High-quality SNPs from genic regions highlight introgression patterns among European white oaks (*Quercus petraea* and *Q. robur*).

.

Authors: Tiange Lang[1,2,3], Pierre Abadie[1,2], Valérie Léger[1,2], Thibaut Decourcelle[1,2,4], Jean-Marc Frigerio[1,2], Christian Burban[1,2], Catherine Bodénès[1,2], Erwan Guichoux[1,2], Grégoire Le Provost[1,2], Cécile Robin[1,2], Naoki Tani[1,2,5], Patrick Léger[1,2], Camille Lepoittevin[1,2], Veronica A. El Mujtar[1,2,6], François Hubert[1,2], Josquin Tibbits[7], Jorge Paiva[1,2,8,9], Alain Franc[1,2], Frédéric Raspail[1,2], Stéphanie Mariette[1,2], Marie-Pierre Reviron[1,2], Christophe Plomion[1,2], Antoine Kremer[1,2], Marie-Laure Desprez-Loustau[1,2], Pauline Garnier-Géré[1,2,§]

Addresses :

[1]INRA, UMR 1202 Biodiversity Genes & Communities, F-33610 Cestas, France

[2]Univ. Bordeaux, UMR 1202, Biodiversity Genes & Communities, F-33400 Talence, France

[3]Big Data Decision Institute, Jinan University, Tianhe, Guangzhou, PR China

[4]GEVES, 25 rue Georges Morel, 49071, Beaucouzé, France

[5] Japan International Research Center for Agricultural Sciences (JIRCAS), Tsukuba, Ibaraki, Japan

[6]Unidad de Genética Ecológica y Mejoramiento Forestal. Instituto Nacional de Tecnología Agropecuaria (INTA) EEA Bariloche, Modesta Victoria 4450 (8400), Bariloche, Río Negro, Argentina

[7] Department of Environment and Primary Industries, Biosciences Research Division, Agribio, 5 Ring Road, Bundoora, Victoria, 3086, Australia

[8] Instituto de Biologia Experimental e Tecnologica, iBET, Apartado 12, Oeiras 2780-901, Portugal

[9] Institute of Plant Genetics, Polish Academy of Sciences, 34 Strzeszynska street, Poznan PL-60-479, Poland

Keywords: SNPs, functional candidate genes, *Quercus robur, Q. petraea*, Sanger amplicon resequencing, introgression, species differentiation

[§]Corresponding author Pauline Garnier-Géré

INRA, UMR 1202 Biodiversity Genes & Communities, F- 33610 Cestas, France; Univ. Bordeaux, UMR 1202 Biodiversity Genes & Communities, Bordeaux, F-33400 Talence, France

Fax +33 (0)35385381, email: pauline.garnier-gere@inra.fr

Running title: High-quality SNPs for *Quercus* species

Abstract

In the post-genomics era, non-model species like most *Fagaceae* still lack operational diversity resources for population genomics studies. We completed the analyses of Sanger sequences produced from over 800 gene fragments covering ~530 kb across the genic partition of European oaks in a range-wide sampling of 25 individuals (11 *Quercus petraea*, 13 *Q. robur*, one *Q. ilex* as an outgroup). Regions targeted represented broad functional categories potentially involved in species ecological preferences, and a random set of genes. Using a high-quality dedicated pipeline, we provide a detailed characterization of over 14500 polymorphisms, including ~12500 SNPs -218 being triallelic-, over 1500 insertion-deletions, and ~200 novel di- and tri-nucleotide SSR loci. This catalog also includes various summary statistics within and among species, gene ontology information, and standard formats to assist loci choice for genotyping projects. The distribution of nucleotide diversity and differentiation across genic regions are also described for the first time in those species (mean $\theta\pi$ close to ~0.0049 in *Q. petraea* and to ~0.0045 in *Q. robur* across random regions, and mean F_{ST} ~0.13 across SNPs). Robust patterns were observed which emphasize a slightly but significantly higher diversity in *Q. petraea*, across a random gene set and in the abiotic stress functional category, and a heterogeneous landscape of both diversity and differentiation. These patterns are discussed in the context of both species documented introgression history despite strong reproductive barriers. The quality, representativity in terms of species genomic diversity, and usefulness of the resources provided are discussed for possible applications in medium scale landscape ecology projects, and as a reference resource for validation purposes in larger-scale re-sequencing projects. These are preferentially recommended in oaks in contrast to SNP array development, given the large nucleotide variation and low levels of linkage disequilibrium revealed.

Introduction

High-throughput (HT) techniques of the next-generation sequencing (NGS) era and increased genome sequencing efforts in the last decade have greatly improved access to genomic resources in non-model forest tree species, but these have only been applied recently to large-scale ecological and population genomics research (Neale and Kremer 2011, Neale *et al*. 2013; Plomion *et al*. 2016, Holliday *et al*. 2017). One notable exception are studies undertaken in the model genus *Populus* (e.g. Zhou *et al*. 2014, Geraldes *et al*. 2014, Christe *et al*. 2016b) that benefited from the first genome sequence completed in 2006 in *P. trichocarpa* (Tuskan *et al*. 2006). In *Fagaceae*, previous comparative mapping and "omics" technologies

(reviewed in Kremer *et al*. 2012) with recent development of genomic resources (e.g. Faivre-Rampant *et al*. 2011; Tarkka *et al.* 2013; Lesur *et al*. 2015; Lepoittevin *et al*. 2015, Bodénès *et al*. 2016) set the path to very recent release of genome sequences to the research community (*Quercus lobata*, Sork *et al.* 2016; *Q. robur*, Plomion *et al*. 2016, 2018; *Q. suber*, Ramos *et al*. 2018; *Fagus sylvatica*, Mishra *et al*. 2018), and these provide great prospects for future evolutionary genomics studies (Petit *et al*. 2013; Parent *et al*. 2015; Cannon *et al*. 2018; Lesur *et al*. 2018).

Recently, building from the European oaks genomic resources (*Quercus Portal*, htpps://arachne.pierroton.inra.fr, and references therein), natural populations of 4 *Quercus* species (*Q. robur, Q. petraea, Q. pyrenaica, Q. pubescens*) were genotyped for ~4000 single-nucleotide polymorphisms (SNPs, from an initial 8K infinium array, Lepoittevin *et al*. 2015). The data were further analysed (Leroy *et al*. 2017), with results extending previous knowledge on their likely diversification during glacial periods, as well as their recolonization history across Europe and recent secondary contacts (SC) after the last glacial maximum (Hewitt 2000; Petit *et al*. 2002; Brewer *et al*. 2002). Using recent model-based inference allowing for heterogeneity of migration rates (Roux *et al*. 2014; Tine *et al.* 2014), Leroy *et al*. (2017) showed that the most strongly supported scenarios for all species pairs included very recent SC, due to a much better fit of the patterns of large heterogeneity of differentiation observed across SNP loci. These recent SC events have been documented in many patchily distributed hybrid zones where current *in situ* hybridization can occur among European oak species (e.g. Curtu *et al*. 2007; Jensen *et al.* 2009; Lepais and Gerber 2011; Guichoux *et al*. 2013). The resulting low levels of differentiation among *Q. robur* and *Q. petraea* in particular is traditionally linked to a model of contrasted colonization dynamics, where the second-in-succession species (*Q. petraea*) is colonizing populations already occupied by the earlier pioneering *Q. robur* (Petit *et al.* 2003). This model predicts asymmetric introgression towards *Q. petraea* (see Currat *et al*. 2008), as often observed in interspecific gene exchanges (Abbott *et al*. 2003), and a greater diversity in *Q. petraea* was documented at SNP loci showing higher differentiation (Guichoux *et al.* 2013). The directionality of introgression in oaks was also shown to depend on species relative abundance during mating periods (Lepais *et al*. 2009, 2011).

Nethertheless, oaks like other hybridizing taxa are known for the integration of their species parental gene pools and strong reproductive isolation barriers (Muir *et al*. 2000; Muir and Schlötterer 2005; Abadie *et al.* 2012, Lepais *et al*. 2013; Ortiz-Barrientos and Baack 2014;

Christe *et al*. 2016a), raising essential questions about the interacting roles of divergent (or other types of) selection, gene flow, and recombination rates variation in natural populations, and their imprints on genomic molecular patterns of variation (e.g. Zhang *et al*. 2016; Christe *et al*. 2016b; Payseur and Rieseberg 2016). These issues will be better addressed with genome-wide sequence data in many samples (Buerkle *et al*. 2011), which will be facilitated in oaks by integrating the newly available genome sequence of *Quercus robur* to chosen HT resequencing methods (Jones and Good 2016; e.g. Zhou and Holliday 2012; Lesur *et al*. 2018 for the first target sequence capture in oaks).

However, obtaining high quality haplotype-based data required for nucleotide diversity estimation and more powerful population genetics inferences will likely require the development of complex bioinformatics pipelines dedicated to high heterozygosity genomes and solid validation methods for polymorphism detection (e.g. Geraldes *et al*. 2011; Christe *et al*. 2016b). Indeed, very stringent filters are often applied in practice to limit error rates and avoid false-positives, hence limiting the impact of variable read depth and possible ascertainment bias risks, which altogether significantly decrease the number of informative loci compared to either initial fixed amounts (in genotyping arrays, e.g. Lepoittevin *et al*. 2015) or potential amounts (in reference genomes, e.g. Pina-Martins *et al*. 2018 in *Quercus* species; see also Van Dijk *et al*. 2014).

Therefore, the objectives of this work were first to provide a detailed characterization of sequence variation in *Quercus petraea* and *Quercus robur*. To that end, we validated previous unpublished Sanger sequence data for fragments of targeted gene regions in a panel of individuals sampled across a large part of both species geographic range. Both functional and expressional candidate genes potentially involved in species ecological preferences, phenology and host-pathogen interactions were targeted, as well as a reference set of fragments randomly chosen across the last oak unigene (Lesur *et al*. 2015). These data were obtained within the framework of the EVOLTREE network activities (http://www.evoltree.eu/). Second, we aimed at estimating the distributions of differentiation and nucleotide diversity across these targeted gene regions for the first time in those species, and further test the robustness of comparative diversity patterns observed in the context of both species contrasted dynamics and introgression asymmetry. We discuss the quality, representativity and usefulness of the resources provided for medium scale genotyping landscape ecology projects or as a reference resource for validation purposes in larger-scale resequencing projects.

Material and methods

Sample collection

The discovery panel (*DiP*) included 25 individuals from 11 widespread forest stands with 2 to 4 individuals per location (13 from *Q. robur*, 11 from *Q. petraea*, 1 from *Q. ilex* to serve as outgroup, in Table 1). These stands occur across a large part of both *Quercus* species natural distributions, spanning ~20° in longitude (~2200 km) and ~11° in latitude (~1250 km) in western and central Europe (Fig. S1, Supporting Information). Individuals were chosen either on the basis of their differing leaf morphology among *Q. robur* and *Q. petraea* species (Kremer *et al*. 2002), or as parents of mapping pedigrees (e.g. Bodénès *et al*. 2016, see Table 1). Leaves were sampled, stored in silica gel and sent to INRA (Cestas, France) for DNA extraction following Guichoux *et al*. (2013). DNA quality and concentration were assessed with a Nanodrop spectrophotometer (NanoDrop Technologies, Wilmington, 152 DE, USA) and by separating samples in 1% agarose gels stained with ethidium bromide. Extractions were repeated until we obtained at least 20 micrograms of genomic DNA per sample, which was needed for a few thousands individual PCRs.

Choice of genic regions

Genic regions were chosen from over 103 000 Sanger sequences available in expressed sequence tags (EST) databases at the start of the project. These sequences correspond to 14 cDNA libraries obtained from various tissues and developmental stages (bud, leaf, root and wood-forming tissues), and thus likely to target a large range of expressed genes. Overall, 146 individuals were sampled in 3 different French regions (South-West, North-East and North-West). We performed the first working assembly for those sequences, with the main aim of avoiding paralog assembly while limiting split contigs with overlapping homolog sequences, the final assembly including 13477 contigs and 74 singletons (Appendices S1 and S2, Fig. S2-A, Supporting information). The libraries used in this assembly have since been named A, B, F to O, and S, and were included in larger transcriptome resources for *Quercus* species (Ueno *et al*. 2010).

In parallel, expressional and functional candidate genes information was compiled for targeting those potentially involved in white oaks' local and/or divergent adaptive traits (Fig. S2-B and Table S1, Supporting information). Briefly, model species databases were searched for gene accessions by gene ontology (GO) and metabolic pathways keywords. Those sequences were first Blasted against our oak assembly (Altschul *et al*. 1990, 1997). Second,

the sequences from their best hits were extracted (see filtering criteria in Fig. S2-B, Supporting information) and re-Blasted against the non-redundant protein (NR) database at NCBI. Third, their annotation was compared to those of the initial gene accessions, allowing 95% of hits from the oak assembly to be validated (step 2 in Fig. S2-B, Supporting information). Expressional candidate genes sequences from bud tissues or stress treatment libraries and a random set of ESTs were also directly sampled across the oak assembly generated above (see Table S1, column F, Supporting information). Primers were designed with the OSP software (Hillier and Green 1991) by setting up homogenous melting temperatures constraints and excluding low-complexity propositions. Predicted amplicons were Blasted against each other and onto our assembly to exclude those with potential amplification problems and multiband patterns. They were also checked for their depth and presence of polymorphisms in contigs alignment, yielding finally 2000 amplicons for resequencing (Fig. S2-B, Supporting information).

Data production and polymorphism discovery

All the sequencing work was performed by Beckman Coulter (Agencourt Bioscience Corporation, Beverly, MA, USA) on ABI3730 capillary sequencers (Applied Biosciences) after preparing DNA samples according to the company's guidelines. Data quality steps were designed throughout the process in order to maximize the amount and quality of the sequences finally obtained (Fig. 1-A). Forward and reverse sequences were produced for 981 amplicons across 25 individuals (100+881 in steps 2 and 3, Fig. 1-A), and more than 85% of them yielded at least 12 high-quality sequences (Fig. 1-B and column L in Table S1, Supporting information). All amplicon assembly steps, merging, trimming, and filtering/masking based on quality were performed with Bioperl scripts from our *SeqQual* pipeline, available at https://github.com/garniergere/SeqQual with examples of data and command files. This repository compiles and extends former work dealing with 454 data (Brousseau *et al*. 2014; El Mujtar *et al.*2014), providing scripts used here that automatically deal with Sanger haploid or diploid DNA sequences and allow fasta files post-processing in batch (Fig. 1-C). Polymorphism discovery was finally performed on nucleotide data with an error rate below 0.001 (i.e. Phred score above 30, Ewing *et al*. 1998, and see Appendix S1, Supporting information for more details). Simple sequence repeat (SSR) patterns were further detected or confirmed from consensus sequences using the *mreps* software (Kolpakov *et al.* 2003, Fig. 1-D). Various additional steps involving the treatment of insertion-deletion

polymorphisms (indels) and heterozygote indels (*HI*) in particular, allowed missing data from polymorphic diploid sequence to be minimized (see Appendix S1, Supporting information).

Functional annotation

BlastN best hits for our working assembly (*orict)* original contigs and for the amplified expected fragments (*orict-cut*) were first retrieved using Lesur *et al*. (2015)' most recent oak assembly (*ocv4*, see Table S2-C, Supporting information). Consensus of candidate regions originated from both *orict* and *ocv4* (396 and 368 respectively, see Table S2-A, S2-B, and Appendix S1, Supporting information), aiming at retrieving the longest consensus sequences that included the resequenced gene regions, while avoiding to target those with possible chimeric sequences likely different to the regions being successfully resequenced. Functional annotation was then performed via homology transfer using BlastX 2.6.0+ program at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with parameters to optimize speed, hits' annotation description and GO content (Fig. 1-E and Table-S2, Supporting information). Retrieval of GO terms were performed with Blast2GO (Conesa *et al.* 2005 free version at https://www.blast2go.com/blast2go-pro/b2g-register-basic) and validation of targeted annotations with Fisher Exact enrichment tests (Appendix S1, supporting information).

Characterization of diversity and genetic clustering

Using the *SNP-stats* script for diploid data (see above), simple statistics were computed across different types of polymorphisms (SNPs, indels, SSRs…) including minimum allele frequencies (*maf*) and heterozygote counts, Chi-square tests probability for Hardy-Weinberg proportions, G_{ST} (Nei 1987) and G_{ST}' standardized measure (Hedrick 2005). Complex polymorphisms (involving *HI* and/or SSRs) were also further characterized, and data formatted or analyzed using either Arlequin 3.5 (Excoffier and Lischer 2010), *SeqQual* (e.g. for Arlequin input file with phase unknown, Fig. 1-C), or R scripts. Nucleotide diversity $\theta\pi$ (Nei 1987), based on the average number of pairwise differences between sequences, and its evolutionary variance according to Tajima (1993), were also estimated and compared among species and across candidate genes grouped by broad functional categories (see column F in Table S1, Supporting information), and Weir and Cockerham (1984) F_{ST} estimates of differentiation were computed among species for SNP data along genic regions using analyses of molecular variance (Excoffier 2007).

The initial morphological species samples were compared to the genetic clusters obtained with the STRUCTURE v2.3.3 inference method (Falush *et al.* 2003) in order to test possible

levels of introgression across individuals. We used the admixture model allowing for mixed ancestry and the correlated allele frequencies assumption for closely related populations as recommended defaults, and since they best represent previous knowledge on both species divergence across their range (e.g. Guichoux *et al.* 2013). Preliminary replicate runs using the same sample of loci produced very low standard deviation across replicates of the data log likelihood given K (ln Pr(X/K), see Fig. S3-A, Supporting information). We thus resampled loci at random for each of 10 replicate datasets in 3 different manners to add genetic stochasticity: 1) one per region, 2) one per 100 bp block, and 3) one per 200 bp block along genes (see Appendix S1, Supporting information and https://github.com/garniergere/Reference.Db.SNPs.Quercus/tree/master/STRUCTURE.files for examples of STRUCTURE files as recommended by Gilbert *et al.* (2012), along with R scripts for outputs). Statistical independence among loci within each species was verified with Fisher's exact tests implemented in Genepop 4.4 (Rousset 2008).

Results

Polymorphisms typology and counts

Among the amplicons tested, 986 were successful, 13 did not produce any data and 23 were excluded because of paralog amplifications (Fig. 1-C and Table S1, Supporting information). Around 25% of the successful amplicons overlapped and were merged, consistently with their original design across contigs. Despite the presence of *HI* patterns due to SSR or indels, most amplicons were entirely recovered with forward and reverse sequencing. Several (5% of the total) were however kept separate, either because of functional annotation inconsistency, or because amplicon overlap was prevented by the presence of SSRs or putative large introns (see "Final gene region ID" column with -F/-R suffix in Table S1, Supporting information). We finally obtained 852 genic regions covering in total ~529 kilobases (kb), with an average size of 621 bp per region, ranging from 81 to 2009 bp (Table 2, and Appendix S4, Supporting information, for genomic consensus sequences). Compared to the EST-based expected total fragment size of ~ 357 kb, around 187 kb of intron sequence was recovered across 460 resequenced regions (assuming intron presence if an amplicon size was above its expected size by 40 bp). Introns represented ~35% of genic regions in length and ~51% of those including introns.

We observed 14102 polymorphisms in both species across 852 gene regions, 15 of those regions (<2%) being monomorphic (Table 2). This corresponds to 1 polymorphism per ~38 bp, or 1 per ~30 bp when considering the total number of variant positions in both species

(17594 bp, Table 2). Remarkably, variant positions involving larger indels, SSRs and mixed complex polymorphism patterns represented ~30% of the total variant positions (Table 2, and see their exhaustive lists with various statistics in Table S3 and S4, Supporting information). We observed 12478 SNPs (88.5% of all polymorphisms), 1 SNP per 42 bp, and 218 triallelic SNPs (~1.75% of SNPs) were confirmed by visual examination of chromatograms. Considering only one species, 1 variant position per ~48 bp, 1 polymorphism per ~60 bp, and 1 SNP per ~68 bp were observed on average. Among indels, 1213 (8.6% of all polymorphisms) were single base, 309 ranged from 2 to 10 bp, and 102 had sizes above 10 bp which were mostly shared among species (Table 2). In this range-wide sample, there were 4334 singletons among all single base polymorphisms and 506 of them were indels. Overall, indels were present in 69% of gene regions and non-single base ones across ~30% of them. Excluding homopolymers (see Appendix S1, Supporting information), we detected 201 SSRs occurring on 163 gene regions by considering a minimum repeat numbers of 4 and a mismatch rate among repeats below 10% (Table 2, Table S1 and Table S5, Supporting information), and 55% (111) were polymorphic in our sample of individuals (Table 2). Among them, 89 (44%) had dinucleotide repeats and 65 (32%) trinucleotide repeats. The SSRs with the lowest number of repeats (<5) had a majority (59%) of repeat sizes between 4 and 7, the rest being trinucleotides (Table S5, Supporting information).

Given the PCR conditions for sequence acquisition and their high similarity *a priori*, homologous sequence data were obtained for one individual of the outgroup *Quercus ilex* across 37% of the gene regions (~197 kb, 397 sequences, 676 heterozygous sites in Table 2, Table S1 column Q, and see Appendix S5, Supporting information, for *Q. ilex* genomic sequences).

Annotations and GO term distributions

BlastX matches with *E*-values below 10^{-30} were found for ~97% (738/764) of the contig consensus, only 11 sequences (1.4%) having hits with *E*-values above 10^{-10} that were all among the reference random sample (see BlastX criteria in Table S2, Supporting information). The most represented species among the best hits with informative annotations were *Prunus persica (111)*, *Theobroma cacao (91)*, *Morus notabilis (57)* and *Populus trichocarpa (45)* (Appendix S6-A, Supporting information), which probably illustrates both the close phylogenetic relationships among *Quercus* and *Prunus* genera, consistently with results obtained on the larger *ocv4* assembly (Lesur *et al.* 2015), and the quality and availability of *P. persica* genome annotation (Verde *et al*. 2013, 2017).

Between 1 to 30 GO terms could be assigned to 761 sequences, with EC codes and InterProScan identifiers for 343 and 733 of them respectively (Fig. 1, and Table S2, Supporting information). The most relevant GO terms were then retained using the Blast2GO "annotation rule" (Conesa *et al*. 2005) that applies filters from the Direct Acyclic Graph (DAG) at different levels (Fig. S4-A- to-F, Supporting information). At biological level (BP) 3, apart from general terms involving "metabolic processes", a large number of sequences (between ~100 and ~150) were mapped to "response to..." either "...stress", "...abiotic stimulus" or "...chemical", and also to categories linked to developmental processes (Fig. S4-D, Supporting information). Enrichment tests also revealed a significant increase at both BP levels 2 and 3 for the following GO categories: "response to stress" or "external stimulus" for *bud* and *biotic* gene lists, "response to abiotic stimulus" for the *bud* list, and "immune" and "biotic stimulus" responses for the *biotic* list (see Fig. 2-B to 2-D compared to Fig. 2-A, and Fig. S5, Supporting information). Most of these exact tests (>80%) were still significant when selecting genes attributed exclusively to one particular list, which adds to the relevance of our gene lists in targeting particular functional categories.

Species assignment and introgressed individuals

In both species, the proportion of significant association tests among the loci used for clustering (> 2 million within each species) was generally one order of magnitude below the type-I error rates at 5% or 1%. This indicates a very low background LD within species at their range levels, consistently with the underlying model assumptions used in STRUCTURE. Based on both ln Pr(X/K) and ΔK statistics and as expected, the optimal number of genetic clusters inferred was 2, whatever the number of polymorphisms and type of sampling (Fig. 3, Fig. S3 and S6, Supporting information).

Most individuals (20) clearly belonged to either cluster with a mean probability of cluster assignment above 0.9 , which was not significantly different from 1, based on mean values of 90% Bayesian credible intervals (BCI) bounds across replicates, and for different types of sampling or SNP numbers (Fig. 3 and Fig. S6, Supporting information). Two individuals from Roudsea Wood in UK, the most northerly forest stand of this study, were significantly introgressed, each from a different cluster, with mean probabilities within 0.125 and 0.875. These values can be considered as typical for back-crosses and later-generation hybrids (Guichoux *et al*. 2013), suggesting a mixed ancestry with the other species a small number of generations in the past. In the initial morphological *Q. petraea* group, two individuals were clearly of recent mixed ancestry: one from the easternmost forest stand of Sopron (S444), and

another one (Qs28) from central France, considered previously to be a *Q. petraea* parental genotype in two oak mapping pedigrees (Bodénès *et al*. 2012, 2016; Lepoittevin *et al.* 2015). However, Qs28 shows here a clear F1 hybrid pattern, given its probability values close to 0.5 and its BCI maximum upper and minimum lower bound values of 0.30 and 0.61 respectively across runs (Fig. 3 and Fig. S6-A to S6-J, Supporting information). Testing 3 or 4 possible clusters showed the same ancestry patterns for the introgressed individuals with 2 main clusters and similar *Q*-values (data not shown), which does not support alternative hypotheses of introgression from different species in those individuals.

Large heterogeneity of diversity and differentiation across genes

Nucleotide diversity was estimated in each parental species after excluding the 4 most introgressed individuals from each initial morphological group (see above). We then checked how the remaining samples represented species' diversity. Starting with one individual, we observe a dramatic drop in the mean proportion of new variant positions brought by each new individual in any species (*Mpn*) as a function of the initial sample size, followed by a subsequent stabilization (Fig. 4-A, and see Fig. S7-B, Supporting information). Indeed, *Mpn* was only around 11% when going from 4 to 5 individuals in both species, and stabilized below 5% after 8 individuals in *Q. robur* **(**Fig. 4-A). We thus decided to retain 726 gene regions with at least 8 gametes per species (listed in column L in Table S1, Supporting information). The larger *Q. robur* sample after excluding the most introgressed individuals (24 versus 16 gametes in *Q. petraea*) only exhibited slightly higher polymorphism counts than in *Q. petraea* overall (Table 3). Also, 48% and 52% of the polymorphisms observed were exclusive to *Q. petraea* and *Q. robur* respectively in our panel, the rest being shared among species (Table 3). Among exclusive polymorphisms, 46% and 44% were singletons in *Q. petraea* and *Q. robur* respectively, suggesting that they might be either rare in both species, or more polymorphic in local populations from which few individuals were sampled across the species wider ranges. Overall and within both species, we observed a large variation in number of segregating sites per gene size (Fig. S7-A, Supporting information).

The mean nucleotide diversity estimates ($\theta\pi$) across genic regions when considering all polymorphisms were 0.00447 and 0.00425 in *Q. petraea* and *Q. robur* respectively, with up to a 10-fold variation among polymorphic genes overall and in different functional categories (Fig. 4-B and Table 3). When including SNPs only, mean $\theta\pi$ decreased overall by more than 10% (Table 3, and see column D in Table S4, Supporting information). The large variation among genes is also illustrated by the absence of significant differences between mean

diversity among functional categories within species, in most comparisons using non-parametric Wilcoxon rank sum tests (*Wrs*) with similar number of genes. Two notable exceptions were observed when considering all polymorphisms: the *biotic stress* category (358 genes) had on average a lower $\theta\pi$ in *Q. petraea* than in the random gene list (211 genes, *Wrs* Pr<0.042), and the mean $\theta\pi$ of the *reproductive phenology* category was significantly lower in both species than that of the *Bud phenology* category (*Wrs* Pr<0.040 and Pr<0.013 in *Q. petraea* and *Q. robur* respectively, considering exclusive categories from Table S2, Supporting information). Genes with $\theta\pi$ estimates above 0.02 were found across most categories, whether considering all polymorphisms (Fig. 4-B) or SNPs only. The 8 genic regions showing the highest $\theta\pi$ values in both species were annotated for example as disease resistance, transcription factor or membrane transport proteins, half of them being from the original random list.

Mean $\theta\pi$ comparison tests between species across all gene regions were not significant (Table 3, *Wrs* Pr>0.15 for all polymorphisms or SNPs only), nor were they across different categories and between gene pairs, using a 95% confidence interval based on Tajima's evolutionary variance for $\theta\pi$ (Tajima 1983) while assuming underlying Gaussian distributions. Indeed for the same genic regions, many examples can be found of higher $\theta\pi$ estimates in one species or the other. However, comparing diversity estimates across the exact same positions and performing Wilcoxon paired ranked tests (*Wpr*) across all genes, there was a significant pattern of a slightly higher diversity in *Q. petraea* (see Table 3 and Fig. 4-B), whether considering all polymorphisms (*Wpr* Pr<0.028) or SNPs only (*Wpr* Pr<0.036). This pattern remained significant across the 211 genes chosen randomly (*Wpr* Pr<0.037, all polymorphisms), even when excluding the 5% or 10% of genes having the highest $\theta\pi$ values, but it was not robust to considering the other 509 genic regions chosen in functional categories, either together or separately in the different categories (Fig. 4-B), except for the *Abiotic stress* category.

We also observed a very large variation for F_{ST} estimates across gene regions and functional categories, which covered the full range of possible values [0,1], with mean values of ~0.13 whether considering all polymorphisms or SNPs only (Fig. 4-C, and Fig. 4-D for the random genic regions and a representative example of one category). The very few segregating sites with F_{ST} values at 1 had either missing individuals' or strands, possibly caused by polymorphisms within primer regions. Among the sites sequenced for the full sample of gametes, the 20 highest F_{ST} values ranged from 0.6 to 0.9 and belonged to 10 genic regions,

many of which also showed null or very low F_{ST} values within 100 bp. This large variation in differentiation was observed between very close variant sites in many genes, suggesting very high recombination rates at genome-wide and range-wide scales, and consistently with the observed very low LD estimates (see above). Additionally, a large variance is expected around F_{ST} estimates due to the relatively low sample size in both species, in particular for bi-allelic loci (Weir and Hill 2002; Buerkle *et al*. 2011; e.g. Eveno *et al.* 2008).

Discussion

In the NGS era, non-model tree species such as many *Fagaceae* still lag behind model species for easy access to sequence polymorphism data (but see Gugger *et al*. 2016 for *Quercus lobata*). These data are needed for larger scale studies addressing the many diversity issues raised by their combined economic, ecological and conservation interests (Cavender-Bares 2016; Fetter *et al*. 2017; Holliday *et al*. 2017). However, recent achievements and data availability from the *Q. robur* genome sequence project (Plomion *et al*. 2018) opens a large range of applications in many related temperate and tropical *Fagaceae* species due to their conserved synteny (Cannon *et al.* 2018). In this context, we discuss below the representativity of our data in terms of species genomic diversity as well as the robust patterns observed across genes, and further illustrate their past and future usefulness for *Quercus* species.

Genic resources content, quality, and representativity

We provide a high-quality polymorphism catalog based on Sanger resequencing data for more than 800 gene regions covering ~530 kb, using a discovery panel (*DiP*) from mixed *Q. robur* and *Q. petraea* populations located across a large part of their geographic range. This catalog details functional annotations, previous published information, allele types, frequencies and various summary statistics within and across species, which can assist in choosing novel polymorphic sites (SNPs, SSRs, indels…) for genotyping studies (Tables S1 to S5, Supporting information, see also https://github.com/garniergere/Reference.Db.SNPs.Quercus). Among genomic SSRs, more than 90% are new (17 already detected in Durand *et al*. 2010; 3 in Guichoux *et al.* 2011), so they constitute an easy source of potentially polymorphic markers in those white oak species. This catalog corrects and largely extends the SNP list at https://arachne.pierroton.inra.fr/QuercusPortal (see "Quercus petraea / robur" field) which was previously used to document a SNP diversity surrogate for both *Quercus* species in the oak genome first public release (Plomion *et al*. 2016). Standard formats for high-density genotyping arrays and primer information are also provided, making these resources readily

operational for medium scale molecular ecology studies while avoiding the burden of bioinformatics work needed for SNP development.

Thanks to a high quality dedicated pipeline, we could perform a quasi-exhaustive characterization of polymorphism types in our *DiP* and across part of the genic partition of these *Quercus* species (see Fig. 1). Although base call error rates below 1/1000 were used (as originally developed for Sanger sequencing), most variant sites were located in regions with lower error rates (below 1/10000) so that true singletons could be identified. At the genotypic level, a Sanger genotyping error rate below 1% was previously estimated using a preliminary subset of around 1200 SNPs from this catalog (~5800 data points, Lepoittevin *et al.* 2015). This rate can be considered as an upper bound for the present study, given all additional validation and error correction steps performed. Although little used now with the advent of NGS methods, Sanger data have served for genome sequencing projects in tree species before 2010 (Neale *et al*. 2017), and have been instrumental in combination to NGS for BAC clones sequencing, which helped ensuring assembly long-distance contiguity in large genomes such as oaks (Faivre-Rampant *et al*. 2011, Plomion *et al*. 2016). Sanger sequencing has often provided reference high-quality data to estimate false discovery or error rates, and validate putative SNPs in larger scale projects (e.g. Geraldes *et al*. 2011 in *Populus trichocarpa*; Sonah *et al*. 2013 in Soybean; Cao *et al*. 2014 in *Prunus persica*).

Finding an optimal balance between the number of samples and that of loci is critical when aiming to provide accurate estimates of diversity or differentiation in population genetics studies. Given the increasing availability of markers in non-model species (usually SNPs), it has been shown by simulation (Willing *et al*. 2012, Hivert *et al*. 2018) and empirical data (Nazareno *et al*. 2017) that sample sizes as small as 4 to 6 individuals can be sufficient to infer differentiation when a large number of bi-allelic loci (> 1000) are being used. A broad-scale geographic sampling is however required if the aim is to better infer genetic structure and complex demographic scenarios involving recolonization and range shifts due to past glacial cycles, such as those assumed for many European species (Lascoux and Petit 2010, Keller *et al*. 2010, Jeffries *et al*. 2016, Sousa *et al*. 2014). Our sampling design is likely to have targeted a large part of both species overall diversity and differentiation across the resequenced genic regions. This is first suggested by the small proportion of additional polymorphisms once an initial sample of 8 gametes was included for each species (~10% and decreasing as sample size increases, Fig. 4-A and 7-A). Considering the *DiP* within each species, each individual brings on average ~166 new variants (~1% of the total). Second, the

large variance observed across gene nucleotide diversity estimates (see Table 3) is mostly due to stochastic evolutionary factors rather than to sampling effects so unlikely to be impacted by sample sizes over 10 gametes (Tajima 1983). Also, sampling sites are located in regions which include 4 out of the 5 main cpDNA lineages along white oaks recolonization routes (Petit *et al.* 2002), so only the less frequent *D* lineage from South-western Spain might not be represented in our *DiP*. Therefore, if new populations were being sampled within the geographical range considered, they would likely include many of the alleles here observed within species. Indeed, older and more recent reports showed a low genetic structure and high gene flow among distant populations within each species, and a much larger overall differentiation component than the local one (Bodénès *et al.* 1997; Mariette *et al*. 2002; Petit *et al*. 2003; Muir and Schlötterer 2005; Derory *et al.* 2010; Guichoux *et al*. 2013; Gerber *et al*. 2014).

We further tested the frequency spectrum representativity of our *DiP* by comparing genotypes for a subset of 530 SNPs (*sanSNP*) to those obtained using the Illumina infinium array technology (*illuSNP*) for ~70 individuals from Southern France local populations (Lepoittevin *et al*. 2015, using their Table S3). The *illuSNP* excluded SNPs showing compressed clusters (*i.e*. potential paralogs) and those showing a high number of inconsistencies with control genotypes, as recommended by the authors. For SNPs exclusive to one species in *sanSNP*, more than 68% of *illuSNP* showed either the same pattern or one where the alternative allele was at a frequency below 5% in the other species. Less than 8% of those SNPs were common in both species in *illuSNP*. Similarly, for singletons in *sanSNP*, more than two-third showed very low to low frequency (<10%) in *illuSNP*, and only 11% in *Q. petraea* and 9% in *Q. robur* showed a *maf* above 0.25. This confirmed the singletons reality in our *DiP*, and also that some may represent more frequent polymorphisms in local populations. The correlations among *maf* in both datasets were high and significant (0.66 and 0.68 respectively for *Q. petraea* and *Q. robur*, both Pr< 0.0001). These results suggest a small risk of SNP ascertainment bias if the new resources were to be used in populations within the geographic distribution surveyed, in contrast to panels with much less individuals (see Lepoittevin *et al*. 2015 for a discussion on the consequences of such bias in *Quercus* species).

Overall, we obtained sequence data for 0.072% (~530 kb) of the haploid genome of *Q. robur* (size of ~740 Mb in Kremer et al. 2007). We also targeted ~3% of the 25808 gene models described in the oak genome sequencing project (www.oakgenome.fr), and around 1% of the gene space in length. Interestingly, both randomly chosen genic regions and those covering

different functional categories have been mapped across all linkage groups (columns F and X in Table S1, Supporting information). Due to the absence of observed background LD, their diversity patterns can be considered independent. The genes studied represent a large number of categories, as illustrated by very similar distributions for level 2 GO terms to those obtained with the larger *ocv4* assembly (Lesur *et al*. 2015, comparing their Figure 2 to Fig. 4-A to 4-C, Supporting information).

Diversity magnitude and heterogeneity highlight species integrity and introgression patterns

Using a detailed polymorphism typology, we characterized for the first time in two oak species a high proportion of variant positions (30%) that included 1 bp to medium-sized indels and sequence repeats, compared to the more common and commonly reported SNP loci (Table 2). The proportions of indels observed (11.5% of all polymorphisms) is in the range of results available in model tree species (e.g. 13.8% across the genome in *Prunus avium*, Shirasawa *et al*. 2017; 19% in *Prunus persica*, Cao *et al*. 2014; a minimum of 1.4% in *Populus trichocarpa*, Evans *et al*. 2014). Although less abundant than SNPs they represent an important component of nucleotide variation, often have high functional impacts when locating within coding sequences and have been proposed as an easy source of markers for natural populations studies (Väli *et al*. 2008). Larger-sized indels are also likely to be relatively frequent in intergenic regions of the *Quercus* genome and have been linked to transposable elements (TE, see the BAC clones overlapping regions analyses in Plomion *et al*. 2016). Similarly, large indels and copy number variation linked to TE activity were identified as an important component of variation among hybridizing *Populus* species (Pinosio *et al*. 2017). Here when considering variant positions involved in complex polymorphisms, we observed 1 variant position per 48 bp on average within species (resp. 1 per 30 bp in both), compared to the 1 SNP per 68 bp statistic (resp. 1 SNP per 42 bp across both species). Also, some of the SNPs observed were located within complex polymorphic regions that would have been classically filtered out, and nucleotide diversity (π) estimates were higher by 12% when including all polymorphisms (from 0.0038 to 0.0044 if averaging across both species and all genes, Table 3). These nucleotide diversity estimates are provided for the first time in *Q. petraea* and *Q. robur* across a large number genic regions (> 800), compared to previous candidate genes studies across much smaller numbers (< 10) of gene fragments (Kremer *et al.* 2012 in *Q. petraea*; e.g. Homolka *et al*. 2013). Our estimates are consistent with those obtained from genome-wide data and range-wide panels in angiosperm tree species, available mostly from the model genus *Populus* (e.g. *P. trichocarpa*: 1 SNP per 52 bp and π~0.003

across genic regions, Zhou and Holliday 2012, Zhou *et al.* 2014, Evans *et al*. 2014, Wang *et al*. 2016; *P. tremula*: π~0.008, P. tremuloides: π~0.009 across genic regions, Wang *et al*. 2016; π ~0.0026 to 0.0045 in a panel including wild *Prunus persica* accessions, Cao *et al.* 2014). These diversity levels are also within the range estimated for the long-term perennial outcrosser category in Chen *et al*. (2017, see Fig. 1-D with a mean value of silent π close to ~0.005) and can be considered relatively high in the plant kingdom if excluding annual outcrosser estimates or intermediate otherwise. In oaks as in many other tree species with similar life history traits, these levels would be consistent with their longevity, large variance in reproductive success and recolonization or introgression histories, which could have maintained deleterious loads of various origins (Zhang *et al*. 2016, Chen *et al*. 2017, Christe *et al.* 2016b)

Comparing the nucleotide diversity distributions and examining the range of differentiation across genic regions in our *Dip* reveal several robust patterns that altogether illustrate historical introgression patterns among both *Quercus* species, which have long been considered as iconic examples of species exhibiting high levels of gene flow (e.g: Petit *et al*. 2003; Arnold 2006), despite more recent evidence of strong reproductive barriers (Abadie *et al.* 2012). What has been referred to as "strong species integration" seems nevertheless clearer in our *Dip* for *Q. robur* than for *Q. petraea*, according to genetic clustering inference without any *a priori*. Three individuals (27%) considered as typical morphological *Q. petraea* adults (Kremer *et al* 2002) showed significant levels of introgression (Fig. 3). In contrast, only one *Q. robur* based on morphology was introgressed to a level matching the least introgressed *Q. petraea* individual. Discussing species delimitation, Guichoux *et al*. (2013) also showed more robustness in assigning morphological *Q. robur* individuals to their genetic cluster, illustrating an asymmetry in their introgression levels. We note that among our *Dip* individuals, Qs28, one parent from two mapping pedigrees (Bodénès *et al*. 2016) is a clear F1 hybrid among both species (Fig. 3), making those pedigrees two back-crosses instead of one cross within species and one between species.

Moreover, after excluding the four most introgressed individuals, nucleotide diversity in *Q. petraea* was significantly higher (by ~5% on average) than in *Q. robur*. This effect is small, detectable only with Wilcoxon paired ranked tests, mostly across the same ~200 regions sampled randomly and in the *Abiotic stress* category, despite the very large diversity variance across regions, and robust to excluding the highest diversity values. We also sequentially removed the three individuals with the lowest *Q*-values from the *Q. petraea* cluster (Fig. 3),

since they could still harbor residual heterozygosity due to recent back-crossing events and generate the pattern observed. Remarkably, the same significant patterns of higher diversity in *Q. petraea* were observed. Therefore, with 8 to 10 gametes in *Q. petraea* instead of 8 to 24 gametes in *Q. robur*, and with twice less natural stands sampled, the nucleotide diversity in *Q. petraea* was still slightly and significantly higher than in *Q. robur* (Pr<0.011 and Pr<0.026, using all polymorphisms or SNPs only respectively). Although range-wide population structure within species could differentially affect both species global diversity across our *Dip*, published results show that these are very small (~1% across SNPs) and similar (Guichoux *et al.* 2013). The main hypotheses developed so far to explain this difference in species diversity pattern relate to their disparities in life-history strategies for colonizing new stands and associated predictions (Petit *et al*. 2003, Guichoux *et al*. 2013). The colonization dynamics model and patterns observed also assumes very similar effective population sizes in both species, which is a reasonable assumption due to their shared past history and the strong introgression impact at the genomic level. However, given increasing and recent evidence of pervasive effects of different types of selection across genic regions with HT data (e.g. Zhang *et al*. 2016; Christe *et al.* 2016b in *Populus*; Chen *et al* 2017 for long-term perennials), alternative (and non-exclusive) hypotheses worth considering are ones of a higher genome-wide impact of selective constraints in *Q. robur (*Gillespie 2000; Hahn 2008; Cutter and Payseur 2013; Kern and Hahn 2018; e.g. Grivet *et al.* 2017). Since *Q. robur* is the most pioneering species, it has likely been submitted to very strong environmental pressures at the time of stand establishment. Selection might be efficient given oak tree reproductive capacities across a large number of genes involved in abiotic and biotic responses. This would be consistent with significantly lower levels of *He* in *Q. robur* for genes that were specifically enriched for abiotic stress GO terms (Guichoux *et al*. 2013, see their Table S5). Redoing here the same tests across a larger number of independent SNPs (> 1000), whether comparing Nei's unbiaised locus diversity (*He*) or the mean *He* within regions among species, *Q. petraea* systematically showed the same trend of a slightly higher diversity overall, and significantly so only for the *Abiotic stress* category (*Pr*<0.01) and for a similar outlier SNP category (F_{ST} >0.4, mean He>0.15, *Pr*<0.001) than in Guichoux *et al*. (2013). In summary, the absence of the same pattern in any other functional categories might suggest that these are too broad in terms of corresponding biological pathways, hence mixing possible selection signals of opposite effects among species, while we still detect an overall effect due to linked selection on a random set of genes, and on genes involved in abiotic stress. More analyses comparing the nucleotide diversity patterns at genes involved in both species relevant biosynthesis

pathways for ecological preferences (e.g. Porth *et al*. 2005; Le Provost *et al*. 2012, 2016) are clearly needed in replicated populations, for example to estimate the distribution and direction of selection effects and putative fitness impact across polymorphic sites (Stoletzki and EyreWalker 2011), or to study the interplay between different types of selection and variation in local recombination rates on both diversity and differentiation patterns (Payseur and Rieseberg 2016).

A large proportion of shared polymorphic sites (~50% in any species) highlights the close proximity of species at genomic level, consistently with a low mean differentiation across polymorphic sites (F_{ST}~0.13, Fig. 4-C), and despite the very large heterogeneity observed across differentiation estimates. This has now been classically interpreted (and modeled) as reflecting a strong variance in migration and introgression rates, in oaks in particular (Leroy *et al*. 2017), with islands of differentiation assumed to represent regions resistant to introgression. However, interpretations of such patterns remain controversial and multiple processes might be involved and worth exploring further in oaks, such as the effects of heterogeneous selection (both positive and background) at linked loci (Cruickshank and Hahn 2014; Wolf and Ellegren 2017). These effects could be particularly visible in low-recombination regions (Ortiz-Barrientos *et al*. 2016), and would further interact with the mutational and recombination landscapes during the course of speciation (Ortiz-Barrientos and James 2017) and during their complex demographic history.

Applications and usefulness as reference data

During this project, several studies valued part of these resources, hence illustrating their usefulness. For example, good quality homologous sequences were also obtained for ~50 % of the gene fragments in one individual of *Quercus ilex*. This species is more distant to both *Q. petraea* and *Q. robur*, belonging to a different section, so these data guided the choice of nuclear genes for better inferring phylogenetic relationships across 108 oak species (Hubert *et al.* 2014). Bioinformatics tools and candidate genes annotated during the project were also useful to similar genes and SNP discovery approach in *Quercus* or more distant *Fagaceae* species (Rellstab *et al*. 2016, Lalagüe *et al.* 2014 in *Fagus sylvatica*, El Mujtar *et al*. 2014 in *Nothofagus* species). Given the low ascertainment bias and good conversion rate expected within the range surveyed, those genomic resources would be directly applicable to landscape genomics studies at various spatial scales (reviewed in Fetter *et al.* 2017) in both *Quercus* species. Indeed, easy filtering on SNP statistics provided in the catalog would allow distinguishing among different classes of SNPs (e.g. exclusive to each species, common and

shared by both, linked to particular GO functional categories), or delimiting and tracing species in parentage analyses and conservation studies (e.g. Guichoux *et al*. 2013; Blanc-Jolivet *et al.* 2015), or improving estimates of lifetime reproductive success and aiming to understand how demographic history and ecological drivers of selection affect spatial patterns of diversity or isolating barriers (Andrew *et al*. 2013; e.g. Geraldes *et al*. 2014). This type of spatial studies are rare in the target species, usually included a small number of SSR markers, and all suggested complexity in geographical patterns of genetic variation and importance of the ecological context (e.g Neophytou *et al*. 2010; Lagache *et al*. 2014; Klein *et al*. 2017, Beatty *et al.* 2016 for local or regional studies; Muir and Schlötterer 2005; Gerber *et al*. 2014, Porth *et al.* 2016 for range-wide studies). Their power and scope would likely be greatly improved by using medium-scale genotyping dataset including a few thousands SNPs such as those described in our study.

The robust patterns described above of differentiation heterogeneity and consistent differences in diversity magnitude among species call for more studies at both spatial and genomic scales for unraveling these species evolutionary history, in particular regarding the timing, tempo, dynamics and genetic basis of divergence and introgression. Practically in oaks, genome complexity reduction methods such as RAD-seq and similar approaches (e.g. Elshire *et al.* 2011) might be fairly limiting for the research questions mentioned above (Arnold *et al*. 2013; Henning *et al*. 2014; Zhou and Holliday 2012), especially given the large variance in nucleotide diversity and low overall differentiation characterized here. We therefore do not recommend the development of a very large SNP array in oaks since it is likely to be very costly for a minimal return, especially given the very large and range-wide panel that would be needed to significantly limit ascertainment bias (see Lepoittevin *et al*. 2015). Also the very high SNP density required for targeting of functional variants given low overall levels of LD (indicating potentially high recombination rates), would be technically constrained when controlling for genotyping error rates (shown previously to be high). Indeed, these rates would probably be stronger for high diversity, complex, duplicate or multiple copy genic regions (observed in this study Table S1, supporting information, and shown recently to have evolutionary impact on the *Q. robur* genome structure, Plomion *et al.* 2018), preventing them even to be targeted. In contrast, targeted sequence capture (TSC) strategies for resequencing (Jones and Good 2016), although still uncommon in forest tree species evolutionary studies, might be most useful and efficient since they can be oriented towards recovering long genomic fragments and thus allow more powerful site frequency

spectrum and haplotype-based inferences to be pursued (e.g. Zhou *et al.* 2014; Wang *et al*. 2016), at the same time avoiding most of the above technical issues. TSC approaches will surely be encouraged and tailored to specific evolutionary research questions in oaks in the next decade, given the new *Q. robur* genome sequence availability (Plomion *et al*. 2018; Lesur *et al*. 2018 for the first TSC in oaks). However, the bioinformatics pipelines needed for validating haplotype-based or quality data for population genetics inferences also need constant reassessment according to research questions and chosen technology.

We thus propose, in addition to direct applications to landscape genetics (detailed above) and transferability to other *Quercus* species (see Chen *et al*. 2016, and primer information in Table S1, Supporting information), that the high-quality data characterized in this study serve as a reference for such validation purposes. They could not only help for adjusting parameters of the chosen pipelines for data outputs, but also allow estimating genotyping error rates for SNP and more complex classes of variants, either from general patterns comparisons (e.g maf distribution from Tables S3, S4 Supporting information) or using the same control individuals maintained in common garden that could be included in larger-scale studies. Such a reference catalog of SNPs and other types of polymorphisms within gene fragments could also be very useful for solid cross-validation of variants identification, allele frequency and other derived summary statistics in alternative strategies such as *Pool-Seq*, which allow increasing genomic coverage while sampling cost-effectively by pooling individuals (Schlötterer *et al.* 2014). Indeed, the drawback of these approaches, despite dedicated software (PoPoolation2, Kofler *et al*. 2011) is that they can give strongly biaised estimates, or ones that do not consider evolutionary sampling (Hivert *et al*. 2018). Therefore, they require further validation methods which usually value previously developed high-quality and lower-scale data (e.g. Pool-Seq *versus* Sanger and Rad-Seq in Christe *et al*. 2016b; Illumina GA2 *versus* Sanger in Cao *et al*. 2014; EUChip60K *versus* deep-whole genome resequencing in Silva-Junior *et al.* 2015). Finally such a reference dataset would help optimizing the amount of available data from either TSC or whole-genome resequencing experiments in future research challenges.

Data Accessibility

Original assembly used for selecting contigs is in Appendix S2 (Supporting information). For Sanger trace files (with data on at least 2 individuals), see the Dryad repository (doi:10.5061/dryad.h380d51). Consensus sequences are respectively in Supporting information appendices S1 (used to design primers), S5 (used for functional annotation), and S6 (genomic sequences obtained). Tables S1 and S2 correct and extend and the oak Candidate

Genes Database of the Quercus Portal (www.evoltree.eu/index.php/e-recources/databases/candidate-genes). SNPs, indels and SSRs catalogs and positions within genomic consensus sequences, and ready-to-use format for genotyping essays are provided in Tables S3 to S5 (Supporting information). SNP data correct and largely expand part of the SNP database of the Quercus Portal (www.evoltree.eu/index.php/snp-db/, "Quercus petraea / robur" field).

Bioperl scripts from the SeqQual pipeline are given at https://github.com/garniergere/SeqQual, example of parameter files and scripts for STRUCTURE analyses and parsing MREPS software are given at https://github.com/garniergere/Reference.Db.SNPs.Quercus

Acknowledgments

The authors thank Alexis Ducousso, Jean-Marc Louvet, Guy Roussel, Pablo Goicoecha, Hervé le Bouler, Félix Gugerli, Csaba Matyas, Sandor Bordacs, Hans P. Koelewijn, Joukje Buiteveld, Stephen Cavers, Bernd Degen and Jutta Buschbom for choosing trees and providing dried leaves of individuals from various Intensive Study Populations of previous European projects populations. We are grateful to H. Lalagüe, G. Vendramin, I. Scotti, and L. Brousseau for testing earlier scripts of SeqQual and to I. Lesur for help in using the *ocv4* oak resources. The sequencing work was funded by the EVOLTREE network of Excellence (EU contract n°016322). TL post-doc fellowship was funded by the ANR TRANSBIODIV (06-BDIV-003-04) and LINKTREE (contract n°2008-966). TD salary was funded by the ANR REALTIME (N°59000256). Computing facilities of the Mésocentre de calcul Intensif Aquitain des Universités de Bordeaux, de Pau et des Pays de l'Adour are thanked for providing computer time for this study. We also thank Rémy Petit for funding part of TL fellowship and support in developing SeqQual tools. PA received a Ph.D. grant (2009-2011) from the « Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche » of France, and additional funding from EVOLTREE.

References

Abadie P, Roussel G, Dencausse B, *et al.* (2012) Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *Journal of Evolutionary Biology*, **25**, 157-173.

Abbott RJ, James JK, Milne RI, Gillies ACM (2003) Plant introductions, hybridization and gene flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **358**, 1123–1132.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215,** 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25,** 3389-3402.

Andrew RL, Bernatchez L, Bonin A *et al*. (2013) A road map for molecular ecology. *Molecular Ecology*, **22**, 2605–2626.

Arnold ML (2006) Evolution through genetic exchange. Oxford University Press, Oxford.

Beatty GE, Montgommery 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.

Blanc-Jolivet C, Liesebach M (2015) Tracing the origin and species identity of *Quercus robur* and *Quercus petraea* in Europe: a review. *Silvae Genetica* **64(4)**, 182–193.Bodénès C, Labbe T, Pradère S, Kremer A (1997) General vs. local differentiation between two closely related white oak species. *Molecular Ecology***, 6**: 713-724.

Bodénès C, Chancerel E, Ehrenmann F, Kremer A, Plomion C (2016) High-density linkage mapping and distribution of segregation distorsion regions in the oak genome. *DNA Research*, **23**, 115-124.

Bodénès C, Chancerel E, Gailing O, *et al.* (2012) Comparative mapping in the Fagaceae and beyond with EST-SSRs. *BMC Plant Biology*, **12**, 153.

Bodénès C, Chancerel E, Murat F, *et al*. (2012) Comparative mapping in the Fagaceae and beyond using EST-SSRs. *BMC Plant Biology*, 12, 153.

Branca A, Paape TD, Zhou P, *et al.* (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA*, **108**, E864–E870.

Brewer S, Cheddadi R, De Beaulieu JL, Reille M, Data contributors (2002) The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management***, 156**, 27–48.

Brousseau L, Tinaut A, Duret C, *et al.* (2014) High-throughput transcriptome sequencing and preliminary functional analysis in four neotropical tree species. *BMC Genomics*, **15**, 238.

Buerkle CA, Gompert Z, Parchman TL (2011) The n=1 constraint in population genomics. *Molecular Ecology*, **20**, 1575–1581.

Cannon CH, Brendel O, Deng M *et al*. (2018) Gaining a global perspective on *Fagaceae* genomic diversification and adaptation. *New Phytologist*, *218*, 894-897.

Cao K, Zheng Z, Wang L *et al. (*2014*)* Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biology*, **15**, 415.

Casasoli M, Derory J, Morera-Dutrey C, *et al*. (2006) Comparison of QTLs for adaptive traits between oak and chestnut based on an EST consensus map. *Genetics*, **172**, 533–546.

Cavender-Bares J (2016) Diversity, distributions, and ecosystem services of the North-American oaks. *International oaks*, **27**, 37-48.

Chen J, Glémin S, Lascoux M (2017) Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, **34**, 1417–1428.

Chen J, Zeng Y-F, Liao W-J *et al*. (2016) A novel set of single-copy nuclear gene markers in white oak and implications for species delimitation. *Tree Genetics & Genomes*, **13**, 50.

Christe C, Stolting KN, Bresadola L, *et al*. (2016a) Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. *Molecular Ecology*, **25**, 2482–2498.

Christe C, Stölting KN, Paris M, *et al*. (2016b) Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Molecular Ecology*, **26**, 59-76.

Conesa A, Götz S, Garcia-Gomez JM, *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21(18)**, 3674–3676.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: massive introgression by local genes. *Evolution*, **62**, 1908–1920.

Curtu AL, Gailing O, Finkeldey R (2007) Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. *BMC Evolutionary Biology*, **7**, 218.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

Derory J, Scotti-Saintagne C, Bertocchi E, *et al.* (2010) Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. *Heredity*, **104**, 438-448.

Durand J, Bodénès C, Chancerel E, *et al*. (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics*, **11**, 570.

El Mujtar VA, Gallo LA, Lang T, Garnier-Gere P (2014) Development of genomic resources for *Nothofagus* species using next-generation sequencing data. *Molecular Ecology Resources*, **14**, 1281–1295.

Elshire RJ, Glaubitz JC, Sun Q, *et al*. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, ***6*(5),** e19379. doi:10.1371/journal.pone.0019379

Evans LM, Slavov GT, Rodgers-Melnick E, *et al*. (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, **46**, 1089–1096

Evans LM, Slavov GT, Rodgers-Melnick E, *et al*. (2014) Population genomics of Populus Trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, **46**, 1089-1096.

Eveno E, Collada C, Guevara MA, *et al*. (2008) Contrasting patterns of selection at Pinus pinaster Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25**: 417-437.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, **8**, 175–185.

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

Excoffier L. 2007. Analysis of population subdivision. Pages 980-1020 in Handbook of Statistical Genetics. 3rd ed. DJ Balding, M. Bishop, and C. Cannings, ed. Wiley, Chichester, West Sussex, UK.

Faivre-Rampant P, Lesur I, Boussardon C *et al*. (2011) Analysis of BAC end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome. *BMC Genomics*, **12**, 292.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fetter KC, Gugger PF, Keller SR (2017) Landscape Genomics of Angiosperm Trees: From historic Roots to Discovering New Branches of Adaptive Evolution. In Groover A. and Cronk Q. (eds) *Comparative and evolutionary genomics of angiosperm trees, Plant Genetics and Genomics: Crops and Models*. New York, Springer.

Geraldes A, Farzaneh N, Grassa CJ, *et al.* (2014) Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, **68**, 3260–80.

Geraldes A, Pang J, Thiessen N, *et al.* (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources*, **11** (Suppl. 1), 81–92.

Gerber S, Chadœuf J, Gugerli F *et al*. (2014) High rates of gene flow by pollen and seed in oak populations across Europe. *PLoS ONE*, **9**, e85130.

Gilbert KJ, Andrew RL, Bock DG *et al.* (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular ecology*, **21**, 4925–4930.

Gillespie JH (2000) Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics*, **155**, 909–919.

Grivet D, Avia K, Vaattovaara A, Eckert AJ, Neale DB, Savolainen O, Gonzalez-Martınez SC. 2017. High rate of adaptive evolution in two widespread European pines. *Molecular Ecology*, **26**, 6857–6870.

Grivet D, Deguilloux M-F, Petit RJ, Sork VL (2006) Contrasting patterns of historical colonization in white oaks (*Quercus* spp.) in California and Europe. *Molecular Ecology* **15**, 4085–93.

Gugger PF, Cokus SJ, Sork VL (2016) Association of transcriptome-wide sequence variation with climate gradients in valley oak (*Quercus lobata*). *Tree Genetics and Genomes*, **12**, 15.

Guichoux E, Garnier-Géré P, Lagache L *et al.* (2013) Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, **22**, 450–462.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplex (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.) *Molecular Ecology Resources*, **11**, 578–585.

Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution 62(2):255–265.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD marker for dense linkage mapping. *Molecular Ecology*, **23**, 5224–5240.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913.

Hillier L, Green P (1991) OSP: a computer program for choosing PCR and DNA sequencing primers. *PCR Methods and Applications*; **1,** 124–128.

Hivert V, Leblois R, Petit EJ, Gautier M, Vitalis R (2018) Measuring genetic differentiation from Pool-seq data. bioRXiv doi: https://doi.org/10.1101.282400.

Holliday JA, Aitken SN, Cooke JEK, *et al*. (2017) Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. *Molecular Ecology*, **26**, 706–717.

Homolka A, Schueler S, Burg K, Fluch S, Kremer A (2013) Insights into drought adaptation of two European oak species revealed by nucleotide diversity of candidate genes. *Tree Genetics & Genomes*, **9,** 1179-1192.

Hubert F, Grimm GW, Jousselin E, *et al.* (2014) Multiple nuclear genes stabilize the phylogenetic backbone of the genus Quercus. *Systematics and Biodiversity*, **12**, 405–423.

Jeffries DL, Copp GH, Lawson Handley L, *et al.* (2016) Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, **25**, 2997–3018.

Jensen J, Larsen A, Nielsen LR, Cottrell J (2009) Hybridization between *Quercus robur* and *Q. Petraea* in a mixed oak stand in Denmark. *Annals of Forest Science*, **66**, 706.

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**,185–202.

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Keller SR, Olson MS, Silim S *et al.* (2010) Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*. *Molecular Ecology*, 19, 1212–1226.

Kern AD, Hahn MW (2018) The neutral theory in light of natural selection. *Molecular Biology and Evolution*, **35**, 1366-1371.

Klein EK, Lagache-Navarro L, Petit RJ (2017) Demographic and spatial determinants of hybridization rate. *Journal of Ecology*, **105**, 29–38.

Kofler R, Pandey RV, Schlotterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.

Kolpakov R, Bana G, Kucherov G (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acid Research*, **31**, 3672–3678.

Kremer A, Abbott A, Carlson J, *et al.* (2012) Genomics of Fagaceae. *Tree Genetics & Genomes*, **8**, 583–610.

Kremer A, Casasoli M, Barreneche T, *et al*. (2007) Fagaceae. In: Genome Mapping and Molecular Breeding in Plants (ed. Kole CR), Vol 7 *Forest Trees*, pp. 165–187. Springer, Heidelberg, Berlin, New York, Tokyo.

Kremer A, Dupouey JL, Deans JD, *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**, 777–787.

Lagache L, Klein EK, Ducousso A, Petit RJ (2014) Distinct male reproductive strategies in two closely related oak species. *Molecular Ecology*, **23**, 4331–4343.

Lalagüe H, Csilléry K, Oddou-Muratorio S, Safrana J, de Quattro C, Fady B, Gonzalez-Martinez SC, Vendramin GG (2014) Nucleotide diversity and linkage disequilibrium at 58 stress response and phenology candidate genes in a European beech (*Fagus sylvatica*) population from southeastern France. *Molecular Ecology* **23**, 4696-4708.

Lascoux M, Petit RJ (2010) The'New Wave'in plant demographic inference: more loci and more individuals. *Molecular Ecology*, **19**, 1075–1078.

Le Provost G, Lesur I, Lalanne C *et al.* (2016) Implication of the suberin pathway in adaptation to waterlogging and hypertrophied lenticels formation in pedunculate oak (*Quercus robur* L.). *Tree Physiology*, *36*, 1330–1342.

Le Provost G, Sulmon C, Frigerio JM, *et al.* (2012) Role of waterlogging-responsive genes in shaping interspecific differentiation between two sympatric oak species. *Tree Physiology*, **32**, 119–134.

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, *et al*. (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology*, **18**, 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.

Lepoittevin C, Bodénès C, Chancerel E, *et al*. (2015) Single-nucleotide polymorphism discovery and validation in high density SNP array for genetic analysis in European white oaks. *Molecular Ecology Resources*, **15**, 1446–1459.

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

Lesur I, Alexandre H, Boury C, *et al*. (2018) Development of target sequence capture and estimation of genomic relatedness in a mixed oak stand. *Frontiers in Plant Science (METHODS)*, doi: 10.3389/fpls.2018.00996.

Lesur I, Le Provost G, Bento P, *et al*. (2015) The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. *BMC Genomics*, **16**, 112.

Mariette S, Cottrell J, Csaikl UM, Goikoechea P, Nig A, Lowe AJ, *et al.* (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 Genet.* **51**: 72-79.

Mishra B, Gupta DK, Pfenniger M, *et al*. (2018) A reference genome of the European beech (*Fagus sylvatica* L.) *GigaScience*, **7**:6. https://doi.org/10.1093/gigascience/giy063.

Muir G, Fleming CC, Schlötterer C (2000) Species status of hybridizing oaks. *Nature*, **405**, 1016.

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

Nazareno A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Molecular Ecology Resources*, **17**, 1136–1147.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, **12**, 111–122.

Neale DB, Langley CH, Salzberg SL, Wegrzyn JL (2013) Open access to tree genomes: the path to a better forest. *Genome Biology*, **14**: 120.

Neale DB, Martínez-García PJ, La Torre De AR, Montanari S, Wei X-X (2017) Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Reviews of Plant Biology*, **68**, 457–483.

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

Nei M. (1977) F-statistics and analysis of gene diversity in sub-divided populations. *Annals of Human Genetics*, 41, 225–233.

Neophytou C, Gärtner SM, Vargas-Gaete R, Michiels H-G (2015) Genetic variation of Central European oaks: shaped by evolutionary factors and human intervention? *Tree Genetics & Genomes*, **11**, 79.

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.

Ortiz-Barrientos D, Baack EJ (2014) Species integrity in trees. *Molecular Ecology*, **23**, 4188-4191.

Ortiz-Barrientos D, Engelstädter J, Rieseberg LH (2016) Recombination rate evolution and the origin of species. *Trends in Ecology and Evolution*, **31**, 226–236.

Ortiz-Barrientos D, James ME (2017) Evolution of recombination rates and the genomic landscape of speciation. *Journal of Evolutionary Biology*, **30**,

Parent GJ, Raherison E, Sena J, Mackay JJ (2015) Forest tree genomics: review of progress. In Plomion C, Adam-Blondonpp A-F (eds) Land Plants - Trees. *Advances in Botanical Research*, **74**, 39–92, London: Academic Press, Elsevier.

Payseur BA, Rieseberg LH (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, **25**, 2337– 2360.

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

Petit RJ, Carlson J, Curtu AL, *et al.* (2013) Fagaceae trees as models to integrate ecology, evolution and genomics. *New Phytologist*, **197**, 369–371.

Petit RJ, Csaikl UM, Bordács S, *et al*. (2002) 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.

Pina-Martins JB, Pappas G, Paulo OS (2018) New insights on adaptation and population structure of cork oak using genotyping by sequencing. bioRxiv http://dx.doi.org/10.1101/263160.

Pinosio S, Giacomello S, Faivre-Rampant P, *et al*. (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution*, **33,** 2706–2719.

Plomion C, Aury J-M, Amselem J, *et al*. (2016) Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Molecular Ecology Resources*, **16**, 254–265.

Plomion C, Aury JM, Amselem J, *et al*. (2018) Oak genome reveals facets of long lifespan. *Nature Plants*, **4**, 440–452.

Porth I, Garnier-Géré P, Klapste J, Scotti-Saintagne, El-Kassaby YA, Burg K, Kremer A (2016) Species-specific alleles at a *β*-tubulin gene show significant association with leaf morphological variation within *Quercus petraea* and *Q. robur* populations. Tree Genetics & Genomes **12**: 81.

Porth I, Koch M, Berenyi M, *et al*. (2005) Identification of adaptation-specific differences in mRNA expression of sessile and pedunculate oak based on osmotic-stress induced genes. *Tree Physiology*, **25**, 1317–1329.

Prunier, J., Caron, S., & MacKay, J. (2017). CNVs into the wild: Screening the genomes of conifer trees (*Picea* spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics*, **18**, 97.

Ramos AM, Usié A, Barbosa P, **et al**. (2018) The draft genome sequence of cork oak. Scientific Data, 5, 180069, http://dx.doi.org/10.1038/sdata.2018.69

Rellstab, C., Zoller, S., Walthert, L., *et al*. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present and future climatic conditions. Molecular Ecology, 25, 5907-5924.

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Roux C, Fraïsse C, Castric V, Vekemans X, Pogson GH, Bierne N (2014) Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone. *Journal of Evolutionary Biology*, 27, 1662-1675.

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

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nature Reviews Genetics, 15, 749–763.

Shirasawa K, Isuzugawa K, Ikenaga M, *et al*. (2017) The genome sequence of sweet cherry (*Prunus avium)* for use in genomics-assisted breeding. *DNA Research*, **24(5),** 499-508.

Silva-Junior OB, Faria DA, Grattapaglia (2015) A flexible multi-species 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytologist* **206**, 1527-1540.

Sonah H, Bastien M, Iquira E, *et al.* (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8(1), e54603.

Sork VL, Fitz-Gibbon ST, Puiu D, *et al*. 2016 First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). *G3: Genes Genomes Genetics,* **6 (11),** 3485–3495.

Sousa **VC**, Peischl S, Excoffier L (2014) Impact of range expansions on current human genomic diversity. *Current Opinion in Genetics and Development*, **29**, 22-30

Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Molecular Biology and Evolution*, **28**, 63–70.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Tajima F (1993) Measurement of DNA polymorphism. In: *Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology*, edited by Takahata, N. and Clark, A.G., Tokyo, Sunderland, MA: Japan Scientific Societies Press, Sinauer Associates, Inc., p. 37-59.

Tarkka MT, Herrmann S, Wubet T, *et al*. (2013) OakContigDF159.1, a reference library for studying differential gene expression in *Quercus robur* during controlled biotic interactions: use for quantitative transcriptomic profiling of oak roots in ectomycorrhizal symbiosis. *New Phytologist*, **199**, 529–540.

Tine M, Kuhl H, Gagnaire P-A, *et al*. (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications, **5**, 5770.

Tuskan GA, DiFazio S, Jansson S, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, **313(5793)**:1596–1604.

Ueno S, Klopp C, Leplé JC, *et al.* (2013) Transcriptional profiling of bud dormancy induction and release in oak by next-generation sequencing. *BMC Genomics*, **14**, 236.

Ueno S, Le Provost G, Leger V, *et al.* (2010) Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics*, **11**, 650.

Valbuena-Carabana M, González-Martínez S, Sork V, *et al.* (2005) Gene flow and hybridisation in a mixed oak forest (*Quercus pyrenaica* Willd. and *Quercus petraea* (Matts.) Liebl.) in central Spain. *Heredity*, **95**, 457–465.

Väli U, Brandström M, Johansson M, Ellegren H (2008) Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, **9**, 8.

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics*, **30**, 418–426. https://doi.org/10.1016/j.tig.2014.07.001

Verde I, Abbot GA, Scalabrin S, *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, **45**, 487–494.

Verde I, Jenkins J, Dondini L, *et al.* (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, **18**, 1-18.

Wang J, Street NR, Scofield DG, Ingvarsson PK (2016) Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, **202**, 1185-1200.

Warr A, Robert C, Hume D, et al. (2015) Exome sequencing : Current and Future perspectives. *Genes, Genomes, Genetics*, **5**, 1543-1550.

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

Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721–750.

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. PLoS ONE, **7**, e42649.

Wolf JB, Ellegren H (2017) Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, **18**, 87–100.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology* **14**, 851–865.

Zanetto A, Roussel G, Kremer A (1994) Geographic variation of inter- specific differentiation between *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. *Forest Genetics*, **1**, 111-123.

Zhang M, Zhou L, Bawa R, Suren H, Holliday JA (2016) Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, **33**, 2899–2910.

Zhou L, Bawa R, Holliday JA. 2014. Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (Populus trichocarpa). *Molecular Ecology*, **23**, 2486–2499.

Zhou L, Holliday JA (2012) Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, **13**, 703.

Author contributions

Funding acquisition: AK, PGG, CP, and MLDL; Initial conception and individuals sampling: PGG, AK, CP, MPR, VL; Bioinformatics strategy and experimental design: PGG, TL; DNA extraction and quality check: VL; Sequence Data acquisition: PGG, CP, TL, VL; Individuals' identification checks for quality control VL, CL, PL; Pilot study: VL, PGG; Working assembly: JMF, PGG: Primer design and amplicon choice: PGG, VL, TD; Original candidate gene lists choice: PGG, TL, JMF, CP, AK, TD, CR, MLDL, GLP, ChB, EG, CaB, NT, PA; Bioinformatics tools: TL and PGG (SeqQual pipeline), JMF and AF (Bioperl and R scripts), PA, CL, VelM**,** JT**,** FH, TD (SeqQual tests), FR (website); Visual Chromatogram checks, SNP/assembly validations: PGG, VL, TD, PA, TL, MLDL, CaB, ChB, CL, CR and EG; Bioinformatic and population genetic analyses: PGG, TL, SM, ChB: Functional annotation: TL, PGG, VelM, PA; Manuscript draft: PGG; Manuscript review and edition: PGG, SM, CL, ChB, TL; all authors agreed on the manuscript.

Tables

Table 1 Geographic location of 25 sampled individuals from *Quercus petraea*, *Q. robur* and *Q. ilex.*

Country	Sampling site	Latitude	Longitude	Morphological *Quercus* species	Original Identifier
Spain	Arlaban	42.967	-2.55	*petraea*	Ar18
				robur	Ar22
France	Arcachon	44.663	-1.181	*robur*	A4*
	Pierroton	44.737	-0.776	*ilex*	IL_C
				robur	11P*
				robur	3P*
	Orléans	47.826	1.908	*petraea*	Qs21*
				petraea	Qs28*
				petraea	Qs29*
	Petite Charnie	48.083	-0.167	*petraea*	PC55
				robur	PC229
				robur	PC233
Switzerland	Büren	47.105	7.383	*petraea*	B3
				robur	B179
Hungary	Sopron	47.717	16.642	*petraea*	S444
				robur	S104
The Netherlands	Meinweg	51.181	6.138	*petraea*	M51
				robur	M7
United Kingdom (UK)	Roudsea Wood	54.218	-3.018	*petraea*	RW108
				robur	RW8
				robur	RW11
Germany	Rantzau	53.707	9.765	*petraea*	R100
				petraea	R127
				robur	R300
				robur	R312

Latitude and longitude are given in the WGS 84 coordinate system. Coordinates correspond either to a central point in the mixed forest stand, or the mean of individual trees coordinates. *: parents of controlled crosses used for genetic mapping. *Quercus* species *a priori* assigned from morphological information by persons who sampled the trees.

Table 2 Typology of polymorphisms in successfully resequenced amplicons.

	Both species and introgressed individuals	*Q. petraea*	*Q. robur*	*Q. ilex*
Total length resequenced (bp)	529281	-	-	196676
Number (Nb) of amplicons	986	-	-	486
Nb of genic regions	852	-	-	394
Mean genic region size-N50 size (bp)	621-700	-	-	500-539
Minimum-Maximum genic region size (bp)	81-2009	-	-	198-1285
Estimated intron sequences (bp)	186827	-	-	-
Mean haploid sample size (total sequence)	34.71	13.35	18.28	-
Polymorphism in 852 genic regions				
Mean haploid sample size (variants)	32.16	12.57	13.85	-
Monomorphic genic regions	15 (1.76%)	18 (2.14%)	21 (2.52%)	-
Genes with at least one single base indel	591	345	379	-
" " " " one larger indel (>1 bp)	252	190	214	-
" " " " one SSR (>=di)	163	-	-	-
SNPs only (excluding 1 bp indels)	12478	7511	8078	-
Indels (1 bp)	1213	751	809	-
Indels (2-5 bp)	221	142	161	-
Indels (6-10 bp)	88	72	71	-
Indels (11-50 bp, excl. SSRs)	98	81	79	-
Indels (74,146,219,341 bp, excl. SSRs)	4	3	4	-
Total number of polymorphisms	**14102**	**8560**	**9202**	**676**
Triallelic SNPs	*218*	*141*	*165*	-
...Singletons (incl. 1 bp indels)	*4334*	*1990*	*2151*	-
...Variable SSRs (excl. homopolymers)	*111*	-	-	-
Total length with sequence variant positions	17594	10765	11451	-
Sequence length of indels and complex polymorphisms (Indels and SSRs)	5116	-	-	-

Counts for *Q. petraea* exclude the 2 most introgressed individuals (Qs28 and S444 in Table 1); SSR: simple sequence repeats; "N50 size" is the size for which the cumulative sum of gene amplicons' size equal or higher than this value corresponds to 50% of the total amplicons' size sum; The number of polymorphisms for *Q. ilex* equals the number of heterozygotes in the resequenced individual across amplicons; Numbers of monomorphic regions were computed for those with at least 10 gametes in both species; Some detected SSR patterns were not polymorphic in our samples (detailed in Tables S1 and S5, supporting information).

Table 3 Polymorphism counts and nucleotide diversity in parental species across genic regions with larger sample sizes.

Polymorphism in 726 gene fragments	both species	*Q. petraea*	*Q. robur*
Number of individuals considered	20	8	12
Monomorphic gene fragments	17 (2.34%)	19 (2.63%)	20 (2.87%)
Total number of polymorphisms	11089	7061	7721
SNPs only	9867	6226	6830
All Indels and SSRs	1222	835	891
Exclusive polymorphisms	-	3359	4024
Singletons among them (%)	-	0.456	0.437
Shared polymorphisms	3696	-	-
Mean nucleotide diversity estimates			
SNPs only	3.849E-03	**3.957E-03**	**3.740E-03**
" " diversity range		0-0.03823	0-0.02525
Tajima's evolutionary standard deviation	2.549E-03	2.632E-03	2.465E-03
SNPs only (509 chosen genes)	3.752E-03	3.821E-03	3.682E-03
SNPs only (202 random genes)	4.103E-03	**4.306E-03**	**3.900E-03**
All polymorphisms	4.359E-03	**4.471E-03**	**4.247E-03**
" " diversity range		0-0.03893	0-0.02525
Tajima's evolutionary standard deviation	2.816E-03	2.903E-03	2.729E-03
All polymorphisms (509 chosen genes)	4.214E-03	4.278E-03	4.150E-03
All polymorphisms (202 random genes)	4.716E-03	**4.944E-03**	**4.488E-03**

The 4 most introgressed individuals from Fig. 3 are excluded for computations. Monomorphic regions are defined as in Table 2. Diversity are computed for regions with a minimum of 200 bp overall and at least 8 gametes per species at variant positions. The 509 chosen genes belong to the different functional categories listed in Table S1. Values in the "both species" column for nucleotide diversity estimates are means across each species values.

Figures Legends

Figure 1 Bioinformatics strategy for sequence data production, amplicon assembly, functional annotation, and polymorphism discovery. Scripts used are in italics (see text for further details). GO: Gene Ontology, EC: Enzyme Commission ID. * A **good** sequence is defined as having a minimum of 50% of its nucleotides with a Phred score above 30.

Figure 2 Distributions of GO terms across different gene lists (*bud*, *abiotic* and *biotic*) at Biological Level 3, and Fisher exact tests across pairs of sequence clusters with the same GO terms between the random list and other lists. Significance levels *: P<0.05, **: P<0.01.

Figure 3 Posterior assignment probabilities of individuals into two optimal clusters based on STRUCTURE analyses. Probabilities are sorted in increasing order of belonging to cluster 2 (here *Q. robur* (Qr, in blue/dark grey), the alternative cluster 1 matching *Q. petraea* (Qp, in green/light grey), apart from individuals with higher introgression levels. Each bar represents one individual and includes mean upper and lower bounds of 90% Bayesian confidence intervals around mean *Q*-values across 10 replicates. Each replicate is a different random sample of 1785 polymorphisms.

Figure 4 Mean proportion of new variant sites brought by each new distinct individual added to all possible initial sample size combinations (-A); Mean nucleotide diversity (considering all polymorphisms) in both species across genic regions, and different functional categories (-B) compared between species with Wilcoxon signed-rank tests: significant at Pr<5% (*), non-significant (ns); Histogram of *Fst* estimates across polymorphic gene regions with a minimum of 8 gametes per species, after excluding singletons and grouping negative with null values (-C); Manhattan plot of *Fst* estimates sorted by mean *Fst* values across randomly chosen (black dots) and Bud phenology genic regions (grey dots) (-D).

Fig. 1

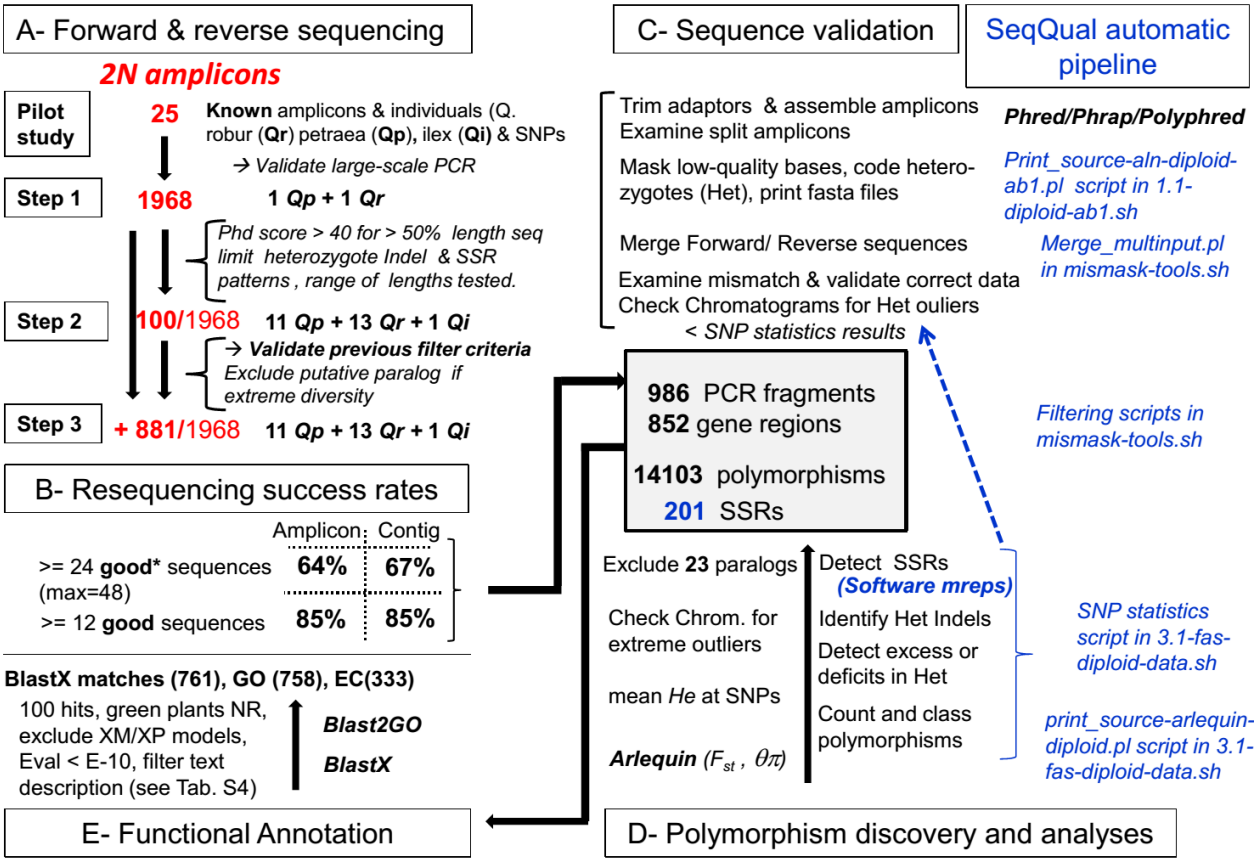


Fig. 2

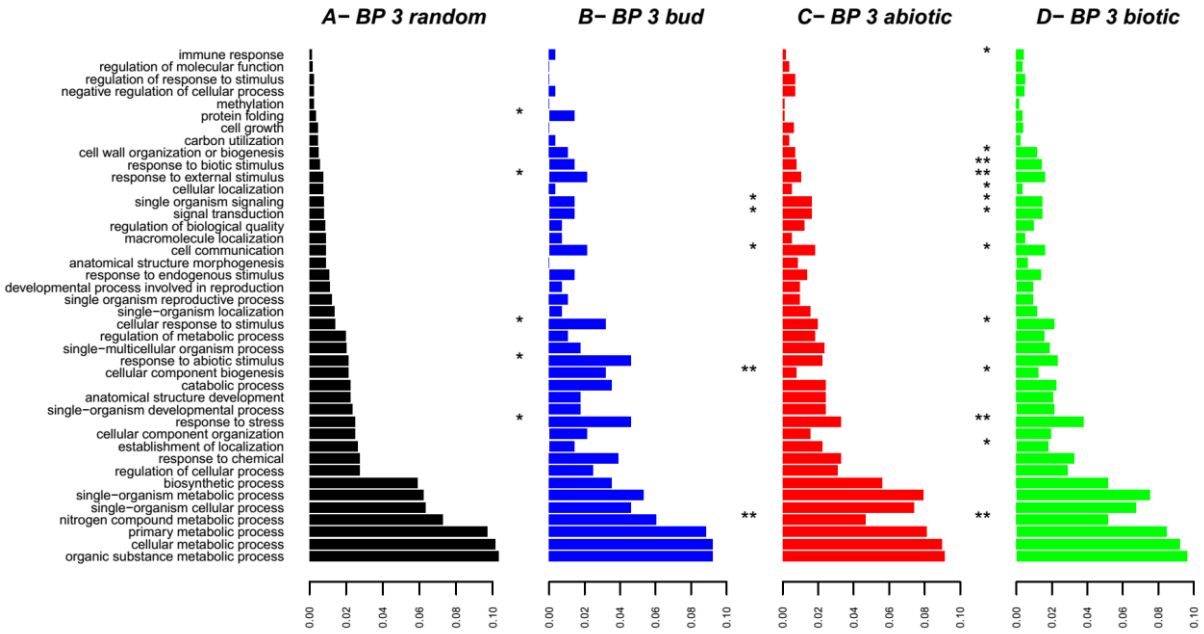


Fig. 3

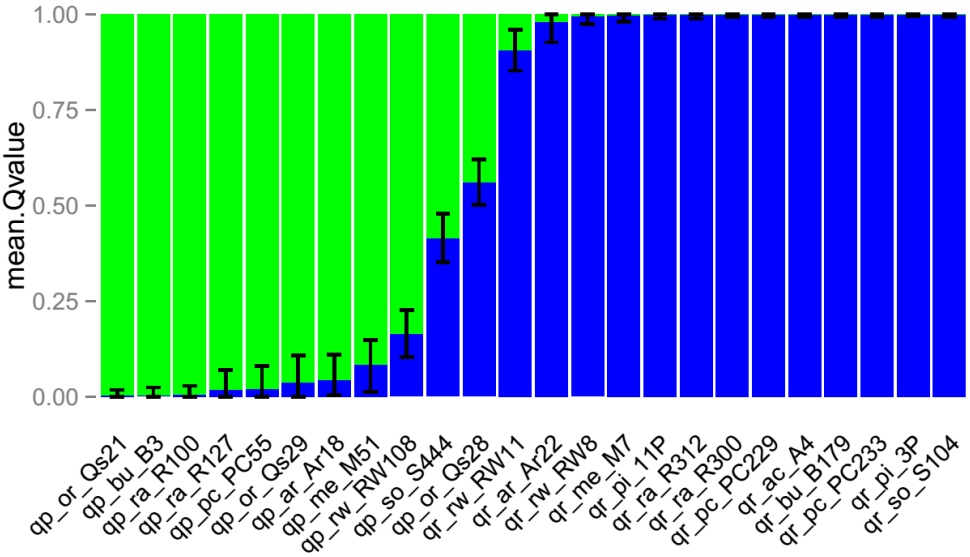
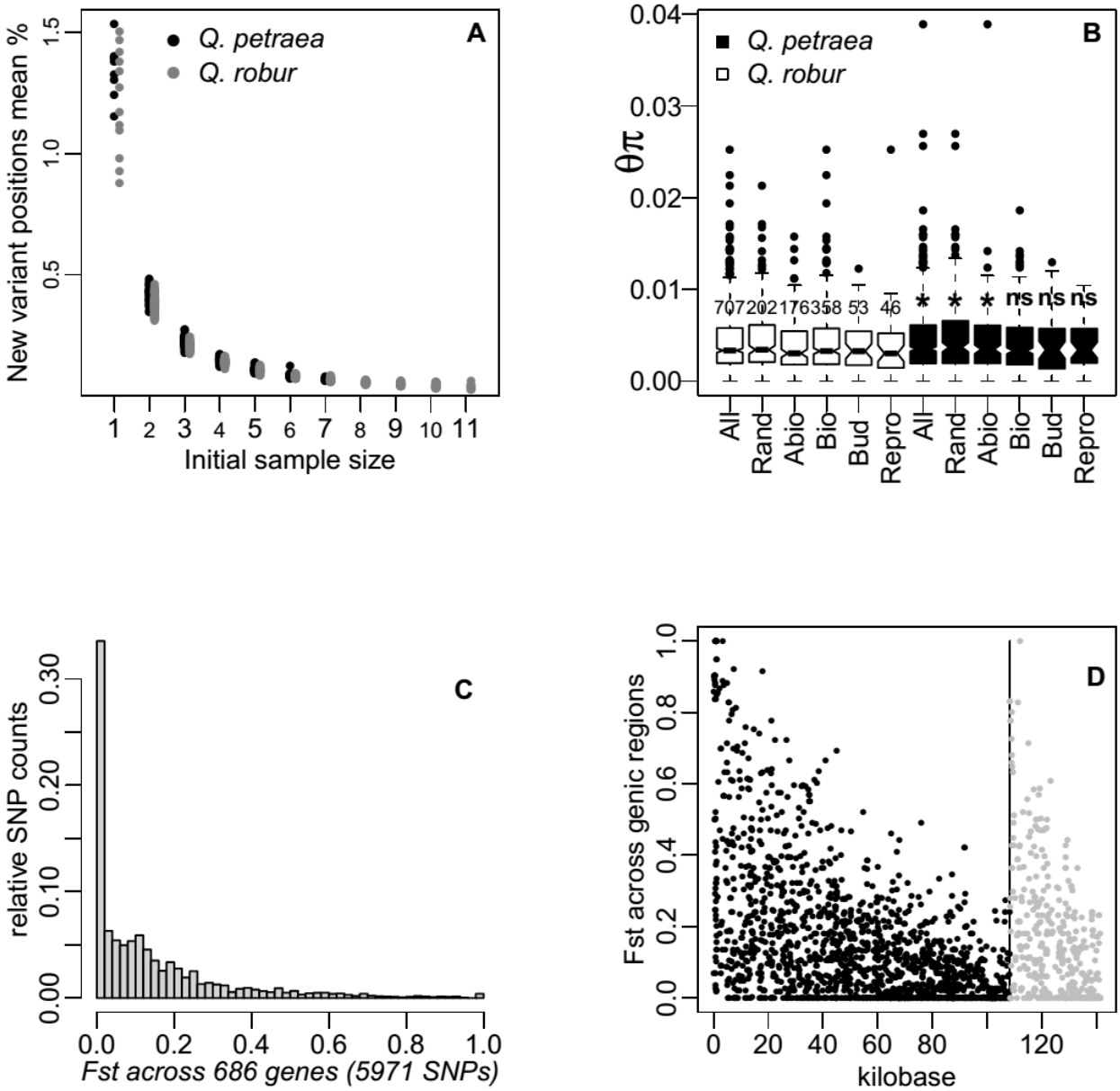


Fig. 4



Supporting Information

Fig. S1 Sampling site locations within the natural geographic distribution of *Q. petraea* and *Q. robur*. Vector map is from *http://www.naturalearthdata.com* and distribution areas from Euforgen (*http://www.euforgen.org/distribution-maps/)*

Fig. S2 Working assembly steps and softwares (A), and bioinformatic strategy for search of candidate genes and amplicon choice (B).

Fig. S3 Plots of the ΔK values from the Evanno *et al*. (2005) method (S3-A, -B, -C, -D, -E), and of the mean values of the estimated probability ln (of the data given K) with standard deviations for K ranging from 1 to 5 (S3-F to S3-J), which show support for K=2. Plots are from the STRUCTURE HARVESTER program.

Fig. S4 Distributions of Gene Ontology (GO) terms for the consensus sequences in Appendix S3, at level 2 (-A, -B, -C) and level 3 (-D, -E, -F): A- and D- for Biological Process, B- and E- for Molecular Function, C- and F- for Cellular Component. Annotation rules: E-value<10^{-30}, annotation cut-off 70, GO weight 5, HSP coverage cutoff 33%. Filtering applies for at least 5 sequences and a node score of 5 per GO term (but see rare exceptions in Table S2).

Fig. S5 Distributions of GO terms across different gene lists (*bud*, *abiotic* and *biotic*) at Biological Level 2, and Fisher exact tests across pairs of sequence clusters with the same GO terms between the random list and other lists. Significance levels *: P<0.05.

Fig. S6 A-J Posterior assignment probabilities (*Q*-values) of 24 individuals attributed to 2 clusters (STRUCTURE analysis) for different numbers of polymorphisms, different sampling of SNP data, and different plots of credible intervals.

Fig. S7 Number of high-quality variant positions per 100 base pair (bp) across 852 gene fragments ranked by their length (bp), overall and for each species (-A); Mean number of new variants brought by each new distinct individual added to all possible initial sample size combinations (-B).

Table S1 Description of amplicons: primer sequences, original candidate gene list, targeted biological functions (see references), candidate gene type, fragment expected size and position in original assembly, preliminary results based nucleotide quality for obtained sequences and final decision after excluding paralog amplifications.

Table S2 Functional annotation results from Blast2GO (-A), comparison of BlastX best hits results (according to *E-values*) between consensus sequences of the *orict* working assembly and the *ocv4* assembly (-B), and comparisons of BlastN results of both *orict* and corresponding amplicon only to *ocv4* (-C).

Table S3 Description of all variants single base positions, with sample sizes, alleles, genotypes counts, various statistics, and generic format for genotyping essays input data. Species samples exclude the 2 most introgressed individuals.

Table S4 Description of all polymorphisms as in Table S3, with additional sequence motifs, contiguous base positions and length for complex ones (derived from Table S3, and see also Table S5 for SSR positions).

Table S5 SSR patterns as detected from the *mreps* software.

Appendix S1 Additional method details.

Appendix S2 Contigs of the original working assembly used for selecting candidate gene regions and design amplicon primers, including consensus sequences and reads where nucleotides with Phred score below 20 have been masked.

Appendix S3 Sequences of chosen contigs consensus and singletons sequences for functional annotation analyses.

Appendix S4 Consensus sequences of 852 genomic regions obtained in this study for *Quercus petraea* and *Q. Robur* individuals. "$(N)^{9}$" : represents a low-quality fragment of a length below ~1 kb separating Forward and Reverse amplicons; "n" position with a majority of nucleotides with phd score below 30. "$(-)^{x}$": insertion is a minor allele at that position.

Appendix S5 Nucleotide sequence data of 394 gene regions for one *Quercus ilex* individual, heterozygote sites being indicated by IUPAC codes.

Appendix S6 Outputs from Blast2GO analyses**.**