubescens* and *Q. pyrenaica*). Co-occurrence of all four species in the same forest is rare. The few reported cases indicate extensive gene flow and admixture between all four species, leading to considerable morphological variations and uncertainties when it comes to taxonomic classification based on morphological characters (Lepais *et al*, 2013; Leroy *et al*, 2017; Viscosi *et al*, 2009). The co-occurrence of the three species *Q. petraea*, *Q. robur* and *Q. pubescens* is more common, especially in the southern parts of the temperate range, for which hybridisation and morphological variation is well documented (Dupouey and Badeau, 1993; Grandjean and Sigaud, 1987; Macejovsky *et al*, 2020; Rellstab *et al*, 2016). Finally, forests with co-occurrences of two species and interspecific admixture have also raised questions about species classification. This is especially true for co-occurrences of *Q. petraea*-*Q. robur* (Bacilieri *et al*, 1995; Jurksiene and Baliuckas, 2014; Kelleher *et al*, 2005; Kremer *et al*, 2002; Yucedag and Gailing, 2013), but also for *Q. petraea*-*Q. pyrenaica* (Lopez de Heredia *et al*, 2009) and *Q. petraea*- *Q. pubescens* (Bruschi *et al*, 2000, Reutimann *et al*, 2020, 2023). This brief overview of species admixture and problems of taxonomic classification based on morphological characteristics highlights the pressing need for a time and cost efficient molecular tool for reliable species assignment within European white oaks for use in forest science and management.

In response to this challenge, molecular tools have been continuously improved and a number of species marker kits have been developed and applied during the last decade (Guichoux *et al*, 2011; Neophytou, 2014; Reutimann *et al*, 2020, Degen *et al*, 2021; Schroeder and Kersten, 2023). These methods have set new milestones for the delimitation of oak species, but their validity has been constrained by some biological and technical limitations. From a biological

point of view, the markers used in the kits are still shared between the species, although interspecific differentiation of the selected markers was higher than in earlier studies. From a technical point of view, the genomic resources explored for selecting the marker candidates was very limited until recently. Using previously published genome-wide data and genome scans targeting genomic positions that maximise differentiation between populations of *Q. robur*, *Q. petraea*, *Q. pubescens* and *Q. pyrenaica*, we overcame these limitations and designed a new single-nucleotide (SNP) marker set for range-wide species identification in European white oaks.

Earlier genome scans for species differences showed that interspecific differentiation (F_{ST}) followed an L-shaped distribution suggesting that there might be highly differentiated markers at an extremely low frequency within the genome (Reutimann *et al*, 2020; Scotti-Saintagne *et al*, 2004). Recent analysis of nucleotide diversity in genes underlying species barriers between European white oaks confirmed these expectations (Leroy *et al*, 2020b). Our approach built on these results by launching a systematic search of so-called species "diagnostic" SNPs within existing genome-wide resources. Ideally a diagnostic SNP contains a diagnostic allele of a given species that is fully fixed in that species and the alternate allele fixed in the other species. Earlier surveys (Scotti-Saintagne *et al*, 2004; Reutimann *et al*, 2020; Lesur *et al*, 2018) in European white oaks indicated that such ideal cases rarely exist. However, some markers exhibit species frequency profiles close to the ideal case (so-called near-diagnostic SNPs; for example an allele with a frequency larger than 0.9 in the target species, and alternate allele frequency larger than 0.9 in all other species) (Schroeder and Kersten, 2023). Only a few of such markers would then be enough to correctly assign trees to the correct species using appropriate analytical approaches. For example, Reutimann *et al* (2020) showed that five SNPs were enough for correctly classifying 95% of *Q. robur* reference trees, although the single SNPs were far from being diagnostic.

In this study we explored pool-sequenced whole-genome libraries of natural populations of four white oak species (*Q. petraea, Q. pubescens, Q. pyrenaica* and *Q. robur*) (Leroy *et al*, 2020b), and genome-wide capture-based sequences of *Q. petraea* and *Q. robur* (Lesur *et al*, 2018) to identify near-diagnostic SNPs for each of the four species. We describe the approaches and methods used to discover near-diagnostic SNPs, and explore the stability of diagnosticity over the distribution range of the four species.

Our main goal was to identify and validate a new set of near-diagnostic SNPs that can be used in the development of an efficient and cost-effective molecular tool for forest research and management. To this end, we focused on the variation of near-diagnostic SNPs across species,

between populations within each species and between SNPs. We finally addressed the evolutionary drivers that may have contributed to the maintenance and/or modification of diagnosticity within the genome, and throughout the distribution range of the four species.

2. Material and Methods

2.1. Discovery of near-diagnostic markers

The discovery of near-diagnostic SNPs was conducted by scanning oak genomic data that have been generated in earlier studies assessing genomic diversity and differentiation in the four sympatric white oak species (*Quercus petraea, Q. pubescens, Q. pyrenaica, Q. robur;* Leroy *et al*, 2017 and 2020b, Lesur *et al*, 2018).

2.1.1. Discovery of near-diagnostic SNPs in whole genome pool-sequenced (pool-seq) resources

2.1.1.1. Pool sequencing

In Leroy *et al*, 2020b, we used leaf and bud samples from up to 20 adult trees of the four species coming from four different forests located at maximum 200 km away from each other in South West of France (Table 4 in Appendix). The sampled stands were of mixed oak composition (generally two or three species) and of natural origin. DNA extracts were pooled in equimolar amounts to obtain a single pool for each species. Libraries were then sequenced on nine to ten lanes for each of the four species (1 pool per species) on a Illumina HiSeq 2000 sequencing platform (Leroy *et al*, 2020b for details). In this study, to reduce the computation load, we only used two lanes per pool from SRA, namely ERR2215923 and ERR2215924, ERR2215937 and ERR2215938, ERR2215909 and ERR2215910, and ERR2215916 and ERR2215917 for *Q. pubescens, Q. petraea, Q. pyrenaica* and *Q. robur* respectively. Raw reads were then trimmed using Trimmomatic (v. 0.33, Bolger *et al*, 2014) to remove low quality bases using the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50.

2.1.1.2. Mapping and SNP calling

Data from two sequencing lanes per species (from up to 10 lanes per species in Leroy *et al*, 2020b) were then mapped against the oak haplome assembly ("PM1N", Plomion *et al*, 2018) using bwa mem (Li, 2013). PCR duplicates were removed using Picard v. 1.140 (http://broadinstitute.github.io/picard/). Samtools v.1.1 (Li, 2011) and PoPoolation2 v. 1.201

(Kofler *et al*, 2011) were then used to call bi-allelic SNPs with at least 10 reads of alternate alleles and a depth between 50 and 2000x at each position. To ensure a reasonably low rate of false positives due to Illumina sequencing errors, all SNPs with a MAF lower than 0.05 were discarded. A total of 24,345,915 SNPs were identified and then screened for their diagnostic value (see next paragraph).

2.1.1.3. Genome scan for near-diagnostic SNPs

Allele frequencies were computed from the SNP-frequency-diff.pl script of PoPoolation2. SNPs exhibiting a high difference in allele frequency (Δp>0.9 between the focal species and all other species) were then selected. All candidate diagnostic SNPs with a coverage lower than 80 in the four populations were discarded, in order to ensure that the high Δp was not associated with inaccurate allele frequency estimation in low coverage regions. Despite the relatively limited linkage disequilibrium in oaks (Coq-Etchegaray *et al*, 2023) even in species barrier regions (Leroy *et al*, 2020b), the relatively high nucleotide diversity in oaks (Plomion *et al*, 2018; Saleh *et al*, 2022) allows several neighboring SNPs to be identified by this screening. We therefore selected the best SNPs per identified region considering the constraints associated with the SNP design (see below).

2.1.2. Discovery of near-diagnostic SNPs in sequence-captured (seq-cap) genomic resources

In addition to the pool-seq resources, we mined a separate genome-wide resource that came from a sequence capture experiment of *Q. robur* and *Q. petraea* aiming at calling SNPs for inferring genomic relatedness among trees (Lesur *et al*, 2018). Here, the discovery population consisted of a far larger panel (245 adult trees in total) equally distributed between *Q. petraea* and *Q. robur* growing in the Petite Charnie forest located in the western part of France (Table 4 in Appendix). We used the capture data in complement of the pool-seq data to ensure a higher diagnosticity of the markers for this specific pair, given the larger panel of *Q. robur* and *Q. petraea* samples available in the capture data. The capture-based assay consisted in sequencing 2.9 Mb (15 623 target regions) on an Ion Proton System (Thermo Fisher, Scientific, Waltham, MA, USA) covering both genic and intergenic regions and resulted in the calling of more than 190,000 SNPs with a coverage of more than 10x (Lesur *et al*, 2018). The study provided allele frequencies of each SNP, and we screened the total set of SNPs for their differentiation between *Q. petraea* and *Q. robur*, by ranking their F_{ST} values to complete the discovery panel. Although limited to two species (*Q. petraea* and *Q. robur*), this data set

corresponded also to a genome-wide exploration of species differentiation implemented on a larger population sample (Lesur *et al*, 2018). It was therefore selected for this study, pending its relevance for selecting near-diagnostic markers for the remaining two species (*Q. pubescens* and *Q. pyrenaica*), which is investigated in this study.

2.2. Training and validation of near-diagnostic SNPs

2.2.1. Training populations

The candidate diagnostic SNPs of the discovery panel were first tested on a limited number of oak individuals, with one sample per population for up to nine populations per species (19 to 48 samples per species) originating mainly from the western part of Europe (Table 5 in Appendix). The training experiment was conducted over two sessions that took place during two periods (training 1 and training 2, Table 5 in Appendix) with different samples (but from the same geographic range). The two sessions differed only by the samples included which was constrained by the availability of the material. The objective of the training step was to check whether the candidate SNPs exhibited near-diagnostic frequency profiles in natural populations originating mainly from the area of the discovery panel. The training step also included quality controls and repeatability assessments of the genotyping assay (see results paragraph 3.3.1).

2.2.2. Validation populations

Given that the discovery and training of diagnostic SNPs was implemented on limited number of trees originating mostly from the western part of the distribution of the four species, we included a round of validation by increasing the sample sizes of the training populations and enlarging the collection of populations, studying the SNP diagnosticity across a larger part of the species' natural distribution (Figure 1). Additionally, the validation step aimed also at reducing the number of SNPs, while still maintaining overall multilocus diagnosticity, in order to produce a low cost and easy to use screening tool in operational forestry. In total, 24 populations of *Q. petraea*, 10 of *Q. pubescens*, 6 of *Q. pyrenaica* and 19 of *Q. robur* were part of the validation set, representing in total 1,123 trees (Figure 1 and Table 5 in Appendix). All samples were collected in natural populations and their taxonomic status was assessed by the local collectors based on leaf morphology. Sampled populations were in most cases of mixed oak composition. Some of the populations were used in earlier large-scale genetic surveys (Gerber *et al*, 2014; Kremer *et al*, 2002), others were purposely collected for this study.

2.2.3. Genotyping assay

Medium-throughput SNP genotyping assays were implemented on single tree DNA extracts using the MassARRAY® technology (Agena Bioscience, San Diego, CA, USA). The assay design, using the MassARRAY Assay Designer version 4.0.0.2, was performed on candidate SNPs from pool-seq and seq-cap resources. Nine multiplexes, for a total of 359 SNP (eight 40-plex and one 39-plex) were designed for identyfing the best markers. Genotyping was performed using iPLEX Gold chemistry following Ellis and Ong (2017) on a MassARRAY System CPM384 (Agena Biosciences) at the PGTB platform (doi:10.15454/1.5572396583599417E12). Data analysis was achieved using MassARRAY Typer Analyzer 4.0.4 (Agena Biosciences). After genotyping, we excluded all markers for which there was evidence that the candidate SNP identified during the discovery step was not recovered, for example when the SNP exhibited fixation across the four species at the same allele. We also discarded loci with weak (magnitude <5) or ambiguous signal (i.e. displaying more clusters than expected or unclear cluster delineation) and loci with more than 20% missing data. Following this selection process, 61 SNPs (in two multiplexes) were selected on the basis of their diagnostic value and their compatibility in one multiplex kit for subsequent genotyping on all the samples.

2.2.4. Diagnosticity of candidate SNPs

Standard genetic statistics (allele frequencies, diversity statistics, differentiation and fixation indices) were estimated using GENEPOP (Raymond and Rousset, 1995) and ADEGENET software (Jombart, 2008).

We defined a metric of species diagnostic accuracy, which we coined « diagnosticity » index (*D*) to screen SNP alleles for their ability to be close to full diagnosticity.

Full diagnosticity requires two properties: fixation of the diagnostic allele in the target species and fixation of the alternate allele in the remaining species. These two properties are included in the metric *D*. Considering a set of *n* species, diagnosticity of an allele for species *x* (D_x) regarding the remaining *(n-1)* species could be expressed as:

$$D_x = p_x - \frac{1}{n-1}\sum_{j=1}^{n-1} p_j$$

Where p_x is the frequency of the candidate diagnostic allele in the target species *x*, and p_j the frequency of the same allele in the alternate species *j*. D_x amounts to the difference of allelic frequencies between species *x* and the remaining (*n-1*) species. D_x is equivalent to the mean Gregorius genetic distance between species *x* and the three other species for a diallelic locus (Gregorius, 1984).

D_x has two components, which account for the two properties of diagnosticity

- p_x: the higher p_x, the closer the near-diagnostic allele to fixation in the target species
- $\frac{1}{n-1}\sum_{j=1}^{n-1} p_j$: the lower the mean value of p_j, the closer the alternate allele to fixation in the remaining (*n-1*) species.

D_x is more appropriate for practical diagnostic assessments than the traditional differentiation metric F_{ST} when more than two species are involved (see (Gregorius and Roberts, 1986) for a comparison of *D* and F_{ST}). To illustrate the discrepancy between *D* and F_{ST} regarding diagnosticity, consider the case of four species with frequency profiles (p_1=1, p_2=1, p_3=0, p_4=0). Addressing diagnosticity for species 1, F_{ST} would yield 1, while D_1 would yield 0.67. D_1 accounts for the the lack of frequeny differences between species 1 and 2, while F_{ST} does not.

By extension of the definition of a diagnostic allele, a near-diagnostic SNP is a SNP bearing near-diagnostic alleles, and diagnosticity of a species (or a population of that species) refers to the mean value of all near-diagnostic SNPs assessed for that species or population. Diagnosticity of candidate SNPs are estimated in the training and validation populations.

2.2.5. Multilocus species clustering.

To validate the selected near-diagnostic SNP for a multilocus species assignment procedure, we implemented an empirical clustering approach using Principal Component Analysis, free of any underlying evolutionary assumptions (ADEGENET, Jombart, 2008)). This method allows to check for the ability of the near-diagnostic SNPs to visually discriminate the 4 species.

3. Results

3.1. Discovery of near-diagnostic SNPs

All together we recovered 61 candidate near-diagnostic alleles, 49 originating from the pool-seq study, and 12 from the seq-cap analysis (Table 6 in Appendix). The candidate SNPs are distributed over all chromosomes (except chromosome 4) and their number ranges from 1 (chromosome 3, 9 and 12) to 17 (chromosome 2, Figure 2). In a few cases near-diagnostic markers of a given species clustered in pairs in a few spots (mainly for *Q. robur* on chromosome 2, 5, 6). In such cases one marker of the pair was discarded during the validation step. Near-diagnostic markers are distributed over 6 chromosomes for *Q. petraea*, *Q. pubescens* and *Q. pyrenaica*, and over 8 chromosomes for *Q. robur*. As indicated by their location on the chromosomes, the minimum physical distance of near-diagnostic SNPs located on the same chromosomes was 17 Kb (Table 6 in Appendix). All except two SNPs are located on scaffolds that are anchored on the pseudo-chromosome assembly of the oak genome as shown in Figure 2.

3.2. Diagnosticity of candidate SNPs in the training set

The 61 candidate near-diagnostic SNPs exhibited allele frequency profiles close to the requisite properties of a diagnostic SNP but did not fulfill entirely criteria of full diagnosticity (Figure 2, Figure 6 in Appendix). *D* values indeed varied between 0.283 and 0.963. Most of the near-diagnostic SNPs (92%, 56/61) exhibit *D* scores greater than 0.50 (mean value 0.758). Among the 61 SNPs, 16 are candidate diagnostic of *Q. petraea*, 11 of *Q. pubescens*, 12 of *Q. pyrenaica* and 22 of *Q. robur*.

Diagnosticity scores were higher in the pool-seq uncovered set (D=0.771) than in the seq-cap uncovered set (D=0.704).

Concerning the near-diagnostic SNPs identified with the pool-seq data, diagnosticity was highest for *Q. pyrenaica* (0.897) and *Q. robur* (0.780) and lower in *Q. petraea* (0.736) and *Q. pubescens* (0.657). Deviations to full diagnosticity in the two latter species are associated with different patterns (Figure 6 in Appendix).:

- Lower diagnosticity in *Q. petraea* was mostly related to the sharing of the diagnostic allele with the other species, especially with *Q. pubescens*.
- Lower diagnosticity for *Q. pubescens* was mainly due to three SNPs (Sc0000170_630013, Sc0000192_329301 and Sc0000482_334917) that showed substantial deviation from fixation within *Q. pubescens* (frequency being respectively 0.468, 0.587, 0.283) while the alternate alleles were fixed in the three other species.

Concerning the seq-cap uncovered SNPs, we selected 12 SNPs that exhibited the highest species differentiation in the Petite Charnie population. As expected, all 12 SNPs showed

strong frequency differences between *Q. petraea* and *Q. robur* in our training panel. Eight out of the 12 SNPs exhibited allele frequency differences among the four species consistent with diagnosticity requirements for four species, with the near-diagnostic marker being almost fixed in the reference diagnostic species and present at very low frequencies in all the three remaining species (Figure 6 in Appendix). The four remaining candidate SNPs exhibited near-diagnostic alleles being almost fixed, not only in one but in two species :

- Sc0000040_1694351 in *Q. petraea* and *Q. pubescens*
- Sc0000481_366275 in *Q. robur* and *Q. pyrenaica*
- Sc0000546_456229 in *Q. robur* and *Q. pyrenaica*
- Sc0000598_295142 in *Q. robur* and *Q. pyrenaica*

3.3. Validation of the near-diagnostic SNPs

3.3.1. Screening of near-diagnostic SNPs

The validation step aimed at verifying the diagnosticity of the candidate SNPs on a larger geographic scale, while at the same time optimizing the assay by selecting the best SNPs according to various genetic and technical criteria. We thus attempted to optimize the MassARRAY® genotyping assays by reducing the number of near-diagnostic SNPs and combine them in one final assay, without limiting the species assignment purpose and reducing its diagnosticity. Indeed given the frequency profiles of near-diagnostic alleles we observed in the training set (Figure 6 in Appendix), the required number of near-diagnostic SNPs for species assignment can be limited to a handful of markers (Reutimann *et al*, 2020). We aimed at selecting about 10 near-diagnostic SNPs per species for the final design of the operational assay. The following criteria were applied (Table 6 in Appendix):

- Repeatability and clarity of the cluster delimitation on the scatter plots
- Diagnosticity of SNPs
- A nearly equal numbers of near-diagnostic SNPs per species

Combining the remaining SNPs within one or two multiplex sets, resulted in amplification incompatibilities among SNPs which lead us to discard additional SNPs. Finally a total of ten near-diagnostic SNPs were selected for *Q. petraea*, seven for *Q. pubescens*, nine for *Q. pyrenaica* and twelve for *Q. robur* (Table 6 in Appendix).

3.3.2. Allele frequency profiles of near-diagnostic SNPs in the validation populations

Overall, the average diagnosticity of the 38 near-diagnostic SNPs was slightly higher in the validation than in the training populations, with the exception of *Q. pyrenaica* (Figure 3,

Figure 6 in Appendix): 0.784 (validation) *vs* 0.715 (training) in *Q. petraea*, 0.747 *vs* 0.690 in *Q. pubescens*, 0.876 *vs* 0.897 in *Q. pyrenaica*, 0.841 *vs* 0.758 in *Q. robur*. The lower diagnosticity of *Q. pyrenaica* in the validation set (*vs* the training set) was due to SNP Sc0000307_852597, which exhibited contrasting values between the training (0.753) and validation set (0.546) (Table 7 in Appendix).

However, the validation populations provided the opportunity to explore the stability of the allele frequency profiles across geographic regions, and thus addressed the maintenance of diagnosticity of individual SNPs across the distribution of the four species. Most near-diagnostic SNPs exhibited larger genetic differentiation between populations within a given species than usually found (Scotti-Saintagne *et al*, 2004) in oak species (Table 1, 2, 3). Mean intraspecific F_{ST} values of near-diagnostic SNPs amounted to 0.104, 0.192, 0.042 and 0.104 for *Q. petraea*, *Q. pubescens*, *Q. pyrenaica*, and *Q. robur*, respectively. Furthermore, F_{ST} values within a species exhibited large variation among SNPs. For example, F_{ST} values of near-diagnostic SNPs of *Q. petraea* between *Q. petraea* populations varied between 0.012 and 0.252. *Quercus pyrenaica* is an exception to these general rules (0.042), as the mean F_{ST} is much lower than for the 3 other species and the range of variation reduced (-0.022 to 0.142, data not shown).

3.3.3. Allele frequency profiles of diagnostic SNPs in *Q. petraea* populations.

We examined the geographic distribution of near-diagnostic alleles between populations within a given species. To illustrate the results we selected populations that are representative of the variation observed among all populations. We first selected a few widely distributed populations that exhibited allele frequencies at all SNPs close to the expected diagnosticity ("EP populations": Tronçais, Lappwald and Bézange), and added all the populations that deviate from the EP frequency profiles, which we called diverging populations ("DP populations"). The DP populations included three extreme southern populations (Pomieri and Aspromonte in Italy, Montejo in Spain) and one population from the northern distribution edge (Killarney). All the remaining *Q. petraea* populations exhibited frequency profiles similar to the selected EP populations, and are not shown in Table 1 and in Figure 4. While the EP populations exhibited almost full fixation in all near-diagnostic SNPs, the DP populations showed substantial polymorphism (i.e. lower diagnosticity) at a few SNPs in Pomieri and Aspromonte (Sc0000043_1651618, Sc0000135_261350, Sc0000274_909817), and moderate polymorphism distributed among more SNPs in Killarney and Montejo.

Additionally, we examined the occurrences of near-diagnostic alleles of the other three species in *Q. petraea* populations (Figure 4). Interestingly the DP *Q. petraea* populations were also diverging in respect to the frequency of near-diagnostic alleles of *Q. pubescens* (Pomieri and Aspromonte), or *Q. robur* (Killarney and Montejo). The EP populations exhibited lower frequencies of near-diagnostic alleles of the other three species (Figure 4).

3.3.4. Allele frequency profiles of near-diagnostic SNPs in *Q. pubescens*, *Q. robur* and *Q. pyrenaica* populations.

To illustrate the intraspecific differentiation of near-diagnostic SNPs in the other three species, we followed the same procedure as for *Q. petraea*. We selected for each species two sets of populations: a subset of populations exemplifying the pattern close to full fixation of near-diagnostic loci at all SNPs (EP populations), and the set of diverging populations (DP populations) that exhibited deviations to this trend.

In the case of *Q. pubescens,* the DP populations (Switzerland and Ventoux) were located at the central northern edge of distribution. These deviations were not evenly distributed across the 7 near-diagnostic SNPs of *Q. pubescens*, but restricted to the same loci in the two populations (Table 2). The two populations Switzerland and Ventoux exhibited also higher frequencies of *Q. petraea* near-diagnostic alleles, in comparison to the two EP populations (Figure 7 in Appendix).

In the case of *Q. robur*, there were also two DP populations located at the south western (Pedro) and north western margin of the distribution (Roudsea) (Table 3). These two populations comprised also larger frequencies of near-diagnostic alleles of other white oak species (*Q. pubescens* and *Q. pyrenaica* in the case of Pedro; *Q. petraea* in the case of Roudsea) (Figure 8 in Appendix). Finally, in *Q. pyrenaica*, all populations behave as EP populations (data not shown), eg all *Q. pyrenaica* populations exhibited frequency profiles similar to those shown for *Q. pyrenaica* in Figure 3 and Table 7 in Appendix.

3.4. Multilocus structure of near-diagnostic SNPs

We used a principal component analysis (PCA) in the validation populations to assess and illustrate species differentiation (Figure 5). We added 13 samples of known first generation hybrid origin to the species samples. Ten samples resulted from controlled interspecific crosses, and three came from parentage analysis conducted in a mixed *Q. petraea*-*Q. robur* stand (Truffaut *et al*, 2017). A combination of the three first components allowed to visually differentiate the four different species. While principal component 1 differentiated mainly *Q.*

petraea and *Q. robur* (Figure 5a), component 3 distinguished *Q. pyrenaica* from the three other species (Figure 5b), and the biplot of component 2 and 3 provided the best visual separation between *Q. pubescens* and *Q. petraea* (Figure 5c).

These multilocus representations showed that there is a small number of samples located at intermediate positions, especially between *Q. petraea* and *Q. robur* (Figure 5a), and between *Q. petraea* and *Q. pubescens* (Figure 5c). These regions of the PCA are also occupied by known interspecific hybrids, suggesting that the species samples, although identified as pure species in the field, represent either hybrids or introgressed forms. These intermediate positions are also preferentially occupied by trees belonging to diverging populations, as shown by the targeted PCA analysis on the two pairs of species sharing intermediate samples: *Q. petraea* and *Q. pubescens* (Figure 9 in Appendix), *Q. petraea* and *Q. robur* (Figure 10 in Appendix).

4. Discussion

We explored large scale existing genomic resources in four European white oaks of the subsection Roburoid (*Q. petraea, Q. pubescens, Q. pyrenaica, Q. robur*) to sceen their genomes for near-diagnostic SNPs that could be used for molecular fingerprinting (species and hybrid identification) in forest research and operational forestry, as wood or seed traceability in the wood chain and in forest nurseries. Despite the widely reported low interspecific genetic differentiation among European white oak species, we were able to identify a subset of SNPs that exhibited near-diagnostic features across their species' distribution ranges. Moreover, mutlivariate analysis showed that these markers can be used for reliable hybrid detection and accurate quantification of admixture levels. However, diagnosticity varied substantially among species, among populations within species, and among SNPs. In the following, we discuss these variations in relation to the known evolutionary history and genetic interactions among and within the four species.

4.1 Variation of diagnosticity among species

Diagnosticity was highest in *Q. pyrenaica* (0.876) and lowest in *Q. pubescens* (0.747) with *Q. robur* and *Q. petraea* showing intermediate values. Near-diagnostic SNPs are likely located in genomic regions that exhibit larger divergence and/or regions prevented from interspecific gene flow. The range of diagnosticity among the four species, may therefore reflect the variation of divergence time and/or the variation of the intensity of gene flow during the ongoing interglacial period.

464 It is striking to notice that higher and lower diagnosticity was observed for species that
465 showed the older (*Q. pyrenaica*, *Q. robur*) and more recent (*Q. petraea*, *Q. pubescens*)
466 divergence, respectively (Leroy *et al*, 2017). Fixation of near-diagnostic SNPs in species with
467 large population sizes as in oaks requires long time periods. Consequently, lower
468 diagnosticity is likely associated with species that diverged more recently. This is illustrated
469 by *Q. pubescens*, which shows lower diagnosticity due to the higher sharing of near-
470 diagnostic alleles with *Q. petraea* than with the other two species (Figure 3 and Figure 6 in
471 Appendix). Diagnosticty may in addition be dependent on the variation of population size
472 (*Ne*) among species and along divergence, for which we lack any estimation today. Our
473 results may therefore be revisited in the light of future evidence of *Ne* differences. Regarding
474 gene flow, we showed earlier that the four species came into contact only recently, during the
475 late last glacial maximum, after being isolated for most of their earlier history (Leroy *et al*,
476 2020b; Leroy *et al*, 2017), resulting in gene flow among species. While interfertility among
477 the four species has been shown experimentally by controlled crosses (Lepais *et al*, 2013),
478 hybridization *in natura* has also been observed among the four species in rare mixed forests
479 where all four species co-occur (Lepais and Gerber, 2011; Lepais *et al*, 2009). Interspecific
480 matings of *Q. pyrenaica* in controlled crosses with the remaining three species were quite
481 successful, however occurrences of natural hybridization were less frequent due to the very
482 late flowering of *Q. pyrenaica* in comparison to the three other species (Lepais and Gerber,
483 2011; Lepais *et al*, 2013). Furthermore *Q. pyrenaica* is mainly distributed in south western
484 Europe, where the other three species are only present in scattered forests, leading, for
485 example, to reported but rare hybridization with *Q. petraea* (Valbuena-Carabana *et al*, 2005)
486 and *Q. robur* (Moracho *et al*, 2016). Altogether, phenological prezygotic barriers and limited
487 overlapping distributions with the other three species may have contributed to reduced genetic
488 exchanges between *Q. pyrenaica* and the other three species, and thus account for the high
489 diagnosticity of the SNPs in of *Q. pyrenaica*. In contrast to *Q. pyrenaica*, no reproductive
490 barriers were observed in *Q. pubescens* when crosses were made with *Q. petraea* as female
491 parent, as interspecific crosses were as successful as intraspecific crosses (Lepais *et al*, 2013).
492 Reduced barriers between these two species were corroborated by frequent admixture
493 detected in genetic surveys conducted in mixed stands of *Q. pubescens* and *Q. petraea*
494 (Alberto *et al*, 2010; Neophytou, 2014; Reutimann *et al*, 2023). As a result, near-diagnostic
495 SNPs of *Q. pubescens* and *Q. petraea* were more frequently shared between the two species
496 (Figure 3 and Figure 6 in Appendix) thus contributing to reduced diagnosticity. Finally
497 interspecific gene exchanges involving *Q. robur* were mainly investigated with regard to *Q.*

petraea. Uneven gene flow has been repeatedly observed in mixed stands with limited pollination from *Q. robur* to *Q. petraea* (Bacilieri *et al*, 1996; Lagache *et al*, 2013; Lepais *et al*, 2013), with a few exceptions in stands of unbalanced mixtures (Gerber *et al*, 2014). Uneven and unidirectional gene exchanges between these two species may have resulted in higher diagnosticity of *Q. robur* in comparison to *Q. petraea*.

4.2 Variation of diagnosticity among populations

There are striking differences of species diagnosticity of the markers among populations within species (Table 1, 2 and 3). In populations of *Q. petraea*, *Q. pubescens* and *Q. robur* located in the central part of their distributions, high levels of diagnosticity (mean values of SNP diagnosticity of the population) could be observed, while in populations located at the margins of the distributions, southern as well as northern, lower diagnosticity was found. We further showed that populations located at the edges of distribution are characterized by higher frequencies of near-diagnostic alleles of the other three congeneric species, suggesting extensive genetic exchanges (Figure 4, Figure 7 and 8 in Appendix). More frequent interspecific gene flow at the northern edge of distribution has been shown earlier in the case *Q. petraea* and *Q. robur* (Beatty *et al*, 2016; Jensen *et al*, 2009; Gerber *et al*, 2014), and has been interpreted as a driver of the succession dynamics at the northern colonization front of the two species (Kremer and Hipp, 2020; Petit *et al*, 2003). In our study, the sessile oak population Killarney (Figure 4) and the pedunculate oak population Roudsea (Figure 8 in Appendix) are typical examples illustrating interspecific gene flow between the two species. Similar observations of more frequent hybridization were made in the case of *Q. petraea* and *Q. pubescens* at the northern edge of distribution of *Q. pubescens* (Neophytou *et al*, 2015; Reutimann *et al*, 2020), which may have as well contributed to the expansion of *Q. pubescens*.

In populations located at the southern edge of distribution (Pomieri, Aspromonte, and Montejo for *Q. petraea*, Figure 1 and Figure 4), the lower diagnosticity may have resulted from more ancient genetic exchanges with *Q. pubescens* and *Q. robur*. Indeed the two italian populations (Pomieri and Aspromonte) in Sicilia and Calabria consist today in almost pure stands, where *Q. pubescens* is extremely rare, if not absent (Bagnato *et al*, 2012; Modica, 2001), while our results indicated introgression of *Q. pubescens* into *Q. petraea* (Figure 4). Similarly the sessile oak population Montejo, in central Spain, is introgressed by *Q. robur* (Figure 4), where the latter species is absent today and where contemporary hybridization has rather been detected with *Q. pyrenaica* (Valbuena-Carabana *et al*, 2005). Finally, a similar

scenario holds for the pedunculate oak population Pedro, which is located at the extreme southern edge of distribution of *Q. robur* (Figure 1; Table3, Figure 8 in Appendix). Hybridization has been observed with *Q. pyrenaica* which is today the most frequent species in the area (Moracho *et al*, 2016) and is confirmed by our results revealing the presence of *Q. pyrenaica* near-diagnostic alleles in the *Q. robur* population (Figure 8 in Appendix). However, introgression by *Q. pubescens* is even more pronounced in our data despite the today's absence of *Q. pubescens* in Extremadura (Figure 8 in Appendix). To sum up, when comparing our results with previous investigations on interspecific gene flow, recent and/or ancient gene exchanges have faded diagnosticity in the so-called diverging populations, which are located at the northern or southern margins of the distribution.

4.3 Variation of diagnosticity among SNPs

Frequency profiles of near-diagnostic alleles differed markedly across SNP in diverging populations. There were cases where lack of diagnosticity affected mainly the same limited number of loci in a given species (Aspromonte and Pomieri in *Q. petraea*, Table 1; Pedro in *Q. robur*, Table 3; and to a smaller extend Switzerland and Ventoux in *Q. pubescens*, Table 2). In the remaining diverging populations (Killarney for *Q. petraea*, Table 1; and Roudsea for *Q. robur*, Table 3), reduced diagnosticity is more evenly distributed across more if not all loci. Contrasting diagnosticity distribution across loci may likely correlate to the timing of hybridization and introgression among the congeneric species. Recent gene exchanges, as first generation hybridization and subsequent backcrosses will indistinctly impact all loci during the early phase of secondary contact among species, and result in reduced diagnosticity of alleles in sympatric species. Such a scenario may hold for the two northern *Q. petraea* (Killarney) and *Q. robur* (Roudsea) populations. Continuous gene exchanges over multiple generations may ultimately result in heterogeneous genomic landscapes, shaped by variable permeability to gene flow along the chromosomes due to the presence of prezygotic or postzygotic barriers and the heterogeneous recombination landscape. This scenario leads ultimately to the maintenance of near-diagnostic loci in genomic regions impermeable to gene flow, while the remaining part of the genome will become poorly differentiated. While this scenario was supported by ABC simulations (Leroy *et al*, 2020b; Leroy *et al*, 2017), our results further suggest that the genomic distribution of near-diagnostic loci is environment-dependent. It is striking that a very limited number of near-diagnostic alleles discovered in western populations of *Q. petraea* show poor diagnosticity in the southern populations Pomieri and Aspromonte (Table 1). Our results further indicated that this low diagnosticity

may be due to more interspecific gene flow with *Q. pubescens*, which suggest preferentially introgression in specific genomic regions–whether adaptive or not- resulting ultimately in heterogeneous genomic distribution of near-diagnostic SNPs especially in marginal range parts. In a recent paper we showed that introgressed regions between *Q. robur* and *Q. petraea* may be more frequent at higher altitudes (Leroy *et al*, 2020a) while in another case study in two Asian oak species the authors found that the genomic landscape of introgression changed in different ecological settings (Fu *et al*, 2022). A similar picture holds for the diverging southern *Q. robur* population Pedro, where diagnosticity is substantially reduced at a few near-diagnostic SNPs in comparison to other *Q. robur* populations (Table 3), most likely due to introgression by *Q. pubescens* and *Q. pyrenaica* (Figure 8 in Appendix). Anecdotally the diverging status of Aspromonte, Pomieri and Pedro echoes with the taxonomic subspecies status that has been assigned to the Sicilian and Calabrian *Q. petraea* populations (*Q. petraea* ssp *austrothyrrenica*, Bagnato *et al*, 2012; Lupini *et al*, 2019; Merlino *et al*, 2014) and to the extreme southern spanish *Q. robur* populations (*Q. robur* ssp *estremadurensis*, Vazquez-Pardo *et al*, 2009).

Conclusions and outlook

Here we showed that near-diagnostic marker development for species identification is feasible despite few species barriers, extensive secondary contact, and, consequently, frequent hybridization and introgression. Recently we demonstrated that the set of near-diagnostic markers resolved species assignment on fossil and archeological oak wood remains, where anatomical features do not allow to discriminate the four deciduous species (Wagner *et al*, 2023). With the steadily ongoing availability of whole genomes in non model species including oaks (Lazic *et al*, 2021), the search of near-diagnostic markers could be extended to the whole Roburoid subsection facilitating white oak species assignment throughout Europe, beyond the subset of four species that we considered here.The near-diagnostic SNPs for the four white oak species could not only be used in forest research and management for reliable and affordable species assignment, but also to identify admixed individuals and accurately quantify admixture levels in natural populations (Reutimann *et al*, 2020). Because the presented alleles are often almost fixed for the target species, these SNPs also allow the identification of hybrid state (F1, F2, backcrosses, later generation hybrids, etc.) with methods like NEWHYBRIDS (Anderson 2008), and altogether help to understand the importance of hydribization and introgression in evolutionary processes. Together with prospect of emergence of field-based genotpying techniques (Urban *et al*, 2021), such near-diagnostic

markers would even allow fast fingerprinting *in-situ* to make decision for forest managers and scientists.

References

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

Anderson EC (2008) Bayesian inference of species hybrids using multilocus dominant genetic markers. *Phil. Trans. R. Soc. B* **363**: 2841–2850

Bacilieri R, Ducousso A, Kremer A (1995). Genetic, morphological, ecological, and phenological differentiation between *Quercus petraea* (Matt) Liebl and *Quercus robur* L in a mixed stand of Nortwest of France. *Silvae Genetica* **44**(1)**:** 1-10.

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

Bagnato S, Merlino A, Mercurio R, Solano F, Scarfo F, Spampinato G (2012). Le basi conoscitive per il restauro forestale: il caso di Bosco Pomieri (Parco Regionale delle Madonie, Sicilia). *Forest@* **9:** 8-19.

Beatty GE, Montgomery WI, Spaans F, Tosh DG, Provan J (2016) Pure species in a continuum of genetic and morphological variation: sympatric oaks at the edge of their range. *Annals of Botany* **117**: 541-549

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120

Bruschi P, Vendramin GG, Bussotti F, Grossoni P (2000). Morphological and molecular differentiation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens* Willd. (Fagaceae) in Northern and Central Italy. *Annals of Botany* **85**(3)**:** 325-333.

Camus A (1938). *Les chênes. Monographie du genre Quercus. Tome II. Genre Quercus. Sous-genre Euquercus (Section Lepidobalanus et Macrobalanus).* Editions Paul Lechevalier: Paris.

Coq-Etchegaray D, Bernillon S, Le-Provost G, Kremer A, Ducousso A, Lalanne C, Bonne F, Moing A, Plomion C, Brachi B (2023). Extensive variation of leaf specialized metabolite production in sessile oak (*Quercus petraea*) populations is to a large extent genetically

640 determined but not locally adaptive. *Preprint at bioRxiv*
641 https://doi.org/10.1101/2023.04.07.536008
642
643 Degen B, Blanc-Jolivet C, Bakhtina S, Ianbaev R, Yanbaev Y, Mader M, Nurnberg S,
644 Schröder H (2021). Applying targeted genotyping by sequencing with a new set of nuclear
645 and plastid SNP and indel loci for *Quercus robur* and *Quercus petraea*. *Conservation*
646 *Genetics Resources* **13**: 345-347.
647
648 Denk T, Grimm GW, Manos PS, Deng M, Hipp AL (2017). An updated infrageneric
649 classification of the oaks: review of previous taxonomic schemes and synthesis of
650 evolutionary patterns. In: Gil-Pelegrin E, Peguero-Pina JJ and Sancho-Knapik D (eds) *Oaks*
651 *physiological Ecology*. *Exploring the functional diversity of the genus Quercus L.* Springer pp
652 13-38.
653
654 Dupouey JL, Badeau V (1993). Morphological variability of oaks (*Quercus robur* L, *Quercus*
655 *petraea* (Matt)Liebl, *Quercus pubescens* Willd) in Northeastern France. Preliminary results.
656 *Ann For Sci* **50**: 35s-40s
657
658 Ellis JA, Ong B (2017). The MassARRAY® System for Targeted SNP Genotyping. In: White
659 S, Cantsilieris S (eds) Genotyping. Methods in Molecular Biology, vol 1492. Humana Press,
660 New York, NY. https://doi.org/10.1007/978-1-4939-6442-0_5
661
662 Fu R, Zhu Y, Liu Y, Feng Y, Lu R, Li Y *et al* (2022). Genome-wide analyses of introgression
663 bewteen two sympatric Asian oak species. *Nature Ecology & Evolution* **6:** 924-935.
664
665 Gerber S, Chadoeuf J, Gugerli F, Lascoux M, Buiteveld J, Cottrell J *et al* (2014). High Rates
666 of Gene Flow by Pollen and Seed in Oak Populations across Europe. *Plos One* **9**(1).
667
668 Grandjean G, Sigaud P (1987). Contribution à la taxonomie et à l'écologie des chênes du
669 Berry. *Ann For Sci* **44** : 35-66
670
671 Gregorius HR (1984). A unique genetic distance. *Biometric J* **26:** 1-14.
672
673 Gregorius HR, Roberts JH (1986). Mesuring genetic differentiation in subpopulations.
674 *Theoretical and Applied Genetics* **71:** 826-834.
675
676 Guichoux E, Lagache L, Wagner S, Leger P, Petit RJ (2011). Two highly validated
677 multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks
678 (Quercus spp.). *Molecular Ecology Resources* **11**(3)**:** 578-585.
679
680 Hipp AL, Manos PS, Hahn M, Avishai M, Bodénès C, Cavender-Bares J *et al* (2020). The
681 genomic landscape of the global oak phylogeny. *New Phytologist* **226:** 1198-1212.
682
683 Jensen J, Larsen A, Nielsen LR, Cottrell J (2009) Hybridization between *Quercus robur* and
684 *Q. petraea* in a mixed oak stand in Denmark. *Annals of Forest Science* **66**: 706
685
686 Jombart T (2008). Adegenet: a R package for the multivariate analysis of genetic markers.
687 *Bioinformatics* **24:** 1403-1405.
688

Jurksiene G, Baliuckas V (2014). Leaf morphological variation of sessile oak (*Quercus petraea* (Matt.)Liebl.) and pedunculate oak (*Quercus robur* L.) in Lithuania.. In: Treija S and Skujeniece S (eds) *Research for Rural Development 2014, Vol 2*, pp 63-69.

Kelleher CT, Hodkinson TR, Douglas GC, Kelly DL (2005). Species distinction in Irish populations of *Quercus petraea* and *Q.robur*: Morphological versus molecular analyses. *Annals of Botany* **96**(7)**:** 1237-1246.

Koffler R, Pandey RV, Schlotterer C. (2011). POPOOLATION2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**: 3435–3436

Kremer A, Dupouey JL, Deans JD, Cottrell J, Csaikl U, Finkeldey R *et al* (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.

Kremer A, Hipp AL (2020). Oaks: an evolutionary success story. *New Phytologist* **226:** 987-2011.

Lagache L, Klein EK, Guichoux E, Petit RJ (2013). Fine-scale environmental control of hybridization in oaks. *Molecular Ecology* **22**(2)**:** 423-436.

Lazic D, Hipp AL, Carlson JE, Gailing O (2021). Use of genomic resources to assess adaptive divergence and introgression in oaks. *Forests* **12**: 690

Le Hardy de Beaulieu A, Lamant T (2006). *Guide illustré des chênes. Tome 1*, Vol 1. Editions du 8ième: Paris.

Lepais O, Gerber S (2011). Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution* **65**(1)**:** 156-170.

Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, Kremer A *et al* (2009). Species relative abundance and direction of introgression in oaks. *Molecular Ecology* **18**(10)**:** 2228-2242.

Lepais O, Roussel G, Hubert F, Kremer A, Gerber S (2013). Strength and variability of postmating reproductive isolating barriers between four European white oak species. *Tree Genetics & Genomes* **9**(3)**:** 841-853.

Leroy T, Louvet JM, Lalanne C, Le Provost G, Labadie K, Aury JM *et al* (2020a). Adaptive introgression as a driver of local adaptation to climate in European white oaks *New Phytologist* **226:** 1171-1182.

Leroy T, Rougemont Q, Dupouey JL, Bodénès C, Lalanne C, Belser C *et al* (2020b). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytologist* **226:** 1183-1197.

Leroy T, Roux C, Villate L, Bodenes C, Romiguier J, Paiva JAP *et al* (2017). Extensive recent secondary contacts between four European white oak species. *New Phytologist* **214**(2)**:** 865-878.

Lesur I, Alexandre H, Boury C, Chancerel E, Plomion C, Kremer A (2018). Development of Target sequence capture and estimation of genomic relatedness in a Mixed Oak Stand. *Frontiers in Plant Science* **9**: 996

Li H (2011). Improving SNP discovery by base alignment quality. *Bioinformatics* **27** : 1157-1158

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint at arxiv* https://arxiv.org/abs/1303.3997 (2013).

Lopez de Heredia U, Valbuena-Carabana M, Cordoba M, Gil L (2009). Variation components in leaf morphology of recruits of two hybridising oaks *Q. petraea* (Matt.) Liebl. and *Q. pyrenaica* Willd. at small spatial scale. *European Journal of Forest Research* **128**(6)**:** 543-554.

Lupini A, Aci M, Mauceri A, Luzzi G, Bagnato S, Menguzzato G *et al* (2019). Genetic diversity in old populations of sessile oak from Calabria assessed by nuclear and chloroplast SSR. *Journal of Mountain Science* **16:** 1111-1120.

Macejovsky V, Schmidtova J, Hrivnak M, Krajmerova D, Sarvasova I, Gomory D (2020). Interspecific differentiation and gene exchange among the Slovak Quercus sect. Quercus populations. *Dendrobiology* **83:** 20-29.

Merlino A, Baliva M, Di Filippo A, Piovesan G, Solano F. (2014). Analisi strutturali e dendroecologiche su popolamenti di *Quercus Petraea* subsp. *austrothyrrenica* Brullo, Guarino e Siracusa nel parco regionalle delle Madonie (Sicilia). *Second International Congress of Silviculture*, pp 183-189.

Modica G (2001). La rovere (*Quercus petraea* (Matt.) Liebl.) in Aspromonte. *Monti e Boschi* **3/4:** 13-18.

Moracho E, Moreno G, Jordano P, Hampe A (2016). Unusually limited pollen dispersal and connectivity of Pedunculate oak (*Quercus robur*) refugial populations at the species southern range margin. *Molecular Ecology* **14:** 3319-3331.

Neophytou C (2014). Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. *Tree Genetics & Genomes* **10**(2)**:** 273-285.

Neophytou C, Gartner SM, Vargas-Gaete R, Michiels HG (2015). Genetic variation of Central European oaks: shaped by evolutionary factors and human intervention? *Tree Genetics & Genomes* **11**(4).

Petit RJ, Bodenes C, Ducousso A, Roussel G, Kremer A (2003). Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**(1)**:** 151-164.

Plomion C, Aury JM, Amselem J, Leroy T, Murat F, Duplessis S *et al* (2018). Oak genome reveals facets of long lifespan. *Nature Plants* **4**(7)**:** 440-452.

Raymond M, Rousset F (1995). GENEPOP(version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86:** 248-249.

Rellstab C, Buhler A, Graf R, Folly C, Gugerli F (2016). Using joint multivariate analyses of leaf morphology and molecular-genetic markers for taxon identification in three hybridizing European white oak species (*Quercus* spp.). *Annals of Forest Science* **73**(3)**:** 669-679.

Reutimann O, Dauphin B, Baltensweiler A, Gugerli F, Kremer A, Rellstab C (2023). Abiotic factors predict taxonomic composition and genetic admixture in populations of hybridizing white oak species (*Quercus* sect. *Quercus*) on a regional scale. *Tree Genetics & Genomes* **19:** 22.

Reutimann O, Gugerli F, Rellstab C (2020). A species-discrimnatory single-nucleotide polymorphism set reveals maintenance of species integrity in hybridizing European white oaks (Quercus spp.) despite high levels of admixture. *Annals of Botany* **125:** 663-676.

Saleh D, Chen J, Leple JC, Leroy T, Truffaut L, Dencausse B *et al* (2022). Genome-wide evolutionary response of European oaks during the Anthropocene. *Evolution Letters* **6**(1)**:** 4-20.

Schroeder H, Kersten B (2023). A small set of nuclear markers for reliable differentiation of the two closely related oak species *Quercus robur* and *Q.petraea*. *Plants* **12** (3), 566

Scotti-Saintagne C, Mariette S, Porth I, Goicoechea PG, Barreneche T, Bodenes K *et al* (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.

Truffaut L, Chancerel E, Ducousso A, Dupouey JL, Badeau V, Ehrenmann F *et al* (2017). Fine-scale species distribution changes in a mixed oak stand over two successive generations. *New Phytologist* **215**(1)**:** 126-139.

Urban L, Holzer A, Jotautas Baronas J, Hall MB, Braeuninger-Weimer P, Scherm MJ, Kunz DJ, Perera SN, Martin-Herranz DE, Tipper ET, Salter SJ, Stamnitz MR (2021) Freshwater monitoring by nanopore sequencing. *eLife* **10** : e61504

Valbuena-Carabana M, Gonzalez-Martinez S, Sork V, XCollada C, Soto A, PGoicoechea P *et al* (2005). Gene flow and hybridisation in a mixed oak forest (*Quercus pyrenaica* Willd. and *Quercus petraea* (Matt.) Liebl.) in central Spain. *Heredity* **95:** 457-465.

Vazquez-Pardo F, Rincon-Hercules S, Gutierrez-Esteban M, Garcia-Alonso M, Marquez-Garcia F, Ramos -Maqueda S *et al*. (2009). *Congreso Forestal Espanol. Montes y sociedad: Saber que hacer*. Leon SECFJdCy (ed.). Socieda Espanola de Ciencas Forestales: Avila, pp 3-13.

Viscosi V, Lepais O, Gerber S, Fortini P (2009). Leaf morphological analyses in four European oak species (Quercus) and their hybrids: A comparison of traditional and geometric morphometric methods. *Plant Biosystems* **143**(3)**:** 564-574.

839 Wagner S, Seguin-Orlando A, Leplé JC, Leroy T, Lalanne C, Aury JM, Poirier S, Wincker P,
840 Plomion C, Kremer A, Orlando L (2023). Tracking population structure and phenology
841 through time using ancient genomes from watterlogged white oak. *Molecular Ecology* (in
842 press) https://doi.org/10.1111/mec.16859
843
844 Yucedag C, Gailing O (2013). Morphological and genetic variation within and among four
845 *Quercus petraea* and *Q. robur* natural populations. *Turkish Journal of Botany* **37**(4)**:** 619-629.
846
847

848

849 Table 1 Frequencies and differentiation of near-diagnostic alleles of *Q. petraea* in *Q. petraea* populations

850

SNP ID	Tronçais	Lappwald	Bezange	**Pomieri**	**Aspromonte**	**Killarney**	**Montejo**	Intraspecific F_{ST}	p value
Sc0000254_6223	0.867	1	0.883	0.975	-	**0.625**	-	0.098	0.000
Sc0000121_355205	0.974	1	0.988	0.947	-	0.906	-	0.029	0.002
Sc0000043_1651618	0.817	0.925	0.888	**0.35**	**0.344**	**0.719**	**0.763**	0.190	0.000
Sc0000083_147504	0.95	0.9	0.898	0.875	0.979	**0.656**	**0.776**	0.049	0.000
Sc0000118_1466708	0.933	0.95	0.929	0.8	0.917	**0.688**	**0.671**	0.052	0.000
Sc0000135_261350	0.948	0.8	0.929	**0.425**	**0.573**	0.875	0.974	0.133	0.000
Sc0000145_700044	0.983	1	1	0.85	0.927	0.875	1	0.039	0.000
Sc0000203_707735	1	1	0.981	0.875	0.979	0.8	**0.622**	0.176	0.000
Sc0000274_909817	0.933	0.875	0.949	**0.325**	**0.333**	**0.688**	0.816	0.259	0.000
Sc0000481_343721	0.983	0.9	0.949	0.816	0.958	0.9	0.973	0.012	0.063
Mean Diagnosticity	0.852	0.848	0.852	0.637	0.664	0.686	0.737		

851

852 Populations in bold characters correspond to DP populations (Diverging populations, see text) and frequencies in bold characters correspond to loci
853 exhibiting deviations to expected frequencies of diagnostic alleles. Geographic locations of the populations are shown in Figure 1a.

854

Table 2. Frequencies and differentiation of near-diagnostic alleles of *Q. pubescens* in *Q. pubescens* populations

SNP ID	Auros	Pantano	**Switzerland**[a]	**Ventoux**	Intraspecific F_{ST}	P value
Sc0000314_149731	0.821	-	**0.531**	**0.521**	0.113	0.003
Sc0000047_2398879	0.946	1.000	0.694	0.868	0.127	0.000
Sc0000088_1796044	1.000	1.000	0.806	0.812	0.128	0.000
Sc0000109_800763	0.839	0.882	**0.528**	**0.692**	0.096	0.000
Sc0000111_693153	1.000	0.987	1.000	1.000	-0.004	0.682
Sc0000170_630013	0.907	0.987	**0.222**	**0.400**	0.506	0.000
Sc0000192_329301	0.880	0.986	**0.333**	**0.487**	0.376	0.000
Mean Diagnosticity	0.865	0.926	**0.540**	**0.635**		

Populations in bold characters correspond to DP populations (Diverging populations. see text) and frequencies in bold characters correspond to loci exhibiting deviations to expected frequencies of near-diagnostic alleles. Geographic origins of populations are shown in Figure 1b.

[a]Switzerland population assembles data of populations Ayent. Cordola. Remigen and Saillon described in Table 5 in Appendix.

866 Table 3 in Appendix. Frequencies of near-diagnostic alleles and differentiation of *Q. robur* in *Q. robur*
867 populations

868

SNP ID	Zivinice	Sigmunds-herberg	Charnie	Escherode	**Pedro**	**Roudsea**[a]	Intraspecific F_{ST}	P value
Sc0000013_2578823	0.711	0.925	0.875	0.750	**0.273**	**0.553**	0.140	0.000
Sc0000053_1639108	1.000	1.000	1.000	0.857	0.818	**0.500**	0.098	0.009
Sc0000099_1839376	0.868	0.868	0.889	0.972	0.818	0.711	0.040	0.000
Sc0000158_462639	0.763	0.975	0.889	0.806	0.864	**0.658**	0.071	0.000
Sc0000203_689887	0.842	0.750	1.000	0.806	0.955	**0.632**	0.095	0.000
Sc0000339_4638	1.000	1.000	0.944	0.941	**0.227**	0.853	0.359	0.000
Sc0000381_206331	1.000	1.000	1.000	1.000	**0.500**	**0.737**	0.233	0.000
Sc0000447_521057	1.000	0.950	0.944	1.000	1.000	0.816	0.042	0.001
Sc0000517_258593	0.938	0.975	0.938	0.917	1.000	0.974	-0.003	0.166
Sc0000695_225347	0.947	1.000	0.889	0.917	**0.773**	**0.763**	0.050	0.000
Sc0000796_82698	0.842	0.925	0.944	0.972	0.864	0.789	0.021	0.016
Sc0000967_33996	1.000	1.000	1.000	1.000	1.000	**0.763**	0.106	0.000
Mean Diagnosticity	0.851	0.889	0.885	0.854	0.700	0.671		

869

870 Populations in bold characters correspond to DP populations (Diverging Populations, see text) and
871 frequencies in bold characters correspond to loci exhibiting deviations to expected frequencies of
872 near-diagnostic alleles. Geographic origins of populations are indicated in Figure 1d.

873 [a]Roudsea population assembles data of populations Dalkeith and Roudsea described in Table 5 in
874 Appendix.
875

876 Table 4a in Appendix. Discovery samples of the whole genome pool-sequenced resources

877

Species	Sampling site	Latitude	Longitude	Sample size
Q. petraea	Laveyron	43.9747	0.2297	13
Q. pubescens	Branne	44.8399	-0.2049	12
	Blaignan	45.3192	-0.8559	6
				18
Q. robur	ISS Landes	44.2263	1.0112	20
Q.pyrenaica	ISS Landes	44.2701	1.0697	20

878

879 Table 4b in Appendix. Discovery samples of the sequence captured genomic resources

880

Species	Sampling site	Latitude	Longitude	Sample Size
Q. petraea	La Petite Charnie	48.086	-0.168	110
Q. robur	La Petite Charnie	48.086	-0.168	135

881

882

883

884

885 Table 5 in Appendix. Geographic origins of training and validation samples.

886

Population	Species	Country	Latitude	Longitude	Training1	Training2	Validation
Olovo	*Q. petraea*	Bosnia Herzegovina	44.152	18.548	11		51
Artouste	*Q. petraea*	France	42.890	-0.400			10
Berce	*Q. petraea*	France	47.813	0.391		8	20
Bezange	*Q. petraea*	France	48.759	6.493			20
Briouant	*Q. petraea*	France	43.306	1.048	2		
Gabas	*Q. petraea*	France	42.880	-0.420			20
Gedre	*Q. petraea*	France	42.780	0.020			20
Gresigne	*Q. petraea*	France	44.043	1.749			20
Josbaig	*Q. petraea*	France	43.220	-0.730			20
Charnie	*Q. petraea*	France	48.086	-0.168	9		9
Laveyron	*Q. petraea*	France	43.975	-0.280	2	4	20
Le Hourque	*Q. petraea*	France	42.900	-0.430			20
Longchamp	*Q. petraea*	France	47.264	5.310		2	20
Papillon	*Q. petraea*	France	42.920	-0.030			20
Péguères	*Q. petraea*	France	42.870	-0.120		4	18
Saint Sauvant	*Q. petraea*	France	46.380	0.124			20
Tronçais	*Q. petraea*	France	46.680	2.829	11	6	31
Vachères	*Q. petraea*	France	43.983	5.633		2	20
Göhrde	*Q. petraea*	Germany	53.100	10.846		6	20
Lappwald	*Q. petraea*	Germany	52.257	10.988			20
Killarney	*Q. petraea*	Ireland	52.013	-9.504			20
Aspromonte	*Q. petraea*	Italy	38.143	15.938			50
Pomieri	*Q. petraea*	Italy	37.866	14.069			20
Montejo	*Q. petraea*	Spain	41.117	-3. 500	11		51
Val de Seine	*Q. petraea*	France	48.398	3.578			10
Auros	*Q. pubescens*	France	44.492	-0.148	12	3	12
Briouant	*Q. pubescens*	France	43.306	1.048	2		
Blaignan	*Q. pubescens*	France	45.319	-0.856	2		6
Branne	*Q. pubescens*	France	44.840	-0.205			12
ISSVentoux	*Q. pubescens*	France	44.121	5.312	16	8	40
Pantano	*Q. pubescens*	Italy	40.164	16.671			40
Ayent	*Q. pubescens*	Switzerland	46.266	7.398	4	3	4
Cordola	*Q. pubescens*	Switzerland	46.195	8.863	4		4
Remigen	*Q. pubescens*	Switzerland	47.519	8.163	4	5	5

Saillon	*Q. pubescens*	Switzerland	46.171	7.167	4		7
Val de Seine	*Q. pubescens*	France	48.435	3.598			3
Briouant	*Q. pyrenaica*	France	43.306	1.048	2		
ISSLandes Mont de Marsan	*Q. pyrenaica*	France	44.235	-1.088	10		43
ISSLandes	*Q. pyrenaica*	France	44.270	-1.070	6	8	20
Hoya Del Nevazo	*Q. pyrenaica*	Spain	36.957	-3.423	7		7
La Calanchera	*Q. pyrenaica*	Spain	39.572	-4.647	6		6
Pedro	*Q. pyrenaica*	Spain	40.079	-5.739	12	11	12
Rascafria	*Q. pyrenaica*	Spain	40.911	-3.898	3		3
Sigmundsherberg	*Q. robur*	Austria	48.683	15.750		2	20
Livno	*Q. robur*	Bosnia Herzegovina	44.015	16.630	11		51
Zivinice	*Q. robur*	Bosnia Herzegovina	44.446	18.674		5	20
Briouant	*Q. robur*	France	43.306	1.048	2		
ISSLandes	*Q. robur*	France	44.226	-1.011	2		20
ISSLandes Mont de Marsan	*Q. robur*	France	44.221	-1.098			48
ValSeine	*Q. robur*	France	48.398	3.578			8
Charnie	*Q. robur*	France	48.086	-0.168	9		9
Escherode	*Q. robur*	Germany	51.333	9.400			18
Policoro (Pantano)	*Q. robur*	Italy	40.159	16.675		3	20
Pollutri (San Venanzio)	*Q. robur*	Italy	42.146	14.643		5	20
Arbalan	*Q. robur*	Spain	42.967	-2.550		6	20
Pedro	*Q. robur*	Spain	40.079	-5.739	11		11
Birmensdorf	*Q. robur*	Switzerland	47.436	8.255	3		5
Bonfol	*Q. robur*	Switzerland	47.463	7.148	3		5
Bueren	*Q. robur*	Switzerland	47.117	7.383			20
Cureglia	*Q. robur*	Switzerland	46.042	8.950	2	3	5
Rapperswill	*Q. robur*	Switzerland	47.239	8.839	3		5
Dalkeith	*Q. robur*	United Kingdom	55.917	-3.033			8
Roudsea	*Q. robur*	United Kingdom	54.232	-3.026			16

887

888

889

890

Table 6 in Appendix. Genetic and genomic features of near-diagnostic SNPs

SNP ID[f]	Discovery resources	Sample[a]	Screening[b]	Reference diagnostic species	Genotype	Diagnostic nucleotide	Expression	GeneID	Chr[c]	Position[d]	Distance[e]
Sc0000254_6223	Pool-seq	T+V		*Q. petraea*	AT	T	intergenic	NA	1	12964247	3103379
Sc0000121_355205	Pool-seq	T+V		*Q. petraea*	CT	C	intergenic	NA	5	69028945	33570950
Sc0000040_1694351	Seq-cap	T	PD	*Q. petraea*	AT	A	intergenic	NA	2	17875174	8132192
Sc0000043_1651618	Pool-seq	T+V		*Q. petraea*	AG	G	intergenic	NA	2	41927441	2874167
Sc0000055_2262067	Pool-seq	T	VA	*Q. petraea*	CT	T	intergenic	NA	5	27820895	3322468
Sc0000083_147504	Seq-cap	T+V		*Q. petraea*	AG	A	exonic	Qrob_G0609970.2	2	46841435	2526743
Sc0000090_1332487	Pool-seq	T	PQ	*Q. petraea*	TC	T	exonic	Qrob_G0088630.2	7	44619559	28315
Sc0000090_1360802	Seq-cap	T	DisC	*Q. petraea*	CG	C	exonic	Qrob_G0088640.2	7	44647874	28315
Sc0000118_1466708	Pool-seq	T+V		*Q. petraea*	AG	G	exonic	Qrob_G0081080.2	11	4483502	2258901
Sc0000135_261350	Pool-seq	T+V		*Q. petraea*	AC	C	exonic	Qrob_G0222840.2	7	24799454	2918740
Sc0000145_700044	Seq-cap	T+V		*Q. petraea*	CT	T	intergenic	NA	2	31806962	1145225
Sc0000203_707735	Pool-seq	T+V		*Q. petraea*	AT	A	intronic	Qrob_G0237050.2	1	51985124	17848
Sc0000274_909817	Pool-seq	T+V		*Q. petraea*	AG	A	exonic	Qrob_G0320010.2	2	35418361	563863
Sc0000464_236576	Pool-seq	T	PD	*Q. petraea*	CT	T	exonic	Qrob_G0523500.2	5	23753953	744474
Sc0000481_343721	Pool-seq	T+V		*Q. petraea*	AG	G	intronic	Qrob_G0512850.2	5	3222156	22554
Sc0000974_98303	Pool-seq	T	PD	*Q. petraea*	AG	A	exonic	Qrob_G0759540.2	3	47813240	NA
Sc0000485_93093	Pool-seq	T	VA	*Q. pubescens*	AG	A	exonic	Qrob_G0539160.2	1	16067626	3103379
Sc0000314_149731	Pool-seq	T+V		*Q. pubescens*	AT	A	intergenic	NA	2	55831586	5480810
Sc0000314_149731	Pool-seq	T	VA	*Q. pubescens*	CT	T	intergenic	NA	7	39945935	897324
Sc0000047_2398879	Pool-seq	T+V		*Q. pubescens*	AG	A	intronic	Qrob_G0585850.2	9	16825011	NA
Sc0000062_118505	Pool-seq	T	DisC	*Q. pubescens*	CT	T	intronic	Qrob_G0091810.2	6	15935077	4083535
Sc0000088_1796044	Pool-seq	T+V		*Q. pubescens*	AC	A	intronic	Qrob_G0328570.2	12	20880546	NA
Sc0000109_800763	Pool-seq	T+V		*Q. pubescens*	CT	C	exonic	Qrob_G0124910.2	1	43547799	4560446
Sc0000111_693153	Pool-seq	T+V		*Q. pubescens*	CT	T	intergenic	NA	7	28004200	286006
Sc0000170_630013	Pool-seq	T+V		*Q. pubescens*	AC	A	intronic	Qrob_G0200770.2	6	35588545	14530676
Sc0000192_329301	Pool-seq	T+V		*Q. pubescens*	AG	A	intronic	Qrob_G0228240.2	2	39053274	2874167
Sc0000482_334917	Pool-seq	T	PD	*Q. pubescens*	CT	C	intronic	Qrob_G0533790.2	NA	NA	NA

Sc0000403_286465	Pool-seq	T+V		*Q. pyrenaica*	CT	T	intergenic	NA	2	105577540	28340724
Sc0000006_2873224	Pool-seq	T+V		*Q. pyrenaica*	AG	A	intronic	Qrob_G0005870.2	6	21057869	5122792
Sc0000014_2037045	Pool-seq	T+V		*Q. pyrenaica*	AC	C	exonic	Qrob_G0064170.2	2	77236816	6145559
Sc0000053_1344456	Pool-seq	T	DisC	*Q. pyrenaica*	AG	A	intergenic	NA	6	11300343	294652
Sc0000085_73024	Pool-seq	T+V		*Q. pyrenaica*	TG	G	intronic	Qrob_G0563240.2	10	14291154	1973542
Sc0000228_1091905	Pool-seq	T+V		*Q. pyrenaica*	TC	C	intergenic	NA	5	35457995	7637100
Sc0000269_924931	Pool-seq	T+V		*Q. pyrenaica*	AC	A	intronic	Qrob_G0632320.2	2	26007366	561547
Sc0000287_474090	Pool-seq	T	AF	*Q. pyrenaica*	AC	C	exonic	Qrob_G0459590.2	2	71091257	6145559
Sc0000307_852597	Pool-seq	T+V		*Q. pyrenaica*	AG	A	intergenic	NA	7	16871124	7928330
Sc0000517_383812	Pool-seq	T+V		*Q. pyrenaica*	CG	C	intergenic	NA	10	18200824	125219
Sc0000695_157206	Pool-seq	T	AF	*Q. pyrenaica*	AT	A	exonic	Qrob_G0671270.2	8	55531952	68141
Sc0000778_61930	Pool-seq	T+V		*Q. pyrenaica*	CT	T	intronic	Qrob_G0070130.2	10	12317612	1973542
Sc0000013_2578823	Pool-seq	T+V		*Q. robur*	AG	A	intronic	Qrob_G0010260.2	2	61312396	5480810
Sc0000038_794573	Pool-seq	T	PQ	*Q. robur*	TG	G	intronic	Qrob_G0701760.2	1	38987353	4560446
Sc0000053_1639108	Pool-seq	T+V		*Q. robur*	AG	A	intergenic	NA	6	11594995	256547
Sc0000053_1895655	Seq-cap	T	DisC	*Q. robur*	CT	T	exonic	Qrob_G0631440.2	6	11851542	256547
Sc0000099_1839376	Pool-seq	T+V		*Q. robur*	AC	A	intronic	Qrob_G0084290.2	11	6742403	2258901
Sc0000111_979159	Pool-seq	T	AF	*Q. robur*	CT	C	intronic	Qrob_G0135420.2	7	27718194	286006
Sc0000158_462639	Pool-seq	T+V		*Q. robur*	AT	T	exonic	Qrob_G0304430.2	2	26568913	230471
Sc0000158_693110	Seq-cap	T	AF	*Q. robur*	CG	G	exonic	Qrob_G0304580.2	2	26799384	230471
Sc0000203_689887	Pool-seq	T+V		*Q. robur*	TG	T	intronic	Qrob_G0237030.2	1	51967276	17848
Sc0000225_507799	Seq-cap	T	AF	*Q. robur*	CG	G	exonic	Qrob_G0487320.2	5	24498427	744474
Sc0000240_289656	Seq-cap	T	VA	*Q. robur*	AG	G	exonic	Qrob_G0318610.2	5	4277129	1032419
Sc0000270_806328	Pool-seq	T	AF	*Q. robur*	AG	G	exonic	Qrob_G0692980.2	7	33545097	5540897
Sc0000339_4638	Pool-seq	T+V		*Q. robur*	CT	C	intronic	Qrob_G0473660.2	1	3718979	9245268
Sc0000381_206331	Seq-cap	T+V		*Q. robur*	AC	A	intergenic	NA	7	39048611	897324
Sc0000447_521057	Pool-seq	T+V		*Q. robur*	AG	G	exonic	Qrob_G0543330.2	10	22549567	4348743
Sc0000481_366275	Seq-cap	T	PD	*Q. robur*	CT	T	exonic	Qrob_G0512860.2	5	3244710	22554
Sc0000517_258593	Pool-seq	T+V		*Q. robur*	CT	T	intergenic	NA	10	18075605	125219
Sc0000546_456229	Seq-cap	T	PD	*Q. robur*	AG	G	exonic	Qrob_G0761790.2	2	35982224	563863
Sc0000598_295142	Seq-cap	T	PD	*Q. robur*	CT	C	exonic	Qrob_G0575620.2	2	32952187	1145225
Sc0000695_225347	Pool-seq	T+V		*Q. robur*	AT	T	intronic	Qrob_G0671240.2	8	55600093	68141
Sc0000796_82698	Pool-seq	T+V		*Q. robur*	AT	T	intronic	Qrob_G0759570.2	NA	NA	NA
Sc0000967_33996	Pool-seq	T+V		*Q. robur*	AT	T	exonic	Qrob_G0709860.2	2	49368178	2526743

[a] Study samples (T : Training populations ; V : validation population)

[b] Screening criteria from training to validation (PD : Poor diagnosticity; PQ : Poor quality of cluster delimitation ; Disc : Genotype discrepancy between different multiplexes; VA : variable success (numerous missing data) ; AF : Amplification failure after primer redesign

[c] Chr: Chromosome bearing the diagnostic SNP

[d] Position (in bp) on the chromosome

[e] Distance (in bp) with previous diagnostic SNP on the same chromosome

[f] SNP Identification comprise scaffold number (SC#) and position on the scaffold (Plomion *et al,* 2018)

Table 7 in Appendix. Overall frequencies of near-diagnostic alleles in the validation populations.

SNP ID	Reference diagnostic species	Near-diagnostic allele	*Q.petraea*	*Q.pubescens*	*Q.pyrenaica*	*Q.robur*	Diagnosticity
Sc0000254_6223	*Q. petraea*	T	**0.863**	0.116	0.074	0.023	0.792
Sc0000121_355205	*Q. petraea*	C	**0.966**	0.151	0.109	0.068	0.857
Sc0000043_1651618	*Q. petraea*	G	**0.738**	0.156	0.023	0.027	0.669
Sc0000083_147504	*Q. petraea*	A	**0.876**	0.121	0.320	0.046	0.714
Sc0000118_1466708	*Q. petraea*	G	**0.873**	0.117	0.046	0.077	0.793
Sc0000135_261350	*Q. petraea*	C	**0.823**	0.132	0.056	0.082	0.733
Sc0000145_700044	*Q. petraea*	T	**0.949**	0.074	0.015	0.035	0.908
Sc0000203_707735	*Q. petraea*	A	**0.903**	0.060	0.042	0.029	0.859
Sc0000274_909817	*Q. petraea*	A	**0.786**	0.086	0.123	0.018	0.710
Sc0000481_343721	*Q. petraea*	G	**0.926**	0.242	0.085	0.049	0.801
Sc0000314_149731	*Q. pubescens*	A	0.116	**0.660**	0.050	0.054	0.587
Sc0000047_2398879	*Q. pubescens*	A	0.043	**0.896**	0.045	0.019	0.860
Sc0000088_1796044	*Q. pubescens*	A	0.027	**0.906**	0.000	0.007	0.895
Sc0000109_800763	*Q. pubescens*	C	0.121	**0.752**	0.015	0.007	0.704
Sc0000111_693153	*Q. pubescens*	T	0.337	**0.996**	0.061	0.012	0.859
Sc0000170_630013	*Q. pubescens*	A	0.015	**0.657**	0.000	0.005	0.650
Sc0000192_329301	*Q. pubescens*	A	0.010	**0.697**	0.054	0.006	0.674
Sc0000403_286465	*Q. pyrenaica*	T	0.008	0.014	**0.900**	0.011	0.889
Sc0000006_2873224	*Q. pyrenaica*	A	0.012	0.016	**0.962**	0.009	0.950
Sc0000014_2037045	*Q. pyrenaica*	C	0.006	0.000	**0.908**	0.004	0.905
Sc0000085_73024	*Q. pyrenaica*	G	0.004	0.024	**0.946**	0.000	0.937
Sc0000228_1091905	*Q. pyrenaica*	C	0.003	0.004	**0.844**	0.002	0.841
Sc0000269_924931	*Q. pyrenaica*	A	0.003	0.000	**0.900**	0.004	0.898

Sc0000307_852597	*Q. pyrenaica*	A	0.013	0.000	**0.546**	0.022	0.534
Sc0000517_383812	*Q. pyrenaica*	C	0.001	0.022	**0.975**	0.004	0.966
Sc0000778_61930	*Q. pyrenaica*	T	0.004	0.000	**0.969**	0.000	0.968
Sc0000013_2578823	*Q. robur*	A	0.078	0.028	0.047	**0.814**	0.763
Sc0000053_1639108	*Q. robur*	A	0.094	0.015	0.071	**0.925**	0.865
Sc0000099_1839376	*Q. robur*	A	0.017	0.027	0.000	**0.881**	0.866
Sc0000158_462639	*Q. robur*	T	0.043	0.000	0.054	**0.799**	0.767
Sc0000203_689887	*Q. robur*	T	0.072	0.029	0.078	**0.796**	0.736
Sc0000339_4638	*Q. robur*	C	0.076	0.000	0.062	**0.922**	0.876
Sc0000381_206331	*Q. robur*	A	0.054	0.000	0.108	**0.941**	0.887
Sc0000447_521057	*Q. robur*	G	0.213	0.028	0.000	**0.956**	0.876
Sc0000517_258593	*Q. robur*	T	0.104	0.091	0.069	**0.960**	0.872
Sc0000695_225347	*Q. robur*	T	0.140	0.008	0.008	**0.921**	0.869
Sc0000796_82698	*Q. robur*	T	0.083	0.074	0.133	**0.901**	0.804
Sc0000967_33996	*Q. robur*	T	0.077	0.004	0.085	**0.965**	0.910

Frequencies in bold characters correspond to near-diagnostic alleles of the reference diagnostic species

Figure captions

Figure 1

Title: Geographic distribution of the validation populations

Legend: The green area corresponds to the distribution of the species according to Caudullo, Welk et al. (2017).

Red dots correspond to the origins of the validation populations.

Populations identified by their name refer to populations for which frequency profiles of near-diagnostic alleles are later illustrated and discussed (paragaph 3.3.3)..

Figure 2

Title: Genomic location of the near-diagnostic SNPs on the 12 oak (pseudo-)chromosomes of the oak chromosome

Legend: The color code of the marker corresponds to the species name for which the SNP is expected to be diagnostic (our design), with *Q. robur, Q. petraea, Q. pubescens and Q. pyrenaica*, shown in pink, green, blue and yellow, respectively. For each SNP, the diagnosticity of each marker at the training stage is indicated following the proportional and color scale shown. Thin and bold lines both indicate the location of the SNPs, but separates SNPs that were excluded or included in our final set of 38 SNPs, respectively. Note that two diagnostic SNPs are not shown since they are located on scaffolds that are not anchored on the oak pseudochromosomes (see Table 6 in Appendix).

Figure 3

Title: Heat map of frequencies of near-diagnostic alleles and diagnosticity in the validation populations of the four species.

Legend: SNPs are clustered for their diagnostic value for each species (reference species): First to fourth columns correspond respectively to near-diagnostic alleles of *Q. petraea*, *Q.pubescens, Q. pyrenaica* and *Q. robur*.

Figure 4

Title: Frequencies of near-diagnostic alleles of *Q.pubescens, Q. pyrenaica* and *Q. robur* in *Q. petraea* populations

Legend: Populations in red and green correspond to DP and EP populations (see text). Shown are the mean frequencies of all near-diagnostic alleles of a given species (*Q.pubescens, Q. pyrenaica, Q. robur)* in *Q.petraea* populations. Geographic locations of *Q. petraea* populations are shown in Figure 1a.

Figure 5

Title: Biplot of principal components of tree samples based on a Principal Component Analysis (PCA) conducted in the validation populations.

Legend:

Figure 1a : Biplot of component 1 and 2

Figure 5b : Biplot of component 1 and 3

Figure 5c : Biplot of component 2 and 3

Numbers between brackets stand for the percentage of variation explained by the component. Red dots: *Q. petraea* samples; Orange dots : *Q. pubescens* samples; Green dots: *Q. pyrenaica* samples; Blue dots: *Q. robur* samples; Black dots: hybrids, Pet*Pub: *Q.petraea*Q.pubescens* hybrids; Pet*Rob: *Q.petraea*Q.robur* hybrids; Pub*Pyr : *Q.pubescens*Q.pyrenaica* hybrids; Pub*Rob: *Q.pubescens*Q.robur* hybrids.

Figure 6 in Appendix

Title : Heat map of frequencies of near-diagnostic alleles and diagnosticity in the training populations of the four species.

Legend: SNPs are clustered for their diagnostic value for each species (reference species): First to fourth columns correspond respectively to near-diagnostic alleles of *Q. petraea*, *Q.pubescens, Q. pyrenaica* and *Q. robur*.

Figure 7 in Appendix

Title: Frequencies of near-diagnostic alleles of *Q.petraea, Q. pyrenaica* and *Q. robur* in *Q. pubescens* populations.

Legend : Populations in red and green correspond to DP (diverging populations) and EP (expected populations, see text). Shown are the mean frequencies of all diagnostic alleles of a given species (*Q. petraea, Q. pyrenaica, Q. robur*) in *Q.pubescens* populations. Geographic locations of *Q. pubescens* populations are shown in Figure 1b

Figure 8 in Appendix

Title: Frequencies of near-diagnostic alleles of *Q.petraea, Q. pubescens,* and *Q. pyrenaica* in *Q. robur* populations

Legend: Populations in red and green correspond to DP (diverging populations) and EP (expected populations, see text). Shown are the mean frequencies of all diagnostic alleles of a given (*Q.petraea, Q. pubescens, Q.pyrenaica*) species in *Q.robur* populations. Geographic locations of *Q.robur* populations are shown in Figure 1d

Figure 9 in Appendix

Title: Biplot of principal components of tree samples based on a Principal Component Analysis (PCA) conducted in the *Q. pubescens* and *Q. petraea* validation populations

Legend: Red dots: *Q. petraea* samples; Orange dots : *Q. pubescens* samples.

Blue dots : *Q. petraea* Pomieri population. Green dots : *Q. pubescens* Switzerland population

Figure 10 in Appendix

Title: Biplot of principal components of tree samples based on a Principal Component Analysis (PCA) conducted in the *Q. robur* and *Q. petraea* validation populations

Legend: Red dots: *Q. petraea* samples; Blue dots : *Q. robur* samples.

Green dots : *Q. petraea* Killarney population. Orange dots : *Q.robur* Roudsea population

Figure 1a

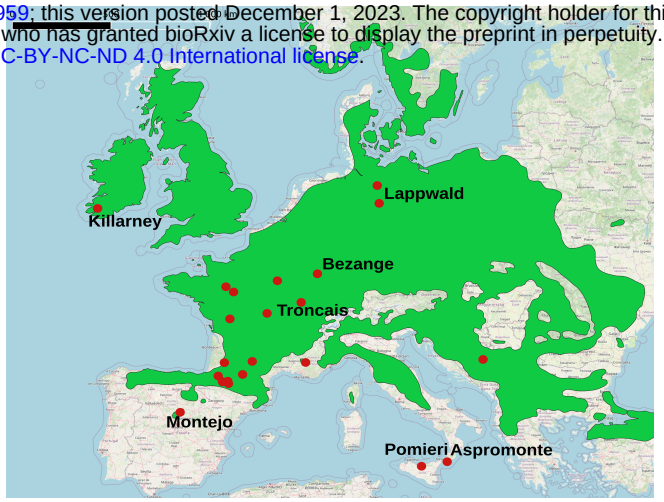


Figure 1b

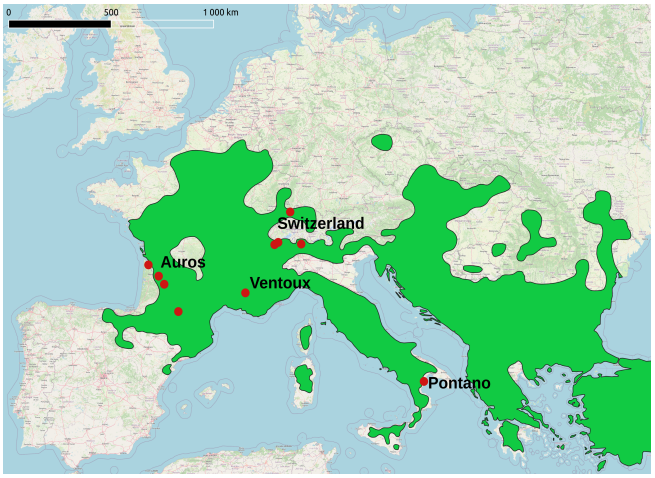


Figure 1c

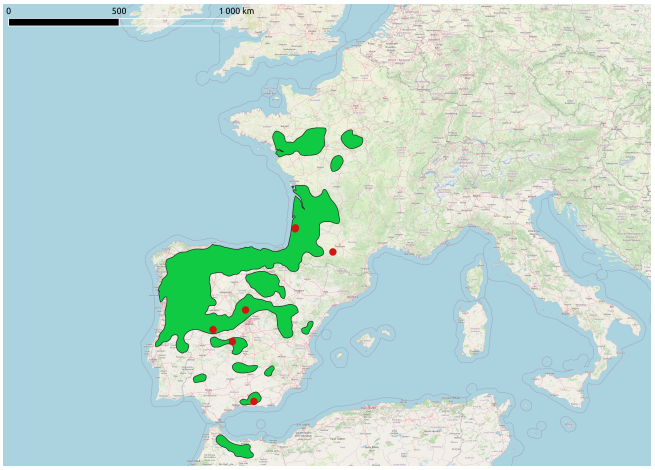
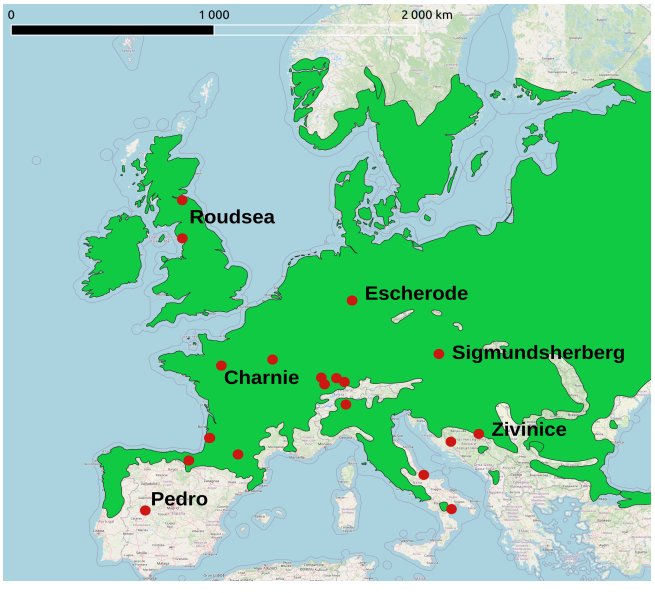


Figure 1d



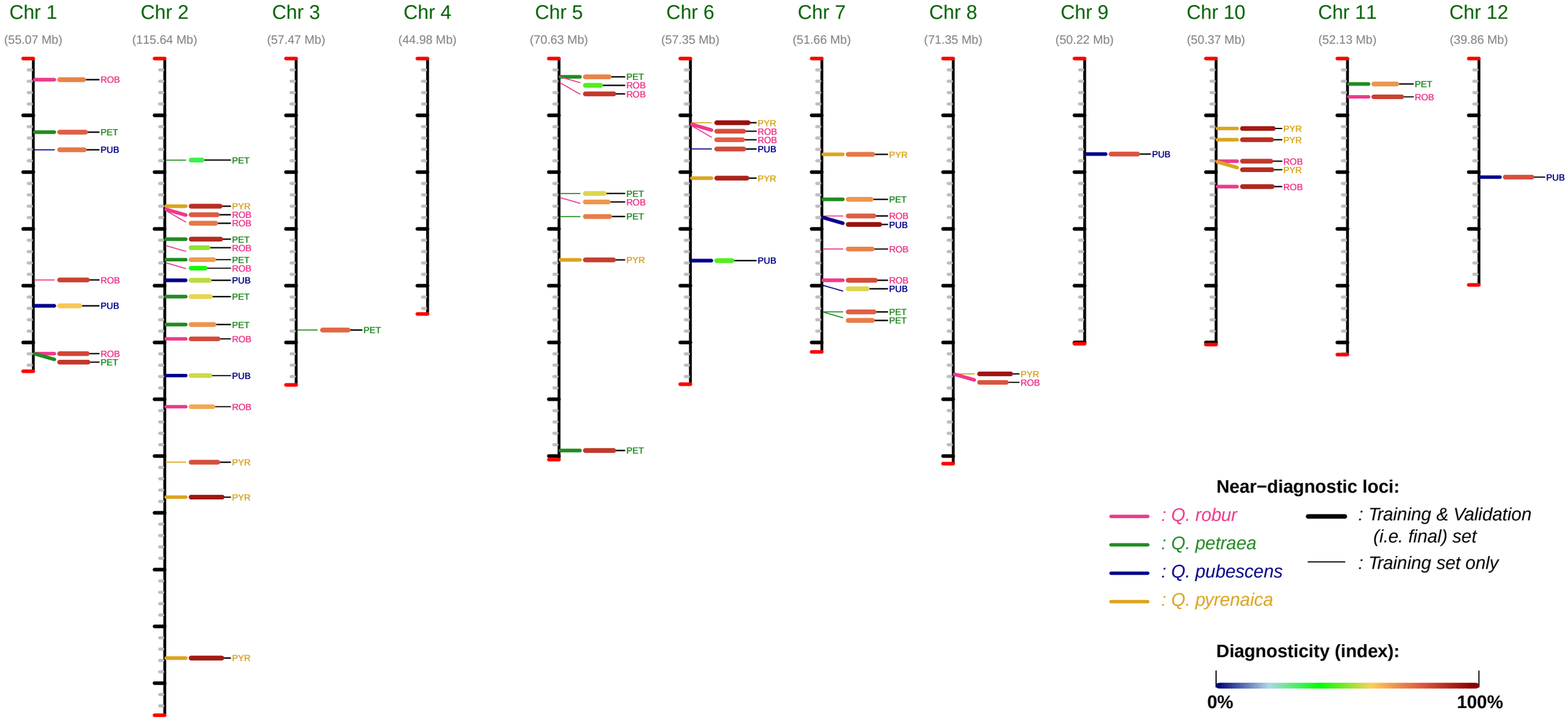

Chr 1
(55.07 Mb)
ROB
PET
PUB
ROB
PUB
ROB
PET
Chr 2
(115.64 Mb)
PET
PYR
ROB
ROB
PET
ROB
PET
ROB
PUB
PET
PET
ROB
PUB
ROB
PYR
PYR
Chr 3
(57.47 Mb)
PET
Chr 4
(44.98 Mb)
Chr 5
(70.63 Mb)
PET
ROB
ROB
PET
ROB
PET
PYR
PET
Chr 6
(57.35 Mb)
PYR
ROB
ROB
PUB
PYR
PUB
Chr 7
(51.66 Mb)
PYR
PET
ROB
PUB
ROB
ROB
PUB
PET
PET
Chr 8
(71.35 Mb)
PYR
ROB
Chr 9
(50.22 Mb)
PUB
Chr 10
(50.37 Mb)
PYR
PYR
ROB
PYR
ROB
Chr 11
(52.13 Mb)
PET
ROB
Chr 12
(39.86 Mb)
PUB
Near−diagnostic loci:
: Q. robur
: Q. petraea
: Q. pubescens
: Q. pyrenaica
: Training & Validation (i.e. final) set
: Training set only
Diagnosticity (index):
0%
100%

Figure 2

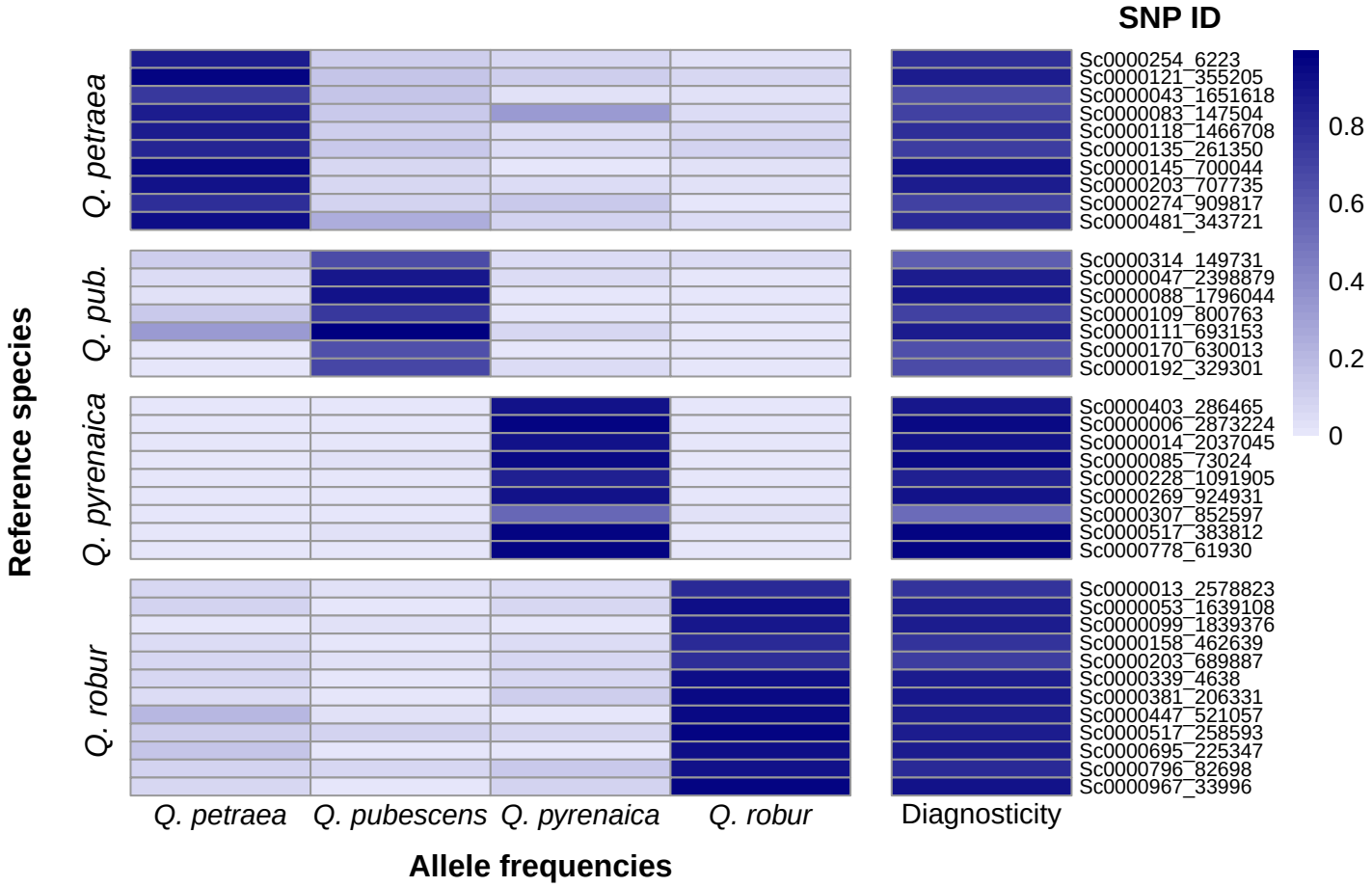

SNP ID
Reference species
Q. petraea
Q. pub.
Q. pyrenaica
Q. robur
Sc0000254_6223
Sc0000121_355205
Sc0000043_1651618
Sc0000083_147504
Sc0000118_1466708
Sc0000135_261350
Sc0000145_700044
Sc0000203_707735
Sc0000274_909817
Sc0000481_343721
Sc0000314_149731
Sc0000047_2398879
Sc0000088_1796044
Sc0000109_800763
Sc0000111_693153
Sc0000170_630013
Sc0000192_329301
Sc0000403_286465
Sc0000006_2873224
Sc0000014_2037045
Sc0000085_73024
Sc0000228_1091905
Sc0000269_924931
Sc0000307_852597
Sc0000517_383812
Sc0000778_61930
Sc0000013_2578823
Sc0000053_1639108
Sc0000099_1839376
Sc0000158_462639
Sc0000203_689887
Sc0000339_4638
Sc0000381_206331
Sc0000447_521057
Sc0000517_258593
Sc0000695_225347
Sc0000796_82698
Sc0000967_33996
0.8
0.6
0.4
0.2
0
Q. petraea
Q. pubescens
Q. pyrenaica
Q. robur
Diagnosticity
Allele frequencies

Figure 3

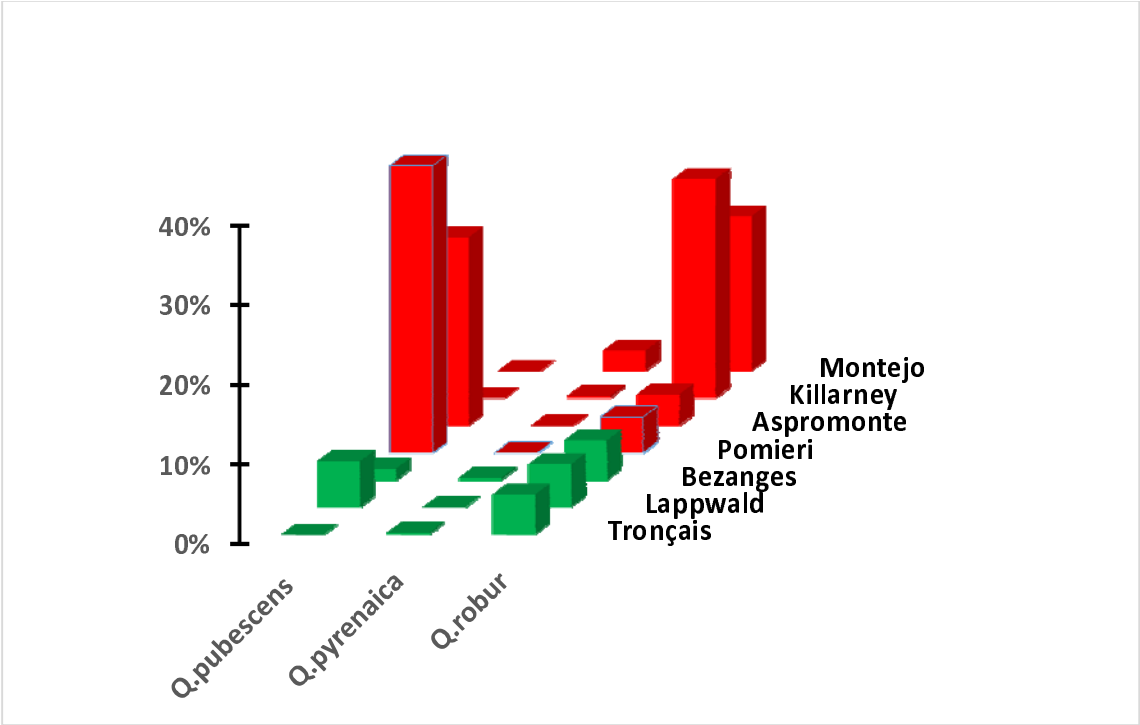


Figure 4

Figure 5a

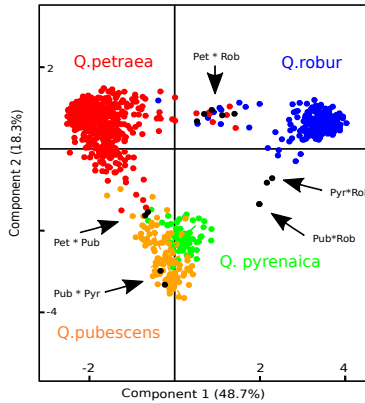


Figure 5b

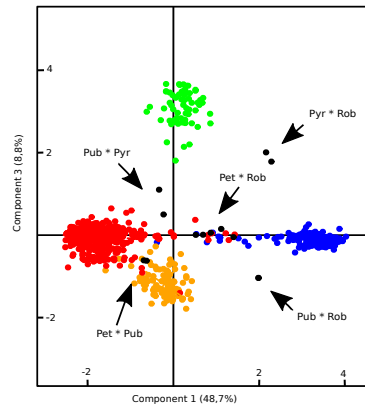
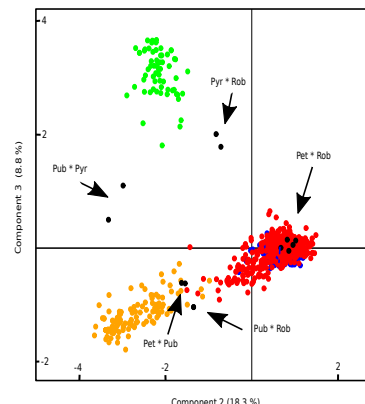


Figure 5c



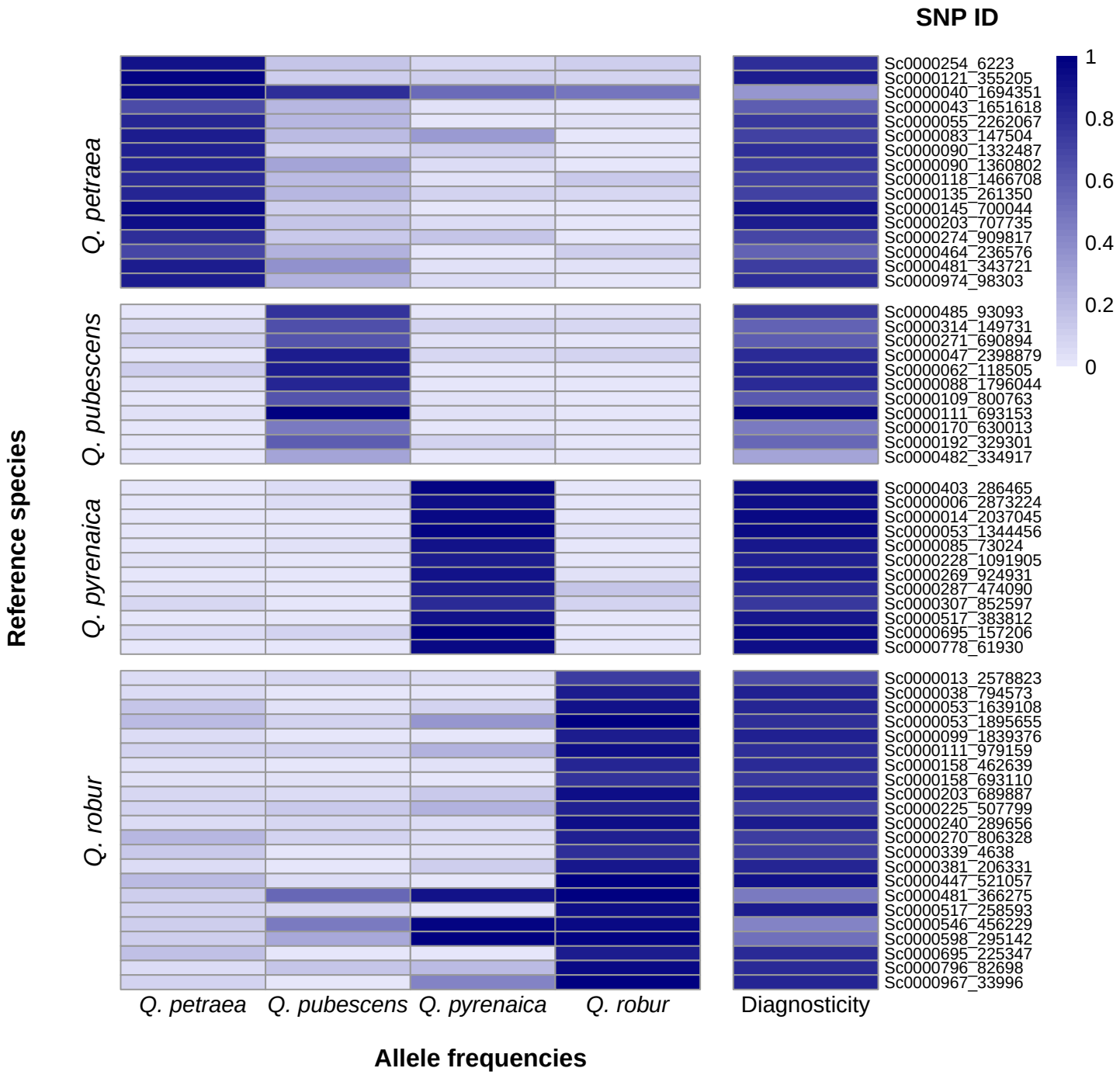


Figure 6 in Appendix

.

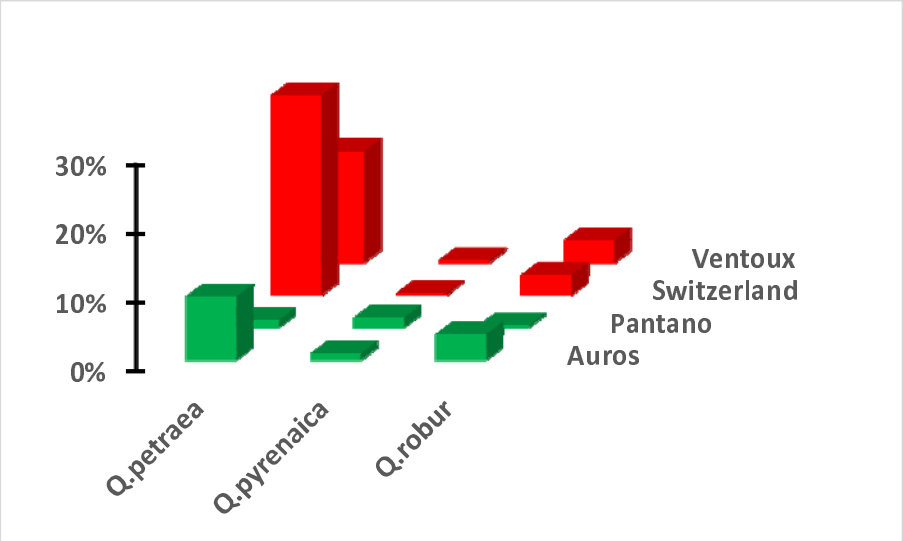


Figure 7 in Appendix

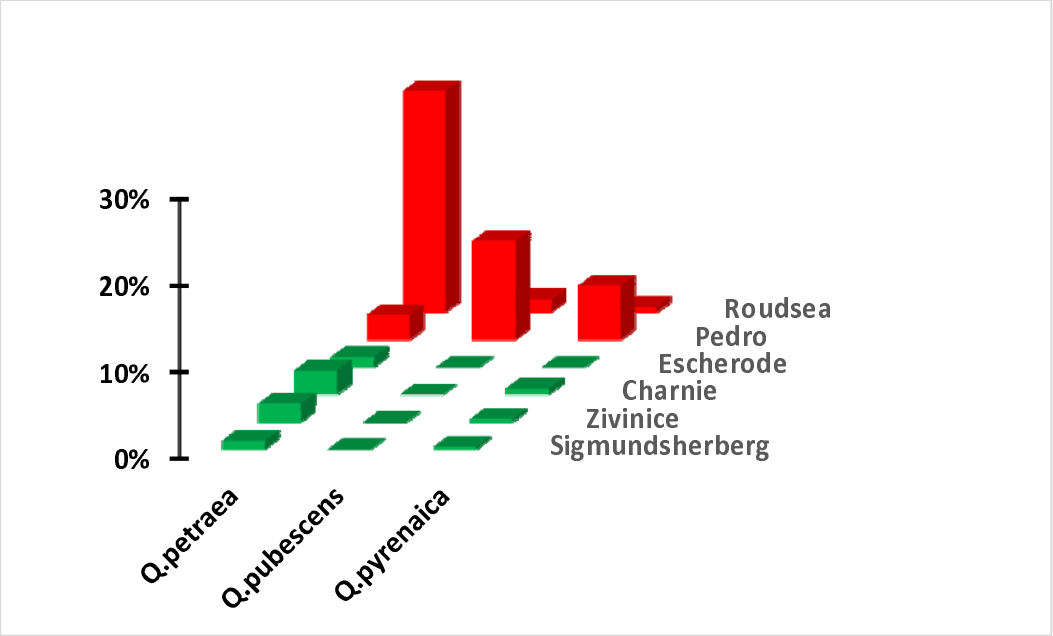


Figure 8 in Appendix

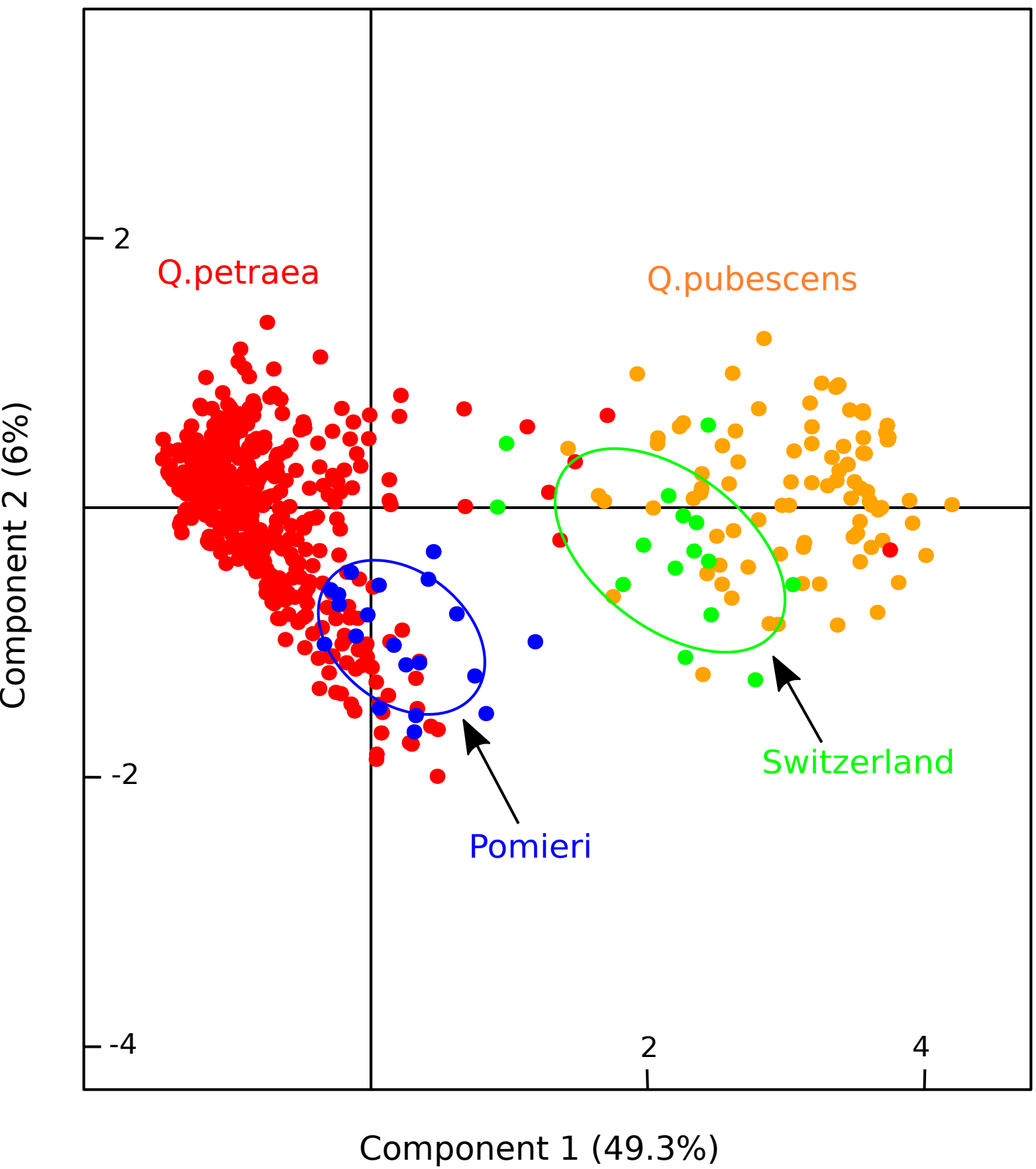


Figure 9 in Appendix

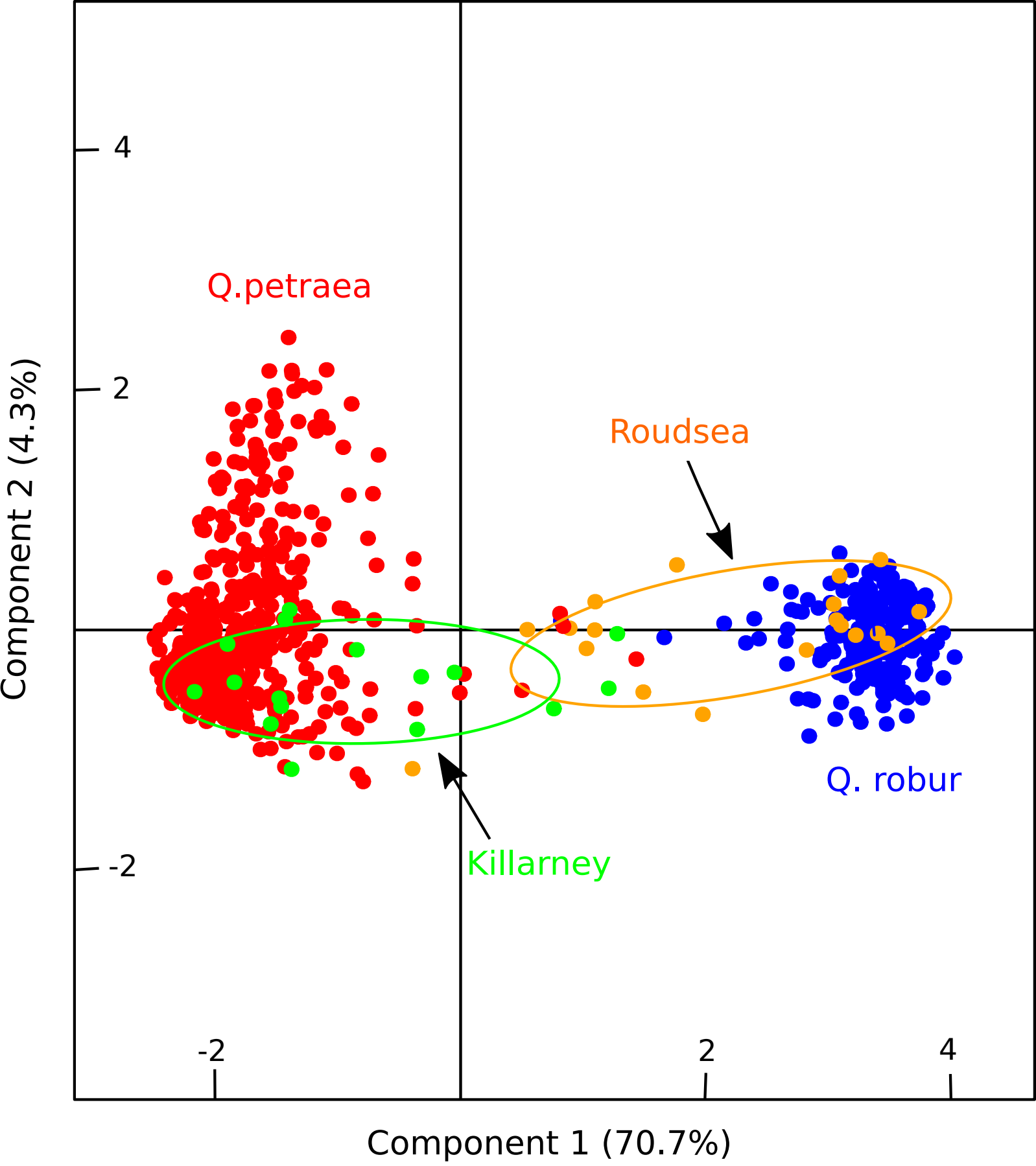


Figure 10 in Appendix