MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 450–462 doi: 10.1111/mec.12125

Outlier loci highlight the direction of introgression in oaks

E. GUICHOUX,*†‡ P. GARNIER-GÉRÉ,*†[1] L. LAGACHE,*† T. LANG,*†§ C. BOURY*† and R. J. PETIT*†[1]

*INRA, UMR1202 BIOGECO, Cestas, F-33610, France, †Univ. Bordeaux, UMR1202 BIOGECO, Talence, F-33400, France, ‡Pernod Ricard Research Center, Créteil, F-94000, France, §Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan, 666303, China

Abstract

Loci considered to be under selection are generally avoided in attempts to infer past demographic processes as they do not fit neutral model assumptions. However, opportunities to better reconstruct some aspects of past demography might thus be missed. Here we examined genetic differentiation between two sympatric European oak species with contrasting ecological dynamics (*Quercus robur* and *Quercus petraea*) with both outlier (i.e. loci possibly affected by divergent selection between species or by hitchhiking effects with genomic regions under selection) and nonoutlier loci. We sampled 855 individuals in six mixed forests in France and genotyped them with a set of 262 SNPs enriched with markers showing high interspecific differentiation, resulting in accurate species delimitation. We identified between 13 and 74 interspecific outlier loci, depending on the coalescent simulation models and parameters used. Greater genetic diversity was predicted in *Q. petraea* (a late-successional species) than in *Q. robur* (an early successional species) as introgression should theoretically occur predominantly from the resident species to the invading species. Remarkably, this prediction was verified with outlier loci but not with nonoutlier loci. We suggest that the lower effective interspecific gene flow at loci showing high interspecific divergence has better preserved the signal of past asymmetric introgression towards *Q. petraea* caused by the species' contrasting dynamics. Using markers under selection to reconstruct past demographic processes could therefore have broader potential than generally recognized.

Keywords: asymmetric introgression, divergent selection, gene flow barriers, genetic assignment, outlier loci, *Quercus petraea*, *Quercus robur*, single nucleotide polymorphism

Received 15 March 2012; revision received 29 September 2012; accepted 4 October 2012

Introduction

In population genetics analyses, loci considered to be under selection are typically discarded in attempts to infer past demographic processes (Beaumont 2005; Helyar *et al.* 2011). The rationale for removing these loci from such analyses is that locus-specific effects caused by selection will bias population genetic inferences that traditionally assume selective neutrality (e.g. Wright 1931; Hudson 1990; Wakeley & Hey 1997; Nielsen & Wakeley 2001). Yet, markers known to be under selection have been used to estimate dispersal in two specific situations: genetic clines along environmental gradients and 'tension' hybrid zones (e.g. Barton & Hewitt 1981; Mallet *et al.* 1990; Szymura & Barton 1991; Lenormand *et al.* 1998). More studies are needed to evaluate the potential utility of genes under selection to reconstruct historical patterns of gene flow in other situations.

In an island model of migration at equilibrium, there is an inverse nonlinear relationship between genetic

Correspondence: Pauline Garnier-Géré, Fax: 33 5 57 12 28 81; E-mail: pauline@pierroton.inra.fr
Rémy J. Petit, Fax: 33 5 57 12 28 81; E-mail: petit@pierroton.inra.fr
[1]These authors contributed equally to this work.

structure and gene flow (Wright 1931). Thus, when gene flow is high, differences in allelic frequencies among populations become very low and genotyping or sampling errors become relatively more important (Waples 1998). In contrast, when gene flow is low, differences in allele frequencies among populations should be large, thereby facilitating the characterization of genetic structure. As selection can reduce effective gene flow and increase divergence (Bengtsson 1985), loci influenced by selection could provide more precise indications of genetic structure than others (Nosil *et al.* 2009). Such loci could be particularly helpful for assessing relative differences in levels of gene flow, especially in high gene flow species (see Appendix S1, Supporting information for a numerical example).

Targeting loci under divergent selection or tightly linked with them could be particularly relevant for reconstructing the main direction of gene flow. If gene flow is asymmetric between two populations, we expect that the population receiving more immigrants will be more variable and harbour more private alleles than the other population (e.g. Quintana-Murci *et al.* 2008; Marsden *et al.* 2011). However, if overall gene flow is high, differences in levels of diversity or in allele frequencies among populations might be slight and error-prone (Waples 1998; Neigel 2002). In contrast, the signature of asymmetric gene flow should be strong at loci under divergent selection.

A prerequisite for testing the potential of selected loci for such purposes is to accurately identify them. Loci showing high allelic-frequency divergence, which are possibly affected by selection in the corresponding genomic region, are typically detected with F_{ST}-based outlier methods (Beaumont & Nichols 1996; Beaumont 2005; Foll & Gaggiotti 2006; Excoffier *et al.* 2009). These methods can identify relatively highly differentiated markers (so-called outlier loci) in comparison to expected levels under neutrality inferred from coalescent simulations (Luikart *et al.* 2003; Li *et al.* 2012). They are increasingly used to study nonmodel species and speciation processes (Butlin 2008; Nosil *et al.* 2009; Garvin *et al.* 2010; Helyar *et al.* 2011).

In this study, we decided to focus on gene flow between closely related plant species rather than between conspecific populations, as divergent selection should be high in this case (Nosil *et al.* 2009). Moreover, interspecific gene flow is often asymmetric in plants (Arnold 1997; Abbott *et al.* 2003). This asymmetry can be caused by differences in fertilization success and offspring survival (Tiffin *et al.* 2001; Lowry *et al.* 2008), differences in abundance at the time of mating (Lepais *et al.* 2009) or differences in population dynamics (Currat *et al.* 2008). We selected a pair of partly interfertile white oak species, pedunculate oak (*Quercus robur*) and sessile oak (*Quercus petraea*), which are widely distributed over Europe and have overlapping distribution ranges (the range of *Q. petraea* being largely included within that of *Q. robur*). These two species are patchily distributed as a function of the environment, resulting in numerous contact zones where hybridization can take place, forming so-called mosaic hybrid zones (Streiff *et al.* 1999; Jensen *et al.* 2009). Despite evidence of hybridization and introgression, *Q. robur* and *Q. petraea* remain ecologically and morphologically differentiated (Kremer *et al.* 2002) and have strong postpollination prezygotic sexual barriers, as revealed by a recent large-scale interspecific crossing study (Abadie *et al.* 2012).

Another important prerequisite for our study was to accurately delimit these two closely related interfertile oak species, which has been a long-lasting goal for botanists and geneticists (Cousens 1963; Carlisle & Brown 1965; Bodénès *et al.* 1997; Muir *et al.* 2000; Coart *et al.* 2002; Kremer *et al.* 2002; Scotti-Saintagne *et al.* 2004; Kelleher *et al.* 2005). Encouraging results have been obtained recently by selecting some of the most discriminating microsatellites identified to date (Guichoux *et al.* 2011). However, greater discriminatory power might be obtained by focusing on F_{ST}-based outlier loci showing high interspecific divergence. The objective is then to use these loci to test hypotheses regarding past demographic events that emerge from considerations of oaks' life histories.

Quercus petraea, a shade-tolerant species, must typically follow the more pioneering oak species, *Q. robur*, during forest successions, as it probably did during the postglacial recolonization of Europe (Petit *et al.* 2003). Thus, there is a phase where immigrant *Q. petraea* trees have to establish in areas already dominated by *Q. robur*. Under such conditions, introgression is expected to be strongly asymmetric towards the late invader, according to neutral models of colonization dynamics (Currat *et al.* 2008). On one hand, alleles from the resident species (*Q. robur*) that leak into the genome of the colonizing species (*Q. petraea*) can rapidly increase in frequency at the time of expansion, resulting in high introgression in the expanding species (*Q. petraea*). On the other hand, less introgression is expected towards *Q. robur*, the resident species, which is already at carrying capacity. Asymmetric introgression would also be consistent with the finding that *Q. robur* female flowers are more easily fertilized by *Q. petraea* pollen than the converse in artificial crosses (Steinhoff 1993). Consequently, late-successional *Q. petraea* should have greater genetic diversity than the early successional *Q. robur*, at least if there are similar initial levels of diversity in the two species. However, if interspecific genetic exchanges are not exceedingly rare,

as might be inferred from previous studies of the species (Streiff *et al.* 1999; Jensen *et al.* 2009; Lepais *et al.* 2009; Lagache *et al.* 2012), the asymmetry signal might be weak or absent. Under such conditions, highly divergent genes that have experienced reduced effective interspecific gene flow might be particularly useful for detecting the signal of ancient asymmetric introgression.

The objective of this work was to use outlier loci to test if the direction of introgression matches predictions from the demo-genetic models described above, thus demonstrating their utility to study demographic processes. The two prerequisites of this study were to accurately identify *Q. robur* and *Q. petraea* purebreds (and remove admixed individuals) and to identify outlier loci. For these purposes, we applied a model-based outlier detection method to a set of single nucleotide polymorphisms (SNPs) enriched with markers showing high differentiation between species in a discovery panel. We compared the ability of outlier SNPs and non-outlier SNPs to delimitate species using existing methods. We then tested for differences in the genetic diversity and structure of the two species using both types of markers, to check if they are consistent with a signature of ancient asymmetric introgression.

Materials and methods

Material

We sampled 855 oak trees in six mixed stands of *Quercus robur* and *Quercus petraea* in northern France (Petite Charnie, Vitrimont, Charmes, Lure, Cuve, Mondon, see Appendix S2, Supporting information for the populations' geographic locations and sample sizes, and Appendix S3, Supporting information for the species' distributions in Europe). One stand (Petite Charnie) includes 278 adult trees and 380 offspring (in 51 half-sib families, see Guichoux *et al.* 2011). Leaves or buds were sampled and stored immediately at −20 °C or in silica gel.

DNA isolation

DNA was isolated from leaves or buds using an Invisorb DNA plant HTS 96 kit (Invitek, Berlin, Germany), following the manufacturer's instructions, except for the lysis step (1 h at 65 °C). DNA quality was estimated by separating the samples in 1% (w/v) agarose gel then staining with bromophenol blue. The DNA concentration in the samples was evaluated using an Infinite 200 microplate reader (Tecan, Männedorf, Switzerland) in conjunction with a Quant-it dsDNA Broad-Range Assay kit (Invitrogen, Carlsbad, CA, USA). The concentration of each sample was then adjusted to 50 ng/μL by a STARlet 8-channel robot (Hamilton, Reno, NV, USA).

SNP choice

Most of the SNPs were chosen from a larger set of 9080 SNPs that had been previously validated by allelic resequencing of 584 gene fragments within the framework of the EVOLTREE network of excellence activities (http://www.evoltree.eu/; SNPs available via the *Quercus* Portal at https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=snp). We selected a subset of 346 validated polymorphic SNPs (Phred Score > 30) from the resequencing study, by applying both technical and biological criteria, using an automatic pipeline developed for these data. In particular, we enriched the list with SNPs expected to better differentiate the species (either from high interspecific differentiation estimates in a small panel of individuals, or from their location in genes putatively involved in drought stress tolerance, a trait that differentiates the two species; see Appendices S4 and S5 for details on functional categories). The aim was also to maximize the number of genes by targeting few SNPs per gene. In addition, 32 SNPs from functional and expressional candidate genes not included in the previous resequencing study were identified by *in silico* analysis (Appendices S4 and S5). In the final list of 384 SNPs (Appendix S5, Supporting information), all markers except those derived *in silico* met stringent technical criteria (successful amplification for at least 2/3 of the sampled individuals in each species, Illumina scores above 0.6, and at least 60-bp spacing between SNPs within genes). These SNPs represent 227 different genes (15 genes for the 32 *in silico* SNPs) with on average 1.7 SNPs per gene.

SNP genotyping

The required SNP format for online submission to the Illumina Assay Design Tool (ADT; Illumina Inc., San Diego, CA, USA) was prepared with a Perl script adapted from Lepoittevin *et al.* (2010), which predicts design feasibility. SNP genotyping was performed with the 384-plex GoldenGate assay (Illumina Inc.) based on the VeraCode technology. We followed the manufacturer's instructions, using 250 ng of DNA as starting quantity for each sample. Three negative controls were added to each batch of the five 96-well plates. The acquired data were analysed (i.e. SNPs were clustered for genotypic class calls) using BeadStudio (Illumina Inc.) according to recommended procedures (Close *et al.* 2009; Lepoittevin *et al.* 2010), except that we initially retained SNPs lacking one homozygote cluster and those showing cluster compression, that is, members of

genotypic classes that were closer to each other than expected on a normalized 0–1 scale in cluster plots. In the Petite Charnie stand, SNP data were validated using parent/progeny relationships determined using microsatellite (SSR) data (Guichoux *et al.* 2011). This allowed *a posteriori* validation of all SNPs, even in cases of cluster compression. Monomorphic loci and loci in total linkage disequilibrium with another locus were discarded from subsequent analyses.

Assignment methods for accurate species delimitation

We used the Bayesian clustering algorithm implemented in STRUCTURE 2.3.3 (Pritchard *et al.* 2000) to classify individual SNP genotypes and compared the results with those for SSR genotypes previously reported (Guichoux *et al.* 2011). After a burn-in of 50 000 steps followed by 50 000 Markov chain Monte Carlo repetitions, we calculated average assignment scores over 10 runs with K (number of groups) set to two, corresponding to the two species. A key step in any such analysis is to choose appropriate threshold values for the assignment scores to identify purebred individuals efficiently (Vähä & Primmer 2006). Purebreds have expected admixture levels of 0 and 1, F1 hybrids of 0.5, and backcrosses of 0.25 and 0.75. Thus, threshold values of 0.125 and 0.875 are optimal for distinguishing between purebreds and first-generation backcrosses, which was deemed sufficient for this study, even though later-generation backcrosses certainly occur in this system. To confirm the relevance of these thresholds under the simplifying assumption that the examined populations consist solely of purebreds, F1s and first-generation backcrosses, we simulated with HYBRIDLAB 1.0 (Nielsen *et al.* 2006) 1000 genotypes for each of the following categories: purebreds (2), F1s and first-generation backcrosses (2). Allelic frequencies of purebreds were used as reference and observed assignment scores were compared to theoretical expectations.

We also tested the repeatability of the assignment scores by performing further clustering analyses using only half of the validated SNPs, grouped into two independent subsets (designated A and B) randomly drawn from the complete set. Finally, we tested the ability of varying numbers of SNPs to assign purebred individuals. For this purpose, we created SNP subsets (2, 4, 8, 16, 32, 64, 128, 256 and all SNPs), with each subset comprising the SNPs with the highest possible interspecific F_{ST}. We then compared the STRUCTURE clustering results for each of these subsets on the basis of a performance index, defined as the efficiency multiplied by the accuracy, as in Vähä & Primmer (2006). Efficiency is 'the proportion of individuals in a category that are correctly identified' (e.g. *Q. robur* identification efficiency = the number of individuals in the *Q. robur* group that are correctly assigned divided by the total number of *Q. robur* individuals, including those falsely assigned to other groups). Accuracy is 'the proportion of an identified group that truly belongs to that category' (e.g. *Q. robur* identification accuracy = the number of individuals in the *Q. robur* group that are correctly assigned divided by the total number of individuals in the *Q. robur* group, including those falsely assigned to the group).

Diversity analyses and outlier detection method

For each species, allelic frequencies, genotypic frequencies, expected heterozygosity (H_e; Nei 1973) and inbreeding coefficients (F_{IS}; Weir & Cockerham 1984) were estimated for each SNP and their average across loci was computed. Only individuals with multilocus genotypes having <10% of missing data were included. Intra- and interspecific F_{ST} estimates (Wright 1951) were computed using ARLEQUIN 3.5.1.2 (Excoffier *et al.* 2009).

The main objective was to contrast diversity patterns between outlier loci and nonoutlier loci. We searched for outlier loci, that is, loci showing higher levels of interspecific genetic differentiation than expected under neutrality, by using the coalescent simulation module implemented in ARLEQUIN, which extends the Beaumont & Nichols method (1996) to a finite number of demes in the symmetrical island migration model and to a variable mutation rate across loci (Excoffier *et al.* 2009).

A main issue was to choose a mean targeted F_{ST} value for the simulations (hereafter called reference F_{ST} value). For that, the ideal would be to have randomly chosen SNPs available across the genome, preferably far away from the influence of coding regions, so that they could be considered to be mostly affected by demographic effects. Unfortunately, such markers are usually not available in nonmodel species. Therefore, the mean F_{ST} is often used as initial reference value, assuming no selection effects overall when using a large number of random markers. In our case, the markers included a large proportion of highly differentiated SNPs and SNPs from candidate genes of ecologically divergent traits among species. Given this choice, using the mean F_{ST} value as reference would assuredly overestimate the 5% quantile of the simulated distribution (Helyar *et al.* 2011). The number of outliers detected with such a reference would therefore be underestimated, which would be very conservative.

To account for the uncertainty in the reference F_{ST} value in oaks, we followed two different approaches for outlier detection: one using a reference F_{ST} value of 0.04, which is based on a multilocus scan from different markers in the same species (Scotti-Saintagne *et al.*

2004), and the very conservative approach described previously, which uses the observed mean F_{ST} value (0.22) as reference. We also derived the neutral envelope differently to the default ARLEQUIN option to better account for our particular case study: first by choosing a trial subset of SNPs with a mean F_{ST} value equal to the reference value; second by adjusting this reference value so that the bias in the mean simulated F_{ST} value for two demes only is accounted for (see Slatkin 1991); finally by retaining only genealogies with one mutation to model SNPs (See Appendix S4, Supporting information for more details on how we ran the outlier tests). We further explored the robustness of outlier detection in our data in more complex situations (Excoffier *et al.* 2009), by testing a hierarchical model with the two species demes each composed of six populations. We also tested for the presence of outliers within each species using as reference intraspecific F_{ST} values the mean observed values (0.012 for *Q. robur* and 0.013 for *Q. petraea*, based on data for all 262 SNPs, see Table 1). In all cases, the null F_{ST} distribution was built as a function of H_{WP}, the mean within-deme heterozygosity value, and observed values were tested as outliers in comparison with the 95th percentile of the simulated distributions.

Graphical comparison of genotype likelihoods

The data set was analysed with the genotype-likelihood approach of Paetkau *et al.* (1995) and Waser & Strobeck (1998), which allows direct, convenient visualization of genetic differences between individuals of two groups. We plotted two likelihoods for each genotype corresponding to their probabilities of generation based on *Q. robur* and *Q. petraea* allelic frequencies, respectively, in the form of biplots. To compute these likelihoods, allele frequencies at each locus in each 'pure' species are first computed. Then, the genotypic likelihood at each locus is estimated as the square of the observed allele frequency for homozygotes or twice the product of the two allele frequencies for heterozygotes, and likelihoods are multiplied across loci assuming that they are independent (Paetkau *et al.* 2004) to yield an overall likelihood. As genotype likelihoods are products across loci, their values are geometrically affected by the number of SNPs included in the computation, so only individuals with nearly complete multilocus genotypes were considered. We compared results obtained using four sets of loci (12 SSRs, all SNPs, nonoutlier SNPs and outlier SNPs) with species and admixed categories previously defined on the basis of all validated SNPs. Genotype likelihoods were computed with GENALEX 6.4 (Peakall & Smouse 2006). For each category, we also plotted the coordinates of the mean likelihood value of all individuals belonging to that category.

Table 1 Genetic parameters for the two oak species (*Quercus robur*, *Quercus petraea*) based on all SNPs

Group	*N*	H_e	F_{IS}	Intraspecific F_{ST}
Q. robur	436	0.221	−0.004	0.012
Q. petraea	329	0.217	0.001	0.013

N, sample size; H_e, mean expected heterozygosity across individuals; F_{IS}, mean value across individuals (within populations and then across them) of fixation indices.

Results

SNP genotyping

A total of 855 individuals of the two species (*Quercus petraea* and *Quercus robur*) were genotyped at 384 SNPs. After all validation steps, 262 out of 384 SNPs were retained for further analyses (68%). We excluded in particular six SNPs that were in complete linkage disequilibrium with another locus. The parent-pair analyses further led to the exclusion of 24 SNPs that did not segregate according to Mendelian expectations, including 11 SNPs that had compressed clusters (16% of this category) and 13 SNPs that had uncompressed clusters (6% of this category; Appendix S7, Supporting information). The retained SNPs all had inconsistency rates lower than 5% in parent-pair analyses. Overall, this validation procedure increased the final success rate by 13% compared with recommended procedures (Close *et al.* 2009; Lepoittevin *et al.* 2010) and decreased the error rate of the selected SNPs.

Species assignment

With the chosen clustering thresholds (0.125 and 0.875, see Materials and methods), we assigned each individual to one of the following categories: (i) 'purebreds' (which should include mostly 'pure' *Q. robur* or *Q. petraea* individuals and, if present, some second and later-generation backcrosses) and (ii) 'admixed individuals'. Assignment results based on the 262 retained loci revealed a low proportion of admixed trees (9%), about half the estimate based on 12 SSRs (17%, see Appendix S8, Supporting information). The stability of assignment values was very high for purebreds when comparing the two subsets of 131 SNPs (97% correspondence, see Fig. 1). Assignment scores for purebreds obtained from the SNP analysis were also very similar to those obtained using SSRs (95% correspondence), despite the lower number of SSR loci (12). In contrast,

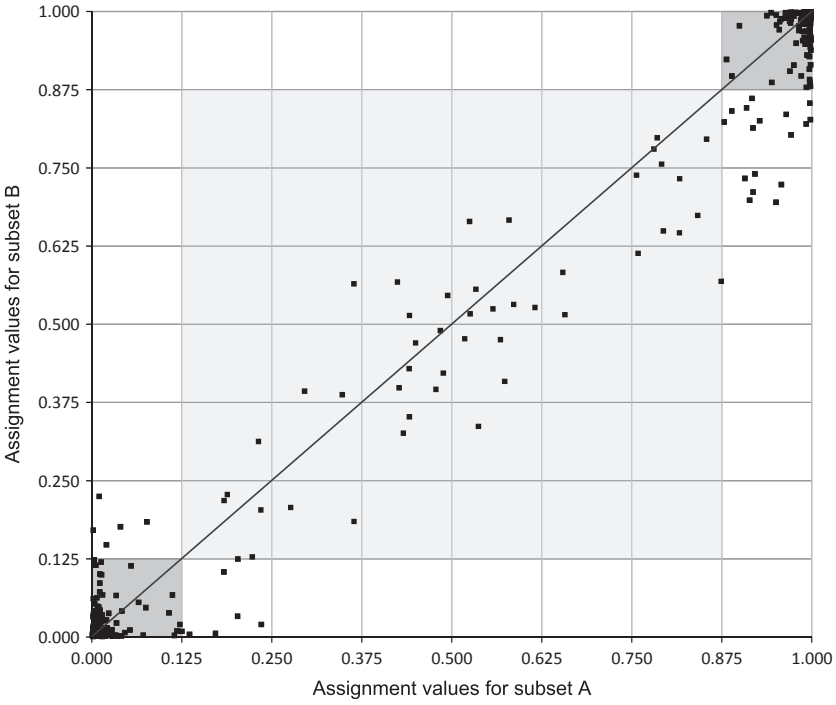

Fig. 1 Correspondence between assignment scores for each individual genotyped with two randomly drawn subsets (A and B) of 131 SNPs from all 262 SNPs. Points close to the diagonal represent individuals with repeatable assignments based on the two subsets. The thresholds for the categories are provided in the text. On the basis of these thresholds, purebred individuals with repeatable assignment scores are found inside the two small grey squares.

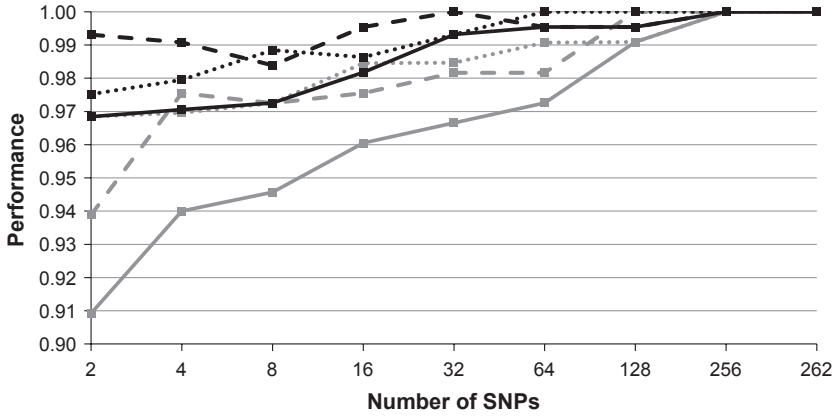

Fig. 2 Efficiency, accuracy and overall performance of assignments for *Quercus robur* (black line) and *Quercus petraea* (grey line) individuals. Efficiency (dashed line), accuracy (dotted lines) and performance (full lines) as functions of the number of SNPs ordered by decreasing interspecific F_{ST} values.

assignments for the admixed category were less stable across the two subsets of 131 SNPs (66% correspondence, Fig. 1), indicating that assignment is less precise in this group. When using few SNPs showing the highest interspecific F_{ST}, assignment performance (Vähä & Primmer 2006) remained high for both species (Fig. 2). Interestingly, the performance for assigning *Q. robur* individuals was always higher than for assigning *Q. petraea* individuals, regardless of the number of SNPs used, due to a better efficiency and accuracy (Fig. 2). Therefore, *Q. robur* individuals require genotyping at fewer SNPs than *Q. petraea* for equally robust assignment.

We also compared assignment values of simulated genotypes with expectations. The results show that all genotypic classes were clearly separated with few incorrect assignments (see Appendix S9, Supporting information).

Genetic structure and outlier detection

The mean expected heterozygosity (H_e) across loci was similar for the two species (0.221 for *Q. robur* and 0.217 for *Q. petraea*, see Table 1). Mean F_{IS} values across loci were very close to zero and did not differ significantly between the species ($P = 0.7$, see Table 1). Within each species, a large number of loci (>90%) were at Hardy–Weinberg equilibrium and all loci included in the analyses were at linkage equilibrium, as required by the initial assumptions of both the STRUCTURE clustering and

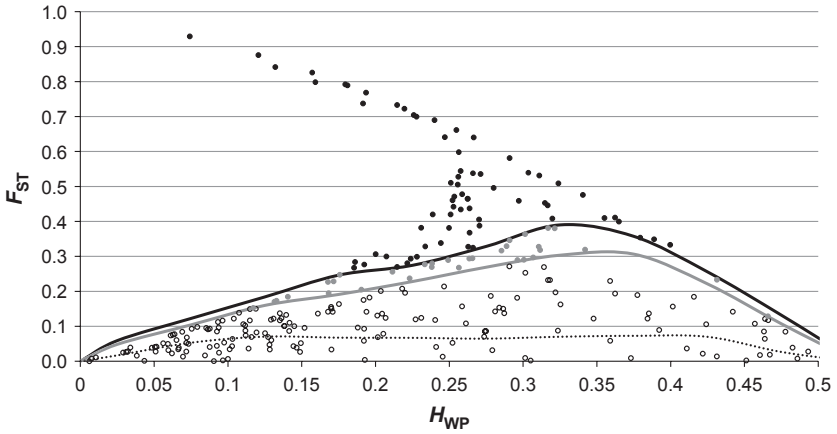

Fig. 3 Distribution of interspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), calculated using a hierarchical island model (two demes and six populations within each deme, reference F_{ST} = 0.04). Outlier and nonoutlier loci are represented by filled and open circles, respectively. Loci represented by grey circles were excluded from the nonoutlier category. Black line, 95th quantile distribution; grey line, 90th quantile distribution; dotted line, 50th quantile distribution.

the genotype-likelihood descriptive methods. As expected given our choice of SNPs, the mean interspecific F_{ST} across loci was much higher (0.22) than previously published estimates (Scotti-Saintagne *et al.* 2004), with some SNPs showing very high values (up to 0.93, see Fig. 3).

The distribution of observed F_{ST} values as a function of mean within-species diversity has a remarkable croissant shape (Fig. 3). We interpret this as a mathematical artefact caused by the fact that F_{ST} at each diallelic locus is constrained to vary within some limits that depend on the minor allele frequency (and thus on diversity) and on the number of populations (e.g. Petit *et al.* 1995; Hedrick 2005). With a reference F_{ST} value of 0.04, the proportions of outlier loci detected (i.e. located above the 95th percentile of the simulated distribution) when using the two demes only or the hierarchical island model were similar (74 and 68 outliers out of 262, respectively). We focus on the latter model in the following as it was slightly more conservative (Fig. 3 and Appendix S5, Supporting information). Moreover, 28 loci located between the 90th and 95th percentiles were excluded from the nonoutlier category, as suggested by Nosil *et al.* (2009). At the intraspecific level, the proportion of outliers detected were 5% (13) outliers in *Q. petraea* and 4% (10) in *Q. robur*, see Appendices S11 and S12. These values are very close to the type I error rate (5%). Eight of these intraspecific outliers were also interspecific outliers and were excluded from subsequent analyses to facilitate interpretations. Thus, a total of 60 interspecific outliers (24%) and 166 nonoutliers (65%) were finally considered. Mean estimates of interspecific F_{ST} were 0.093 for nonoutlier loci, 0.504 for outlier loci and 0.210 across all loci. The levels of genetic differentiation computed among populations within each species did not differ significantly between interspecific outliers and nonoutliers (see intraspecific F_{ST} values in Table 2). As expected, using an initial reference F_{ST} value of 0.22 instead of 0.04 resulted in a much lower proportion of

Table 2 Comparison of genetic diversity and inter- and intraspecific differentiation at outlier and nonoutlier loci, with two different reference F_{ST} values (0.04 and 0.22)

				H_e					Intraspecific F_{ST}		
Reference F_{ST}	Type of loci	*N*	F_{ST}	Total	Mean within species	*Quercus robur*	*Quercus petraea*	*P*†	*Quercus robur*	*Quercus petraea*	*P*
0.04	Outliers	60‡	0.504	0.511	0.255	0.163	0.347	***	0.010	0.010	ns
	Nonoutliers	166	0.093	0.401	0.201	0.228	0.173	**	0.013	0.015	ns
	P		***	***		*	***		ns	ns	
0.22	Outliers	13	0.756	0.393	0.197	0.094	0.299	***	0.003	0.012	ns
	Nonoutliers	247	0.191	0.440	0.220	0.228	0.213	ns	0.013	0.013	ns
	P		***	ns		***	***		***	ns	

N, sample size; H_e, mean expected heterozygosity across loci.

†The significance of differences, obtained from Student *t*-tests, in values between the species (ns, not significant; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

‡For calculating intraspecific F_{ST} values the eight intraspecific outliers were considered, to enable comparison with nonoutliers.

outlier loci (13 interspecific outliers only, out of 262 SNPs).

Genotype likelihoods and diversity patterns at outlier and nonoutlier loci

Log-likelihoods of genotypes were plotted to visualize their similarity to either *Q. robur* (*x*-axis) or *Q. petraea* (*y*-axis; Fig. 4A). Using observed genotypes at the 12 SSRs, several admixed trees could not be distinguished from purebreds, and even some purebreds (indicated by red or blue circles) were not clearly separated on their respective sides of the diagonal. In contrast, the corresponding biplots showed that admixed trees were correctly differentiated from purebreds when the complete SNP data set was used (Fig. 4B).

The total expected heterozygosity (H_t) was significantly higher at outliers than at nonoutliers (0.511 vs. 0.401, $P < 0.001$, see Table 2). At outlier loci (60 SNPs), the mean expected diversity H_e was higher for *Q. petraea* than for *Q. robur* (0.347 vs. 0.163, $P < 0.001$). In contrast, at nonoutlier loci (166 SNPs), H_e was lower for *Q. petraea* than for *Q. robur* (0.173 vs. 0.228, $P < 0.01$). Finally, H_e values calculated in *Q. petraea* and *Q. robur* using data for all SNPs were not significantly different (0.217 and 0.221, $P = 0.85$). Similar results were observed in the conservative approach with a reference F_{ST} value of 0.22: at outlier loci (13 SNPs), H_e was higher for *Q. petraea* than for *Q. robur* (0.299 vs. 0.094, $P < 0.001$). In contrast, at nonoutlier loci (247 SNPs), H_e was slightly lower for *Q. petraea* than for *Q. robur* (0.213 vs. 0.228, $P = 0.35$, Table 2). The same patterns (higher genetic diversity in *Q. petraea* than in *Q. robur* at

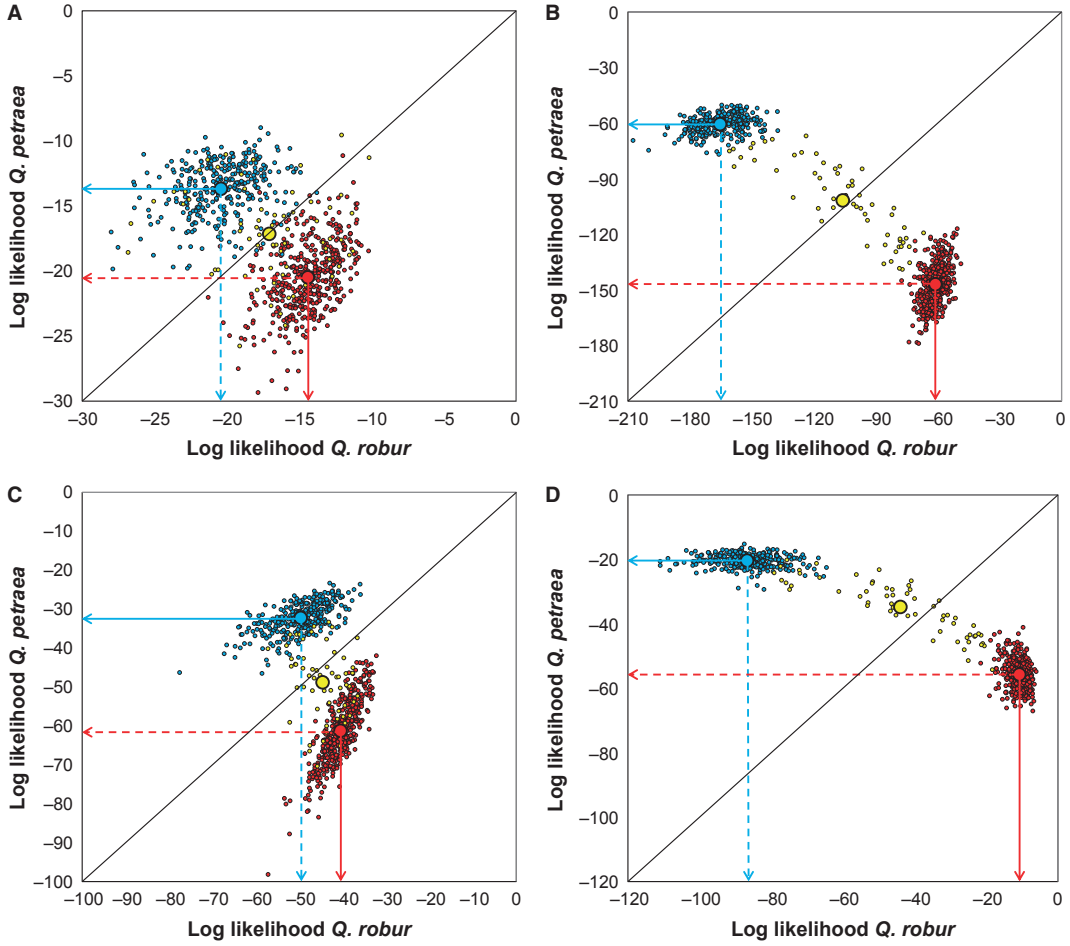

Fig. 4 Biplots of log-likelihoods of assignment to *Quercus petraea* and *Quercus robur* across all individuals. Three groups are distinguished—*Q. petraea* (blue), *Q. robur* (red) and admixed individuals (yellow)—with different categories of markers (A, 12 SSRs; B, all SNPs; C, 166 nonoutlier loci; D, 60 outlier loci). Mean values for each group are indicated by larger circles. Full arrows indicate mean log-likelihoods of conspecific identity and dotted arrows mean log-likelihoods of allospecific identity. The diagonal line helps identify asymmetries in assignment probabilities between species.

outliers but not at nonoutliers) were observed when using two different functional subsets of loci (loci involved in drought stress vs. loci involved in other functions, see Appendix S13, Supporting information).

Similar insights are obtained when considering mean log-likelihood of conspecific identity (full lines in Fig. 4C, D). Even stronger differences between species were detected when comparing mean log-likelihood values of allospecific identity at outlier vs. nonoutlier loci (dotted lines in Fig. 4C, D). For outlier loci, the values differ between the two species by 30 orders of magnitude, implying that *Q. petraea* genotypes are less well affected to *Q. robur* (−86) than *vice versa* (−56). In contrast, for nonoutlier loci, the corresponding difference between mean values of allospecific identity is weaker and in the opposite direction (Fig. 2C). These findings can be related to the counts of private or quasi-private alleles (defined here as alleles present at frequencies higher than 0.5 in one species that are either absent or present at frequencies lower than 0.01 in the other species). There is a higher number of quasi-private alleles in *Q. petraea* (15) than in *Q. robur* (2) among the 60 outlier loci ($P < 0.05$, Fisher test), with the same trend, though not significant, for private alleles. No such trend was found at nonoutlier loci. All the patterns and trends described above were consistent with those found when considering the much smaller number of outliers detected with the very conservative approach or when using different functional subsets of loci.

Discussion

Our increasing ability to isolate large numbers of loci raises questions about the ideal loci to use for reconstructing population structure and demographic history. Loci likely to be affected by the direct or indirect effects of selection are generally excluded to avoid bias when inferring demographic processes, except in the case of spatial gradients and when using loci with known selection intensities. In contrast, interspecific outlier loci were used in our study to better understand past introgression dynamics, considered to reflect an episode of the species' past demography.

Detection and interpretation of outlier and nonoutlier loci

As a prerequisite for exploring the potential interest of outlier loci for characterizing demographic processes, they have to be identified accurately. Using a reference F_{ST} value from the literature, and excluding SNPs that were intraspecific outliers, we detected 60 outlier SNPs (23% of the total, far above the type I error threshold of 5%), with interspecific F_{ST} values ranging from 0.27 to 0.93, that is, on average five times higher than at nonoutliers. This high rate of outliers is consistent with our strategy to deliberately enrich the panel of SNPs genotyped with markers likely to show high divergence between species. Whereas the number of outliers inferred from the reference value could be over-estimated, the number of outliers that were detected directly from the enriched panel (13, that is, 5% of the total) is surely under-estimated, and the reality probably falls between these limits, considering also the large sample size used (around 800 gametes for each species at each locus). In any case, using the smaller set of outliers did not change the diversity patterns and trends observed in comparison with nonoutliers, confirming the robustness of our interpretation. This interpretation was further supported by the fact that the same patterns of higher genetic diversity in *Quercus petraea* than in *Quercus robur* at outliers, but not at nonoutliers, were observed when using two different functional subsets of loci.

Strong outlier patterns have been classically interpreted as being caused by divergent selection affecting the loci themselves or genes strongly linked with them (Storz 2005). Moreover, our selection of SNPs showing a priori high levels of interspecific differentiation or located within candidate genes of ecologically divergent traits between these two oak species might have increased the chance that some outliers are of adaptive significance. Yet, association genetics and functional studies are ultimately required to confirm that particular loci are directly involved in species divergent trait variation. Indeed, alternative explanations for strong genetic divergence at some loci exist and are difficult to rule out (see e.g. Klopfstein *et al.* 2006; Excoffier & Ray 2008; Bierne *et al.* 2011). Problems of interpretations can also arise for nonoutliers (see e.g. Le Corre & Kremer 2003; Latta 2004; Charlesworth 2006; Kremer & Le Corre 2011). Despite these difficulties, the contrast between loci having different levels of divergence should remain informative as long as the average effective gene flow between species is greater at nonoutliers than at outliers.

Species delimitation and SNP discriminatory power

Another prerequisite for our study was to correctly identify individuals belonging to each oak species (i.e. purebreds). Results of the STRUCTURE clustering analysis with 262 SNPs showed that the assignments of individuals to species largely outperformed those from previous studies of the same species based on small sets of SSR loci (Muir *et al.* 2000; Jensen *et al.* 2009; Lepais *et al.* 2009). Validation of assignment performance requires the use of independent samples (Waples 2010). We

therefore confirmed the repeatability of the results for purebreds using independent SNP data sets. We also found that the proportion of admixed trees is prone to overestimation when few loci are used, as previously noted (Vähä & Primmer 2006). Due to the greater abundance of purebred than admixed trees (hybrids *sensu lato*), more purebred trees are likely to be misassigned as hybrids than the converse. Consequently, reducing the precision of assignment by using less loci would artificially increase the proportion of the admixed category, as might have happened in previous studies based on smaller sets of loci or less powerful markers (e.g. Jensen *et al.* 2009; Lepais *et al.* 2009). Moreover, using only the two loci with the highest interspecific F_{ST} (mean = 0.9), assignment performance reached 97% for *Q. robur* and 91% for *Q. petraea*. In contrast, as many as 49 SNP loci with the lowest differentiation (mean interspecific F_{ST} = 0.02) were required to reach similar performance, confirming that locus selection is critical for species delimitation. Overall, our results illustrate the great value of SNPs for assigning individual genotypes: their lower allelic diversity compared with other loci, especially SSRs (Rosenberg *et al.* 2003), can be compensated for by using more loci or selecting outlier loci (Liu *et al.* 2005).

Signals of asymmetric introgression between species

Assignment results based on all SNPs highlight a genetic asymmetry between the two species, *Q. robur* trees being more easily assigned to the purebred category than *Q. petraea* trees. This can be related to the fact that, at outlier loci, *Q. petraea* has higher genetic diversity than *Q. robur*. Altogether, the results based on outlier loci fit well with our expectations for the introgression dynamics between these two species: past asymmetric introgression towards *Q. petraea* should have increased its diversity and decrease the number of private alleles in *Q. robur*. These findings, based solely on data from purebreds and using trees sampled in different populations, point to a relatively ancient and general trend towards asymmetric introgression in the predicted direction (Currat *et al.* 2008).

In the oak colonization model proposed by Petit *et al.* (1997, 2003) to account for shared chloroplast DNA variation across species, a hybrid phase is hypothesized to occur at the time of establishment of *Q. petraea* invading stands already occupied by *Q. robur* through pollen dispersal. Such populations then evolve to yield backcrosses and eventually typical *Q. petraea* trees within a few generations. Thus, there is a stage in the colonization process where the diversity of *Q. petraea* populations would be maximal. Following this, reproductive isolation would rapidly reemerge (see e.g. Gilman & Behm 2011). As loci under divergent selection should be less likely to experience subsequent genetic exchanges between species than other genes, they should retain the initial introgression signal and correspondingly increased genetic diversity in *Q. petraea* most strongly.

At nonoutlier loci, genetic diversity is instead slightly greater in *Q. robur* than in *Q. petraea*, significantly so for the approach using the 0.04 reference F_{ST} value. As nonoutlier loci should behave most closely to neutral expectations, this observation could indicate that *Q. robur* may have a larger effective population size than *Q. petraea*, in line with its greater distribution range and greater dispersal ability through both pollen and seeds (Petit *et al.* 2003). While the latter inference should be confirmed using other methods such as Isolation with Migration coalescent modelling (e.g. Nielsen & Wakeley 2001; Hey & Nielsen 2004), it illustrates the potential benefits of relying on the two different groups of loci to reconstruct particular demographic features of hybridizing species. In fact, in these oaks, the uninformed use of only one class of markers (e.g. only those that are presumably neutral or only those likely to be under divergent selection) would result in opposite conclusions regarding the genetic diversity maintained by each species and the direction of introgression, highlighting the value of jointly considering and comparing results obtained with both groups of markers (Nosil *et al.* 2009). In our study, we were interested in the statistical signals emerging across different sets of loci, not in the behaviour of individual loci. The outlier group potentially includes genes affected by divergent selection, but the approach does not rely on every single locus in that group being actually under divergent selection. Similarly, the approach does not depend on each nonoutlier locus behaving in a strictly neutral manner. Studies aiming at inferring demographic history driven by drift, bottlenecks, gene flow and inbreeding effects are typically based on genome-wide effects of a large number of markers (Luikart *et al.* 2003), whereas selection studies generally focus on particular genes and their locus-specific effects. The approach used here is an original combination of both methods.

Conclusions

We have shown that outlier loci retain signatures of past asymmetric introgression events presumably caused by differences in colonization history, a signature that is missing at other loci. Our approach takes advantage of the nonlinear dependence of genetic structure on levels of gene flow and of the fact that divergent selection can reduce effective gene flow (Bengtsson 1985; Barton & Bengtsson 1986; Nosil *et al.* 2009). Small differences in

gene flow that are usually hard or impossible to detect might become apparent by using loci under divergent selection. While this study dealt with interspecific genetic exchanges, the principles are general. This suggests that loci experiencing reduced effective gene flow due to selection could help reconstruct other aspects of species' demographic histories, providing insights complementary to those obtained with loci evolving closer to neutral predictions. An important perspective is to incorporate variable selection coefficients in model-based approaches aiming at inferring past demographic processes to take advantage of the demographic signal available in the different categories of loci.

Acknowledgements

We thank E. Dreyer and O. Brendel for material and information on the oak stands from eastern France, A. Ducousso and J.-M. Louvet for information on the oak stands from Petite Charnie and S. Wagner for the sampling, and all the persons who contributed to the preliminary candidate gene lists for the oak allelic resequencing project (P. Abadie, C. Bodénès, C. Burban, T. Decourcelle, J. Derory, M.-L. Desprez-Loustau, A. Kremer, G. Le Provost, C. Plomion and C. Robin). We are grateful to J.-M. Frigerio and C. Lepoittevin for their help in developing the new version of the SNP2Illumina perl script, to S. Ueno and I. Lesur for bioinformatic support, to F. Alberto for providing data on previously validated SNPs, to M. Navascués for sharing with us his ideas on the identification of informative loci and to the referees for helpful advices. Genotyping was performed in the Genome-Transcriptome facility of the Functional Genomic Centre of Bordeaux. E.G. was employed during his PhD by the Pernod Ricard Research Center and then as postdoc by INRA (OAKTRACK project, ANR-10-EMMA-0016). T.L. received successive postdoctoral fellowship grants (the first for 18 months and the second for 6 months) from the TRANSBIODIV and LinkTree projects. L.L. was funded by a PhD grant from the EVOLTREE network of Excellence and LinkTree project.

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.

Arnold M (1997) *Natural Hybridization and Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, UK.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.

Barton NH, Hewitt GM (1981) The genetic basis of hybrid inviability in the grasshopper *Podisma pedestris*. *Heredity*, **47**, 367–383.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619–1626.

Bengtsson BO (1985) The flow of genes through a genetic barrier. In: *Evolution: Essays in Honour of John Maynard Smith* (eds. Greenwood JJ, Harvey PH, and Slatkin M), pp. 31–42. Cambridge University Press, Cambridge, UK.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

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

Butlin RK (2008) Population genomics and speciation. *Genetica*, **138**, 409–418.

Carlisle A, Brown AHF (1965) The assessment of the taxonomic status of mixed oak (*Quercus spp.*) populations. *Watsonia*, **6**, 120–127.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, **2**, e64.

Close T, Bhat P, Lonardi S *et al.* (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, **10**, 582.

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

Cousens JE (1963) Variation of some diagnostic characters of the sessile and pedonculate oaks and their hybrids in Scotland. *Watsonia*, **5**, 273–286.

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

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

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

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934.

Gilman RT, Behm JE (2011) Hybridization, species collapse, and species reemergence after disturbance to premating mechanisms of reproductive isolation. *Evolution*, **65**, 2592–2605.

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

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

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel

organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 1–14.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

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.

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

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kremer A, Le Corre V (2011) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, **108**, 375–385.

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, Guichoux E, Petit RJ (2012) Fine-scale environmental control of hybridization in oaks. *Molecular Ecology*, doi: 10.1111/mec.12121.

Latta RG (2004) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, **161**, 51–58.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Lenormand T, Guillemaud T, Bourguet D, Raymond M (1998) Evaluating gene flow using selected markers: a case study. *Genetics*, **149**, 1383–1392.

Lepais O, Petit RJ, Guichoux E *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology*, **18**, 2228–2242.

Lepoittevin C, Frigerio J-M, Garnier-Géré P *et al.* (2010) *In vitro vs in silico* detected SNPs for the development of a genotyping array: what can we learn from a non-model species? *PLoS ONE*, **5**, e11034.

Li JR, Li HP, Jakobsson M, Li S, Sjodin P, Lascoux M (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, **21**, 28–44.

Liu NJ, Chen L, Wang S, Oh CG, Zhao HY (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics*, **6**, S26.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **363**, 3009–3021.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Mallet J, Barton N, Gerardo LM, Jose SC, Manuel MM, Eeley H (1990) Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics*, **124**, 921–936.

Marsden CD, Lee Y, Nieman CC *et al.* (2011) Asymmetric introgression between the M and S forms of the malaria vector, *Anopheles gambiae*, maintains divergence despite extensive hybridization. *Molecular Ecology*, **20**, 4983–4994.

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

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, **70**, 3321–3323.

Neigel JE (2002) Is F_{ST} obsolete? *Conservation Genetics*, **3**, 167–173.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Nielsen EE, Bach LA, Kotlicki P (2006) Hybridlab (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes*, **6**, 971–973.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology*, **13**, 55–65.

Peakall R, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Petit RJ, Bahrman N, Baradat P (1995) Comparison of genetic differentiation in maritime pine (*Pinus pinaster* Ait) estimated using isozyme, total protein and terpenic loci. *Heredity*, **75**, 382–389.

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A (1997) Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9996–10001.

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

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Quintana-Murci L, Quach H, Harmant C *et al.* (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences*, **105**, 1596–1601.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402–1422.

Scotti-Saintagne C, Mariette S, Porth I *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615–1626.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.

Steinhoff S (1993) Results of species hybridization with *Quercus robur* L. and *Quercus petraea* (Matt) Liebl. *Annals of Forest Science*, **50**, 137s–143s.
Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.
Streiff R, Ducousso A, Lexer C, Steinkellner H, Gloessl J, Kremer A (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Molecular Ecology*, **8**, 831–841.
Szymura JM, Barton NH (1991) The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, **45**, 237–261.
Tiffin P, Olson MS, Moyle LC (2001) Asymmetrical crossing barriers in angiosperms. *Proceedings of the Royal Society of London B: Biological Sciences*, **268**, 861–867.
Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.
Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.
Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, **89**, 438–450.
Waples RS (2010) High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology*, **19**, 2599–2601.
Waser PM, Strobeck C (1998) Genetic signatures of interpopulation dispersal. *Trends in Ecology & Evolution*, **13**, 43–44.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

E.G., P.G.-G. and R.J.P. designed the experiment when E.G. was a PhD student under the supervision of R.J.P. P.G.-G. coordinated the resequencing study that provided the main source of SNPs (95%) for this study. T.L. and P.G.-G. developed the bioinformatics tools used to identify and validate these SNPs. L.L., E.G. and P.G.-G. conceived the Illumina genotyping assay. L.L. and C.B. performed SNP genotyping and provided multilocus genotypes. E.G. analysed the data with the help from P.G.-G. (outlier detection) and R.J.P. (species delimitation). E.G., P.G.-G. and R.J.P. wrote the paper. All authors have checked and approved the final version of the manuscript.

Data accessibility

Summary SNP data are available from Appendix S5 (Supporting information).

SNP genotypes are available from Dryad: doi:10.5061/dryad.3g140.

Locus name, sequence, target functional trait, gene annotation and contig reference are available from the *Quercus* Portal: https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=snp.

Contig assembly, contig blast and data mining are available from the GENOTOUL Forest Trees Contig Browser Oak: http://genotoul-contigbrowser.toulouse.inra.fr:9092/Quercus_robur/index.html.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Impact of selection on indirect measures of gene flow.

Appendix S2 Sampling sites in France.

Appendix S3 Distribution of *Quercus petraea* (blue) and *Quercus robur* (red) across Europe (Ducousso & Bordacs, with permission).

Appendix S4 Detailed SNP selection and outlier detection methods.

Appendix S5 Detailed characteristics of the 384 SNPs selected for inclusion in the Illumina array (.xls file).

Appendix S6 Distribution of interspecific F_{ST} values as a function of the mean values of within-demes heterozygosity (H_{WP}), simulated with the FDIST method implemented in ARLEQUIN 3.5.1.2 (Excoffier *et al.* 2009).

Appendix S7 Characteristics of the 384 SNPs used.

Appendix S8 Barplot of the number of individuals (among 855) assigned to different categories from either 12 SSRs (light colours) or 262 SNPs (dark colours).

Appendix S9 Assignment score of 5000 simulated genotypes, i.e. 1000 purebreds *Quercus robur*, 1000 purebreds *Quercus petraea* and 3000 admixed individuals (1000 F1s, 1000 backcrosses *Q. robur* and 1000 backcrosses *Q. petraea*) from 262 SNPs.

Appendix S10 Distribution of observed interspecific F_{ST} values (calculated between purebreds over 262 SNPs).

Appendix S11 Distribution of intraspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), using a finite island model with six demes for *Quercus robur*.

Appendix S12 Distribution of intraspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), using a finite island model with six demes for *Quercus petraea*.

Appendix S13 Genetic diversity and inter- and intraspecific differentiation at outlier vs. non-outlier loci for the subset of loci involved in drought stress (A) and for the subset of loci not involved in drought stress (B).