

ORIGINAL ARTICLE

Detecting the footprints of divergent selection in oaks with linked markers

PG Goicoechea¹, RJ Petit^{2,3} and A Kremer^{2,3}

Genome scans are increasingly used to study ecological speciation, providing a useful genome-wide perspective on divergent selection in the presence of gene flow. Here, we compare current approaches to detect footprints of divergent selection in closely related species. We analyzed 192 individuals from two interfertile European temperate oak species using 30 nuclear microsatellites from eight linkage groups. These markers present little intraspecific differentiation and can be used in combination to assign individual genotypes to species. We first show that different outlier detection tests give somewhat different results, possibly due to model constraints. Second, using linkage information for these markers, we further characterize the signature of divergent selection in the presence of gene flow. In particular, we show that recombination estimates for regions with outlier markers are lower than those for a control region, in line with a prediction from ecological speciation theory. Most importantly, we show that analyses at the haplotype level can distinguish between truly divergent (bi-directional) selection and positive selection in one of the two species, offering a new and improved method for characterizing the speciation process.

Heredity (2012) **109**, 361–371; doi:10.1038/hdy.2012.51; published online 19 September 2012

Keywords: divergent selection; divergence hitchhiking; outlier detection; linkage disequilibrium; score tests; *Quercus*

INTRODUCTION

Divergent selection, defined as selection acting in contrasting directions in different groups of individuals (Schluter, 2000), has been a topic of great interest in recent years because of its central role in studies of ecological speciation. Although work in this area initially focused on the responses of individual genes to divergent selection, there has been a gradual shift towards a genomic approach (Nosil and Feder, 2012), with a growing number of studies using genome/multilocus scans to analyze this process in relation to ecological speciation (Nosil *et al.*, 2009; Apple *et al.*, 2010; Galindo *et al.*, 2010; and references therein). Such scans rely on single-locus outlier detection methods, which are based on population genetics theory (Lewontin and Krakauer, 1973; Bowcock *et al.*, 1991; Beaumont and Nichols, 1996; Schlötterer, 2002; Beaumont and Balding, 2004). Positive selection reduces neutral diversity at linked neutral sites via hitchhiking effects (Maynard-Smith and Haigh, 1974), creating a selective sweep. Other phenomena associated with selective sweeps include changes in linkage disequilibrium (LD) patterns (Kim and Nielsen, 2004; McVean, 2007) and in the site frequency spectrum (Braverman *et al.*, 1995; Nordborg *et al.*, 2005).

Recently, Via and West (2008) and Via (2009) described a new type of hitchhiking that differs from hitchhiking after a selective sweep. This between-population process (Via, 2012) is exclusive to organisms evolving under divergent selection and has therefore been named 'divergence hitchhiking'. In contrast to what happens following a selective sweep in a large panmictic population, where LD created by the sweep is rapidly eroded, LD tends to be more persistent around

genes that are responding to divergent selection among populations due to a reduction in the rate of effective gene flow and recombination (Via, 2009). Eventually, as speciation progresses, divergent selection in the presence of gene flow produces a genetic mosaic of divergent and non-divergent genomic regions (Nosil *et al.*, 2007; Smadja *et al.*, 2008; Via, 2012). Divergence hitchhiking should also operate under other modes of divergent selection, such as when intrinsic (non-environmental) factors drive selection, which should produce distinctive selection footprints (Bierne, 2010). Therefore, there is a need to evaluate suitable approaches to better characterize the molecular signatures of these processes.

Attempts to use outlier detection tests to infer divergent selection have highlighted the lack of available information on several related subjects, such as the actual causes of selection, the potential impact of background differentiation on outlier detection and the likelihood of false positives (Nosil *et al.*, 2009; Michel *et al.*, 2010). Despite these issues, a growing number of researchers have attempted to infer the occurrence of ecological speciation by identifying outlier markers related to changes in environmental variables (Murray and Hare, 2006; Savolainen *et al.*, 2006; Joost *et al.*, 2007) or by identifying cases where the same outliers occur in replicated pairs of populations (Bonin *et al.*, 2006; Egan *et al.*, 2008; Poncet *et al.*, 2010). To date, none of these studies have considered linkage information. However, haplotype tests based on LD mapping have proven to be more powerful (that is, require smaller sample sizes) than standard single-locus methods in association studies (Bader, 2001; Liu *et al.*, 2008) and candidate gene studies (Clark, 2004). Haplotype tests should also

¹Department of Biotechnology, NEIKER-Tecnalia, Vitoria-Gasteiz, Spain; ²INRA, UMR 1202 Biogeco, Cestas, France and ³University of Bordeaux, UMR 1202 Biogeco, Talence, France

Correspondence: Dr PG Goicoechea, Department of Biotechnology, NEIKER-Tecnalia, P.O. Box 46, 01080 Vitoria-Gasteiz, Spain.

E-mail: pgoicoetxea@neiker.net

Received 17 October 2011; revised 13 June 2012; accepted 17 July 2012; published online 19 September 2012

be more robust than single-marker tests because evolutionary forces tend to increase the variability of single markers and the simultaneous analysis of multiple markers in the form of haplotypes should thus yield comparatively simpler patterns (Akey *et al.*, 2001). Moreover, advanced statistical methods for inferring haplotypes from genotypic data have now been developed (Stephens *et al.*, 2001; Excoffier *et al.*, 2003; Stephens and Donnelly, 2003) and divergence hitchhiking is expected to produce LD extending across relatively large chromosomal regions (Via, 2012). This suggests that it may be possible to analyze divergence hitchhiking regions by scanning the genomes of ecologically important non-model species with linked markers.

Oaks represent a particularly good model system to study divergent selection in the presence of gene flow because the reproductive barriers in this genus seem to depend on the ecological contexts (Muller, 1952; Lepais *et al.*, 2009; Zeng *et al.*, 2010). Two of the most extensively studied oak species are the sessile *Quercus petraea* and the pedunculate *Q. robur* (Petit *et al.*, 2004). They are commonly found in sympatry across Europe and Asia, in mixed forests, where hybridization occurs spontaneously (Streiff *et al.*, 1999; Lepais *et al.*, 2009; Lepais and Gerber, 2011). However, there are persistent strong differences between the two species with respect to certain ecologically important traits, such as leaf and fruit morphology (Kremer *et al.*, 2002), water-use efficiency (Ponton *et al.*, 2002; Parelle *et al.*, 2007) and the amount of volatile compounds in their wood (Prida *et al.*, 2007). Scotti-Saintagne *et al.* (2004) used a genetic mapping approach to demonstrate the existence of several 'islands' of strongly differentiated loci surrounded by regions of low differentiation, which is consistent with the expected spatial clustering of outlier markers due to ecological speciation (Via, 2009). Later, Muir and Schlötterer (2005) identified one outlier marker in two such islands using a different outlier detection test.

In the study presented herein, we further investigate the signatures of divergent selection in these oak species. The study was conducted in three stages. First, we used five outlier detection methods based on different underlying assumptions to search for selection signals at individual loci. Second, we used a limited number of loosely linked markers to explore the haplotypic structures around outlier loci from several linkage groups. Clear differences were observed between the haplotypic structures of the selected regions and those of a control region consisting of three markers with average differentiation. Finally, we show that haplotypic methods, in contrast to outlier methods, are capable of distinguishing between truly divergent (bi-directional) selection and directional selection in one species only.

MATERIALS AND METHODS

A total of 192 trees from the two European temperate oak species (*Q. petraea* (Matt.) Liebl. and *Q. robur* L.) were sampled from 10 populations located in eight European countries. Their geographical coordinates, the type of stand involved (mono-specific or mixed-species stands), the morphological classification of the sampled trees and the number of trees and simple sequence repeats (SSRs) analyzed within each population are provided in Table 1. Half of the studied data set has been used in a previous investigation (Scotti-Saintagne *et al.*, 2004). In this work, we increased the number of trees in all mixed stands and added three new stands. For these new samples, we analyzed only a subset of 18 SSRs that displayed significant between-species differentiation (based on F_{ST}) in the first survey.

Leaf morphology was used to guide sampling in mixed forests. In particular, morphologically intermediate individuals were avoided because they could increase the proportion of admixed trees in the sample (see below). The SSRs used in this study were designed by Dow *et al.* (1995), Steinkellner *et al.* (1997) and Kampfer *et al.* (1998). We selected 30 SSRs that had been mapped on a *Q. robur*-controlled cross (Barreneche *et al.*, 1998; Scotti-Saintagne *et al.*, 2004).

Table 1 Characteristics of the sampled populations

Stand (name)	Latitude/longitude	Type	Species (morph)	Nb trees	SSRs
P1 (Petit Charnie)	48.083/−0.167	Mixed	<i>Q. petraea</i>	8	30
				4	18
			<i>Q. robur</i>	8	30
				4	18
P3 (De Meinweg)	51.179/6.14	Mixed	<i>Q. petraea</i>	12	18
			<i>Q. robur</i>	12	18
P4 (Arlaban)	42.967/−2.55	Mixed	<i>Q. petraea</i>	8	30
				4	18
			<i>Q. robur</i>	8	30
			4	18	
P4b (Montejo)	41.117/−3.500	Mixed ^a	<i>Q. petraea</i>	8	30
P5 (Dalkeith)	55.903/−3.055	Mixed	<i>Q. petraea</i>	8	30
				4	18
			<i>Q. robur</i>	8	30
			4	18	
P6 (Sigmundsherberg)	48.680/15.750	Mixed	<i>Q. robur</i>	8	18
P7 (Roudsea Wood)	54.218/−3.018	Mixed	<i>Q. petraea</i>	8	30
				4	18
			<i>Q. robur</i>	8	30
			4	18	
P9 (Gysinge)	60.468/17.310	<i>Q. robur</i>	<i>Q. robur</i>	8	30
P10 (Büren)	47.105/7.383	Mixed	<i>Q. petraea</i>	8	30
				4	18
			<i>Q. robur</i>	8	30
			4	18	
P12 (Rantzau)	54.250/10.520	Mixed	<i>Q. petraea</i>	12	18
			<i>Q. robur</i>	12	18

Abbreviation: SSRs, simple sequence repeats.

^a*Q. petraea*-*Q. pyrenaica* mixed stand.

Their linkage map positions are shown in Figure 1. Protocols for DNA isolation and PCR amplification were slightly modified from those described by Mariette *et al.* (2002). Forward primers were 5' labeled with either the IRD700 or the IRD800 chromophore (Eurofins MWG Operon, Ebersberg, Germany) and alleles were visualized on a LI-COR sequencer (LI-COR Biosciences, Lincoln, NE, USA). The absolute sizes of the alleles were determined by direct comparison with a 10-bp ladder (LI-COR Biosciences) run in triplicate on each gel. Stutter bands and the IMAGEJ (Abramoff *et al.*, 2004) software package were used to match the different alleles within each gel.

Population structure

We inferred the genetic structure in our sample using STRUCTURE v.2.1 (Pritchard *et al.*, 2000; Falush *et al.*, 2003). Inferences were based on a linkage model with correlated allele frequencies, using a unique drift rate from the ancestral population(s) (see Supplementary File 1 for the results obtained with another model and for an admixture test). We ran five independent chain replicates of 10^6 iterations following a burn-in period of 10^5 iterations, for a fixed number of populations ($K=1-5$). Previous analyses on a subset of the data had already demonstrated that when using this model, the probabilities obtained decreased steadily after $K=3$ (data not shown). Model convergence was checked using STRUCTURE likelihood and summary statistics plots. We used the *post-hoc* criterion described by Evanno *et al.* (2005) to determine the most likely number of clusters. The full search algorithm of CLUMPP v.1 (Jakobsson and Rosenberg, 2006) was used to handle the across-chains variation in the proportion of intra-individual mixed ancestry. The resulting final configuration was used to plot the ancestry coefficients with DISTRUCT v.1.1 (Rosenberg, 2004).

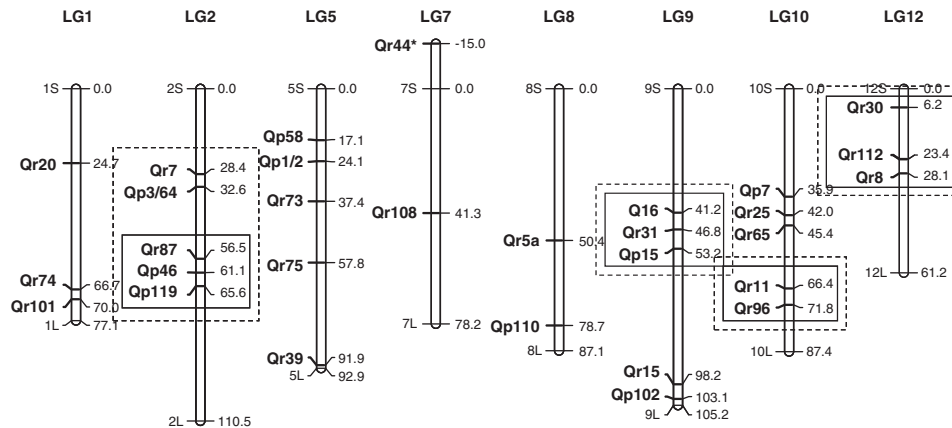


Figure 1 Oak linkage groups showing the location of the 30 SSRs used in this study (after Scotti-Saintagne *et al.*, 2004). Map positions are indicated to the right of each LG. The boxes indicate the markers used for haplotype inferences and recombination estimates (continuous line) and for haplotype score tests (dashed line).

Genetic diversity and differentiation

The data set for ‘pure’ species (see Supplementary File 1) was used to estimate the genetic diversity and interspecific differentiation for the 30 SSRs. Observed (H_o) and expected heterozygosities (H_e) and allelic richness (AR) were calculated using the methods described by Nei (1973) and El Mousadik and Petit (1996), respectively. We measured interspecific differentiation using F_{ST} (Weir and Cockerham, 1984) and D (Jost, 2008), another differentiation parameter that uses a multiplicative partitioning of diversity based on the effective number of alleles. Unbiased estimates of these parameters were calculated with FSTAT (Goudet, 2001) and SMOGD (Crawford, 2010). The significance of each F_{ST} estimate was evaluated using 16 000 permutations. Similarly, 1000 bootstrapped samples were used to calculate the variance of D and its 95% confidence intervals. These results are shown in Supplementary File 2.

Outlier tests

Only data on purebred individuals were retained for these analyses. We used five outlier tests with different underlying assumptions. The first two tests are based on the F_{DI}ST2 method of Beaumont and Nichols (1996), which uses coalescent simulations under either the infinite allele model (IAM, Kimura and Crow, 1964) or the stepwise mutation model (SMM, Ohta and Kimura, 1973) to characterize genetic differentiation (F_{ST}) for a given heterozygosity level. The third test, BAYESFST (Beaumont and Balding, 2004), relies on a Bayesian analysis to characterize differences in migration rates among individual markers. The last two methods, $\ln RV$ and $\ln RH$ (Schlötterer, 2002; Kauer *et al.*, 2003), use the observed data to fit standardized normal distributions whose tails contain the outliers. Further details concerning these analyses are presented in Supplementary File 3.

Linkage disequilibrium

LD estimates were used to search for further selection signals in our data, as an increase in LD between markers is commonly used to infer selection (Hudson *et al.*, 1994; Kohn *et al.*, 2000). Genotypic LD within species was investigated using FSTAT (Goudet, 2001), based on data from purebred individuals. Significance tests were conducted using the log-likelihood ratio G -statistic (Goudet *et al.*, 1996), with 16 000 permutations. Deviations from Hardy–Weinberg equilibrium (H–W equilibrium) were tested for on a per locus basis by randomizing alleles among individuals within samples to determine the departure of F_{IS} from zero.

Obtaining reliable genotypic LD estimates from multiallelic markers, such as SSRs, requires very large sample sizes given the large number of genotypes involved. This difficulty can be eliminated by analyzing haplotypic LD, for which smaller sample sizes are needed. We therefore estimated haplotypic LD in all segments harboring positive selection outliers (linkage group 2 (LG2), LG10 and LG12) and in one control LG without outliers (LG9). Following

haplotype reconstructions (see below), we used ARLEQUIN v.3.5 (Excoffier and Lischer, 2010) to perform LD analysis while accounting for haplotype information. The exact test of Guo and Thompson (1992) was used to detect deviations from H–W equilibrium using 5×10^5 dememorization steps and a final chain of 10^7 steps. The significance of the observed LD was estimated using an extension of Fisher’s exact test on contingency tables (Raymond and Rousset, 1995). In this test, Markov chains were run for 10^7 steps that were preceded by 9×10^5 dememorization steps. Results concerning both genotypic and haplotypic LD estimates are shown in Supplementary File 6.

Inferring haplotypes

We first inferred haplotypes and then performed haplotype-based analyses and studied haplotype sharing among individuals and between species. Haplotypes were inferred using two methods that explicitly allow for recombination: PHASE v.2.1 (Stephens *et al.*, 2001; Stephens and Donnelly, 2003) and the ELB algorithm (Excoffier *et al.*, 2003). Four linkage groups were studied: the three LGs that contain at least one marker that appeared to be under positive selection (LG2 (Qp119, Qp46 and Qr87), LG10 (Qr11 and Qr96) and LG12 (Qr8, Qr112 and Qr30)) and one control region in which we found no outliers (LG9 (Q16, Qr31 and Qp15)). Note that Q16 and Qp15 had low/moderate allelic richness, which facilitated the detection of shared haplotypes in spite of the lower number of genotyped trees at Q16 and Qr31. Other segments with non-outliers situated at appropriate distances could not be used as additional control regions owing to the large number of alleles at the corresponding loci (that is, Qr72 and Qr101 or Qp7, Qr25 and Qr65). Haplotype reconstruction was conducted for markers from each LG using the full data set after excluding trees with missing data. Details on the inferences made using the ELB algorithm are presented in Supplementary File 5.

PHASE haplotype inferences for each linkage group were based on five independent Bayesian chains with lengths of 5×10^8 , chains were sampled every 500 iterations after a burn-in period of 2.5×10^5 iterations. The relative positions of each marker within the linkage groups (Figure 1) were transformed into base pairs using the assumption that 1 cM is roughly equal to 10^8 bp. This is the default value for humans (PHASE documentation) and was used by Koopman *et al.* (2007) in their analyses of gene flow and introgression in apple trees, whose genomes are about the same size as those of oaks. We also tested a larger recombination probability obtained by computing the ratio of the size of the oak genome (Zoldo *et al.*, 1988) to the full linkage length ($c = (0.94 \text{ pg} \times 0.975 \text{ Gb pg}^{-1})/1200 \text{ cM}$). The IAM was used to model mutations at loci Qp46 and Qr31 (because they contain a large number of alleles that differ by only 1 bp), whereas the SMM was used to model mutations at all other markers. Both models were allowed to occasionally deviate from the most stringent assumptions by using δ values of 0.01 and 0.99 (where $\delta = 0$ and $\delta = 1$ correspond to the pure IAM and SMM models, respectively). Our assumptions concerning the variation in the recombination rate are consistent with the general model of Li and Stephens (2003).

We also used PHASE to estimate the population genetics recombination parameter ($\rho = 4Nc$), where c is the recombination probability per base pair and N is the effective population size, using inferred haplotypes as a starting point (Li and Stephens, 2003; Crawford *et al.*, 2004). We compared the recombination estimates for the three LGs containing markers putatively under positive selection (LG2, LG10 and LG12) to that for the control LG without any apparent outliers (LG9). We used the -X100 option and modified the number of burn-in iterations, the thinning intervals and the final chain lengths to increase the accuracy of the estimates (see Supplementary File 4 for details). Recombination rates were estimated using the default number of individuals (100) in all cases save that of LG9 (84). To test the calculations' sensitivity to the recombination priors, we first used the default priors for the population genetics recombination parameter ($\mu = 0.0004$) and for the difference allowed between the estimate and the prior ($f = 10^6$), both of which are appropriate for humans (PHASE documentation). We then used two 'oaks priors' ($\mu = 0.04$, $f = 10^4$) that are expected to be more informative on the basis of experimental results (see Jaramillo-Correa *et al.* (2010), for a ρ estimate in tree populations). In all cases, we tested the convergence of the Bayesian chains (Supplementary File 4) before reporting the medians.

Haplotype score tests

We used haplotype score tests (Schaid *et al.*, 2002) to search for significant associations between haplotypes and species. The score tests are an extension of the trend in proportions tests (Armitage, 1955), which are commonly used to compare haplotypic frequencies between cases and controls in association studies (Devlin and Roeder, 1999). One important advantage of using score tests is that they can incorporate the posterior haplotype probabilities, which are included as covariates. This makes them appropriate for analyzing inferred haplotypes (Schaid *et al.*, 2002). On the other hand, the tests assume that all individuals are unrelated (Sinnwell and Schaid, 2009), which could make them sensitive to the genetic structure in our data (but see below). After excluding missing data, we used the full data set to analyze the five markers in the LG2 (region Qp119–Qr7), the three markers in the LG12 (region Qr8–Qr30), the two markers in the LG10 (region Qr11–Qr96) and the three markers in the LG9 (region Q16–Qp15).

These analyses were performed using the 'haplo.stats' library (Sinnwell and Schaid, 2009) in the R environment (R Development Core Team, 2010). We obtained 100 maximum likelihood estimates of haplotype probabilities for each species with the modified EM algorithm (Excoffier and Slatkin, 1995; Schaid *et al.*, 2002) and the solution with the largest likelihood was retained. Score statistics for the associations between the haplotypes and the binary trait 'species', together with P -values, were obtained using the 'haplo.score' function with the additive haplotype effects model. We used two thresholds for minimum haplotype counts (10 and 5) to avoid introducing bias in the global score due to rare haplotypes. The data set for LG9 was comparatively small and so a threshold of 3 was used in this case. The contributions of consecutive two-marker haplotypes from LG2, LG12 and LG9 were compared with the function 'haplo.score.slide', which uses a sliding-window approach.

RESULTS

The structure of the genetic data

We identified two clusters in our sample (Supplementary Figure S1-1a) using the *post-hoc* criterion of Evanno *et al.* (2005). Furthermore, the plot of intra-individual ancestries showed that the genetic groups closely reflected the trees' morphological classification (Supplementary Figure S1-1b). The ancestry of most of the studied trees derived almost exclusively from just one of the genetic groups ($q_i < 0.15$ or $q_i > 0.85$). These trees were considered to be 'purebreds'. However, a significant number of the trees (13.5%), most of which came from the two British populations, had more balanced levels of shared ancestry ($0.15 < q_i < 0.85$). Simulations including a prior based on morphological classification (see Supplementary File 1 and Supplementary Figure S1-3) provided statistical support for the hypothesis that this shared ancestry was due to admixture.

Heterozygosity and differentiation

The two oak species were found to have similar levels of genetic diversity and allelic richness (Supplementary File 2). The mean heterozygosity values for *Q. petraea* and *Q. robur* were 0.874 and 0.875, respectively, whereas the corresponding mean allelic richness values were 23.0 and 24.8. A total of 19 SSRs had interspecific F_{ST} values differing significantly from zero after the Holm–Bonferroni correction. All LGs studied contained at least one marker with a significant F_{ST} value. The estimated values for the differentiation parameter D (Jost, 2008) were much larger than those for F_{ST} . All of the confidence intervals for D excluded zero, indicating significant interspecific differentiation for all 30 markers.

Outlier tests

The various methods used to search for outlier loci yielded complex results. In particular, none of the markers was identified as an outlier by every test applied. Two tests (F_{DIST2} , using the SMM, and $\ln RV$) were sufficient to discover all five candidate loci for positive selection (Figure 2); the other tests either provided redundant information or

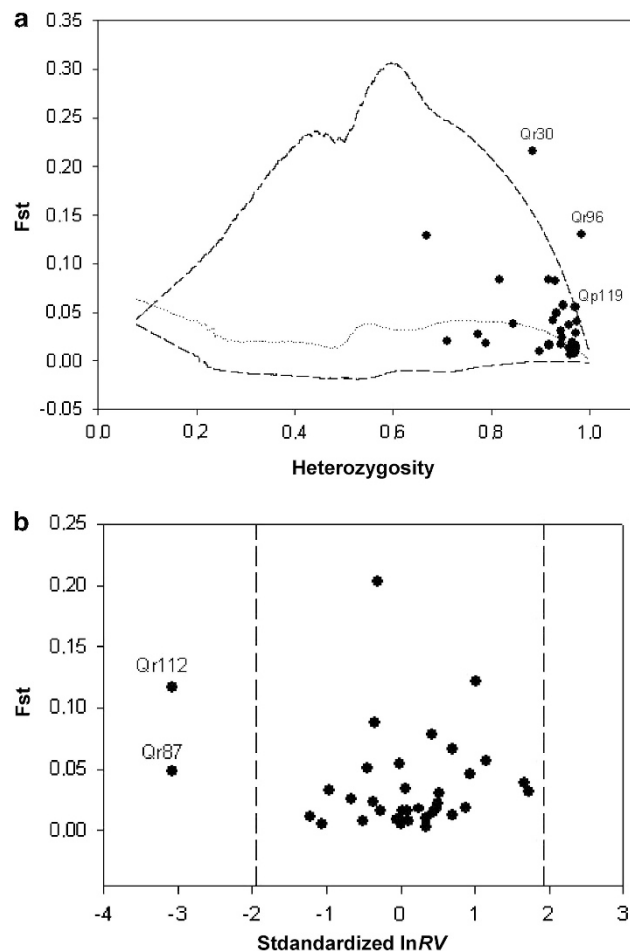


Figure 2 Detection of differentiation outlier markers using two outlier detection methods: F_{DIST2} (using the SMM for coalescent simulations) and $\ln RV$. In the first chart, (a) the observed heterozygosity is plotted against the differentiation values. The 95% envelope (dashed lines) and the median (dotted line) obtained from the coalescent simulations are also plotted. In the second chart, (b) the standardized $\ln RV$ estimates are plotted against the differentiation values. Dashed lines indicate the limits of the 95% confidence interval (-1.96 , $+1.96$).

identified candidate loci for balancing selection (Supplementary Table S3-1).

At least part of the observed heterogeneity in the selection signal is likely to be due to differences in the models underpinning the different tests, as shown in Supplementary Figure S3-2. Markers for which the heterozygosity was much lower in one species than in the other (for example, Qr30, Qr96 or even Qp119) were outliers in tests based on F_{ST} -heterozygosity comparisons. Markers that had fewer alleles at the distribution tails in one species (Qr87, Qr112) were outliers in the $\ln RV$ test statistic. The Bayesian test, which is the only one that models migration, detected one marker with reduced heterozygosity (Qr96) and one marker with reduced variation in the number of repeats (Qr112).

Genotypic linkage disequilibrium

Only two markers located 5 cM apart in LG2 (Qp46 and Qr87) showed significant LD in both species (Supplementary File 6, Supplementary Table S6-1). One of the involved loci (Qr87) was identified as a significant outlier based on the $\ln RV$ test statistic.

Haplotype sharing among individuals and between species

We used three criteria to evaluate the confidence in the haplotypes inferred using PHASE: (1) the coincidence of the haplotypes obtained from five independent Bayesian chains, (2) the goodness of fit between the inferred haplotypes and an approximate coalescent with recombination (Supplementary File 4) and (3) the probabilities assigned by PHASE to the inferred haplotypes (Supplementary Figure S4-1). Furthermore, we found that the two reconstruction methods (PHASE and the ELB algorithm) infer similar haplotypes and similar haplotypic LD patterns. As a result, the following discussion focuses exclusively on the haplotype sharing and haplotypic LD results obtained using PHASE. See Supplementary File 4 and Supplementary Tables S6-2,3 for results obtained using the ELB algorithm.

Despite the low number of shared haplotypes and their low frequencies (Supplementary Table S4-1), almost two-thirds of the trees carried at least one shared haplotype. As expected, shared haplotypes were slightly easier to reconstruct (that is, had higher probabilities) than unique haplotypes (Supplementary Figure S4-1, left panel). The medians of the probability values for all the inferred haplotypes were rather large for all LGs harboring putatively selected loci (LG2: 0.80; LG10: 0.96; and LG12: 0.76). For LG9, three-marker haplotypes were reconstructed with lower confidence (median value = 0.69), this was probably partly due to the lower sample size and partly because of a lower LD in this case (see below). Comparisons between the two species showed that *Q. petraea* haplotypes were reconstructed with higher confidence than *Q. robur* haplotypes in LG2 and LG12, whereas the opposite pattern was found for LG10 (Supplementary Figure S4-1, right panel). We attribute this to heterozygosity differences at the corresponding loci (see Supplementary File 2).

High-frequency haplotypes were species-specific in LG2, LG10 and LG12, whereas in LG9 they were shared across species (Figure 3). Note that in LG10, the most frequent haplotype was shared between the two species but other frequent haplotypes were species-specific.

Recombination estimates

Coalescent modeling was used to estimate background recombination rates using haplotypic data as the starting point. Only Bayesian chains that used the 'oak priors' and the small recombination probability reached convergence (Supplementary File 5). Therefore, we report

only estimates obtained with these priors (Table 2). The rate of recombination within LG10, which includes a segment that is putatively under positive selection, was more than ten times lower than that within control segments from LG9. Bayesian chains reached convergence for recombination estimates in LG12, but longer chains were needed to ensure accurate point estimates (Supplementary File 5). Despite this, the results obtained suggest that the background recombination rate for this region is 15 times lower than that for the control, and that the rate for the Qr112–Qr30 region, which was also inferred to be under directional selection, is even lower.

LD using haplotype inferences

Considering only markers in H–W equilibrium, we found significant LD at LG2 in both species (Supplementary Table S6-2). Several allele pairs showed large contributions to the disequilibrium (Table 3). The only other region in *Q. petraea* with markers at H–W equilibrium is the Qr112–Qr30 segment. The global LD test was not significant for this region (Supplementary Table S6-2), as would be expected for such a large segment. Two other regions in *Q. robur* were at H–W equilibrium (Qr11–Qr96 and Qr8–Qr112). The global LD in the first of these regions was not significant, although one haplotype was not at equilibrium. The last segment showed significant global LD and an excess of several high-frequency haplotypes (Table 3 and Supplementary Table S6-2).

In the case of LG10 in *Q. petraea*, Qr96 displayed only weak deviation from H–W equilibrium frequencies, whereas Qr11 was in H–W equilibrium. Significant global LD and an excess of several high-frequency haplotypes were inferred in this region, assuming that the analyses are robust to weak departures from H–W equilibrium (Table 3 and Supplementary Table S6-2).

Haplotype score tests for association

Global score tests for association between full-length haplotypes and species were significant in LG2, LG10 and LG12, but non-significant in LG9 (data not shown). The sliding-window approach (Figure 4a) showed that the tests were significant for the LG2-S region (Qp119–Qp46–Qr87), with the strongest association found for the Qp46–Qr87 segment. The two-marker haplotypes from LG12 were also significantly associated with species, this association was strongest for the Qr112–Qr30 segment. However, the tests were non-significant for the two segments from LG9.

In contrast to global scores, haplotype-specific scores can be used to determine whether the associations between haplotypes and traits are present in both species or only in one (Figure 4b). The two LG2 segments contain haplotypes that are significantly associated with both species (positive scores for *Q. petraea* and negative scores for *Q. robur*), whereas LG12 haplotypes are significantly associated with *Q. petraea* only (positive scores). LG10 had haplotypes associated with both species (not shown). We did not estimate haplotype-specific scores for LG9 because the global and the sliding-window tests were non-significant in this case.

DISCUSSION

We have used haplotypic tests and recombination estimates to characterize regions surrounding differentiation outliers in oaks. Our results provide stronger evidence for directional selection than was obtained in previous studies that used multilocus data sets to search for selection signals without considering LD information, because LD between markers is expected to increase around gene(s) responding to selection (Kim and Nielsen, 2004). The number of markers used in this methodological study was substantially smaller

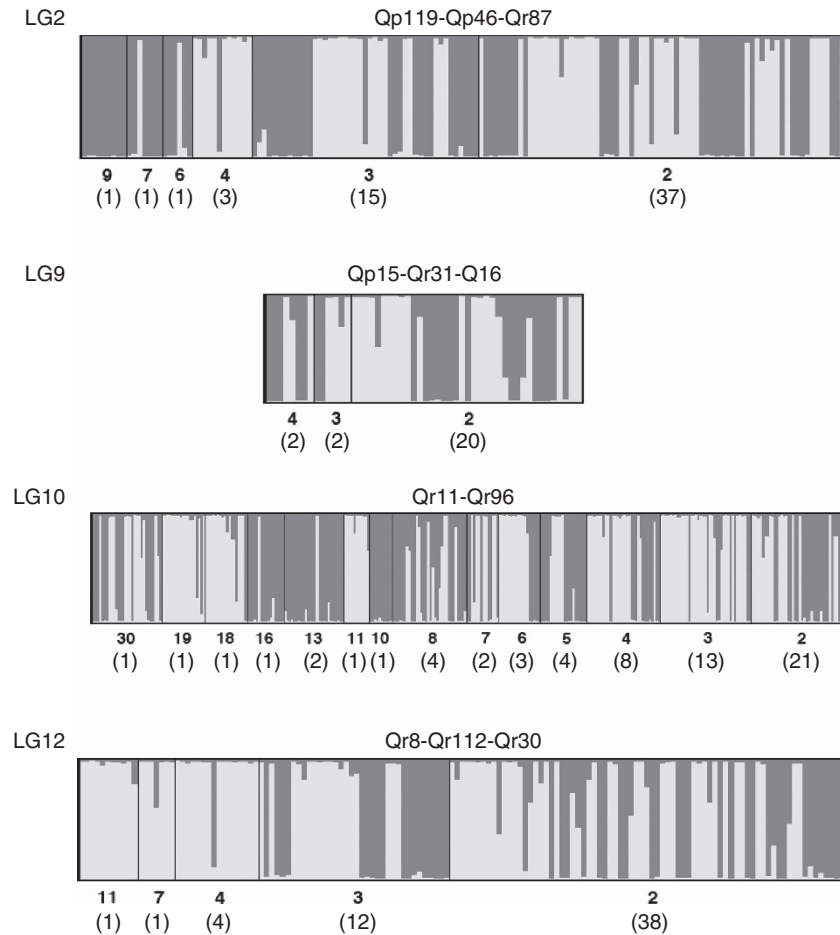


Figure 3 Intra vs interspecific haplotype sharing around chromosome fragments that are probably under selection (LG2, LG10 and LG12) and in the control segment with no selection signals (LG9). Inferred haplotypes are represented by vertical bars (light and dark grays for *Q. petraea* and *Q. robur*, respectively). Haplotype frequencies (in order of decreasing frequency from left to right) and the number of different haplotypes with such frequencies (in brackets) are indicated below the plots.

Table 2 Population genetics recombination estimates (ρ , background recombination) in segments Qp119–Qp46–Qr87 (LG2), Qp15–Qr31–Q16 (LG9), Qr11–Qr96 (LG10) and Qr8–Qr112–Qr30 (LG12)

ρ		$r1^a$	$r2^a$
'Oak priors' ($\mu = 0.04$, $f = 10\,000$)			
LG2 ^b	NA	NA	NA
LG9	0.0337	0.99	0.99
LG10	0.0022	0.73	NA
LG12	0.0014	0.94	0.69

Abbreviations: LG, linkage group; NA, not available.

^aRecombination within each of the two-marker segments ($r1-r2$) is given as a function of the background recombination.

^bBayesian chains failed to converge.

than those typically used in current genome scans, which could create doubts about haplotype reconstructions. However, genetic distances among markers used in the haplotype reconstructions were around 5 cM. Although these distances might be close to the upper limits for haplotype reconstruction, the haplotype probabilities obtained (which depend on the homozygosity of the data) were quite high. This was probably due in part to our decision to reconstruct haplotypes in segments showing strong selection signals and which thus exhibit

reduced heterozygosity. Note that the reduced population recombination rates observed in such segments probably mean that the genetic distances obtained in the controlled crosses progenies were overestimated in relation to the effective recombination rate observed in natural populations. Overall, the clear detection of LD signals, reduced rates of recombination and strong associations with particular species achieved with these haplotypic analyses makes them very promising for the analysis of divergence hitchhiking.

As noted before, outlier detection methods appear to be partly redundant and partly complementary. This suggests the need to use several outlier methods to detect selection signals corresponding to different evolutionary scenarios (Przeworski, 2002; Przeworski *et al.*, 2005; Kim, 2006). For instance, the method based on the loss of rare alleles (ln RV) can probably detect older selection footprints than methods based on heterozygosity differences (F_{DIST2}, BAYESF_{ST}, ln RH), because heterozygosity recovers faster than allelic richness following a bottleneck (Nei *et al.*, 1975). On the other hand, multiple testing increases the risk of false positives (Narum and Hess, 2011). Therefore, we performed three types of tests to further characterize the regions surrounding differentiation outliers: a search for frequent haplotypes with extended LD, a comparison of population genetics recombination estimates in regions that are potentially under selection and in control regions, and score tests for association between frequent haplotypes and species status.

Table 3 LD analyses using haplotypes inferred by Phase

Markers	<i>Q. petraea</i>					<i>Q. robur</i>				
	Haps ^a	Obs ^b	Exp ^b	<i>D'ij</i> ^c	Prob ^d	Haps ^a	Obs ^b	Exp ^b	<i>D'ij</i> ^c	Prob ^d
Qp119-Qp46	13-219	6	2.8	0.20	0.026	10-202	15	14.7	0.01	0.907
	18-199	6	3.3	0.16	0.086	10-194	8	4.3	0.38	0.024
	13-205	4	1.3	0.35	0.009	10-205	5	3.7	0.16	0.388
	22-191	4	1.4	0.39	0.016	10-208	5	2.5	0.46	0.045
	36-191	4	0.9	0.76	0.000	10-193	4	2.5	0.28	0.222
	18-191	0	3.6	-1.00	0.024	17-202	4	2.8	0.16	0.394
						10-191	0	4.0	-1.00	0.013
					10-199	0	2.8	-1.00	0.041	
Qp46-Qr87	199-19	18	4.0	0.70	0.000	202-17	22	11.9	0.34	0.000
	191-22	15	5.4	0.44	0.000	202-26	15	5.1	0.79	0.000
	186-20	9	2.5	0.48	0.000	194-23	9	1.7	0.59	0.000
	219-23	8	1.9	0.47	0.000	198-27	4	0.3	0.65	0.000
	219-22	7	3.8	0.21	0.050	205-24	4	0.9	0.28	0.000
	195-20	6	1.7	0.46	0.000	216-23	4	0.9	0.51	0.000
	191-21	4	3.6	0.02	0.803	202-23	2	5.9	-0.66	0.042
	191-19	0	4.3	-1.00	0.012	191-17	0	3.2	-1.00	0.032
	199-22	0	5.0	-1.00	0.006	194-17	0	3.5	-1.00	0.025
						202-24	0	3.4	-1.00	0.024
Qr11-Qr96 ^e	6-07	16	9.9	0.29	0.009	12-03	16	15.4	0.10	0.746
	12-09	12	10.1	0.11	0.397	7-03	14	14.0	0.01	0.981
	12-03	12	9.3	0.16	0.232	15-03	13	11.2	0.38	0.294
	6-09	9	9.0	0.00	0.991	16-03	11	14.7	-0.25	0.064
	12-08	6	4.7	0.16	0.421	17-03	10	10.5	-0.04	0.784
	6-08	6	4.2	0.21	0.255	18-03	7	7.7	-0.09	0.647
	6-05	4	4.2	-0.04	0.918	8-03	6	4.2	1.00	0.101
	7-07	4	1.2	0.58	0.003	14-03	5	5.6	-0.10	0.647
	12-07	2	11.1	-0.82	0.000	26-03	5	3.5	1.00	0.135
	6-03	3	8.3	-0.64	0.014	27-03	5	3.5	1.00	0.135
						11-03	4	4.9	-0.18	0.458
						19-03	4	4.9	-0.18	0.458
						17-10	4	1.0	0.30	0.001
Qr8-Qr112 ^e	30-13	17	16.4	0.10	0.760	18-13	9	3.7	1.00	0.000
	18-13	14	13.4	0.12	0.745	39-13	9	4.5	0.69	0.005
	31-13	8	6.0	1.00	0.090	31-13	6	4.1	0.32	0.214
	39-13	8	6.7	0.56	0.311	41-13	6	6.2	-0.03	0.918
	19-13	6	4.5	1.00	0.145	42-18	6	1.6	0.59	0.000
	29-13	6	5.2	0.44	0.491	43-13	5	4.1	0.15	0.562
	36-13	6	6.7	-0.11	0.569					
	38-13	5	4.5	0.34	0.618					
	33-13	3	5.2	-0.43	0.047					
	40-13	3	5.2	-0.43	0.047					
Qr112-Qr30 ^f	13-07	48	53.0	-0.26	0.057	18-22	10	3.4	0.422	0.000
	14-07	17	10.2	0.63	0.001	13-15	9	5.8	0.392	0.067
	13-35	8	7.5	0.21	0.687	13-13	9	5.8	0.392	0.067
	13-32	6	4.5	1.00	0.145	13-14	5	4.1	0.149	0.562
	13-26	6	4.5	1.00	0.145	13-22	5	7.8	-0.362	0.159
	13-13	4	3.7	0.21	0.780	16-11	4	0.8	0.448	0.000
	13-27	4	4.5	-0.11	0.646	17-27	4	0.9	0.507	0.000
	12-07	4	5.8	-0.31	0.268					

^aHaplotypes in highest disequilibrium are in bold characters.

^bOnly haplotypes with highest observed (or expected) frequencies are shown.

^cStandardized disequilibrium estimates.

^dChi-square probabilities (exact *P*-values).

^eBiased estimates in *Q. petraea* due to single marker Hardy-Weinberg disequilibrium.

^fBiased estimates in *Q. robur* due to single marker Hardy-Weinberg disequilibrium.

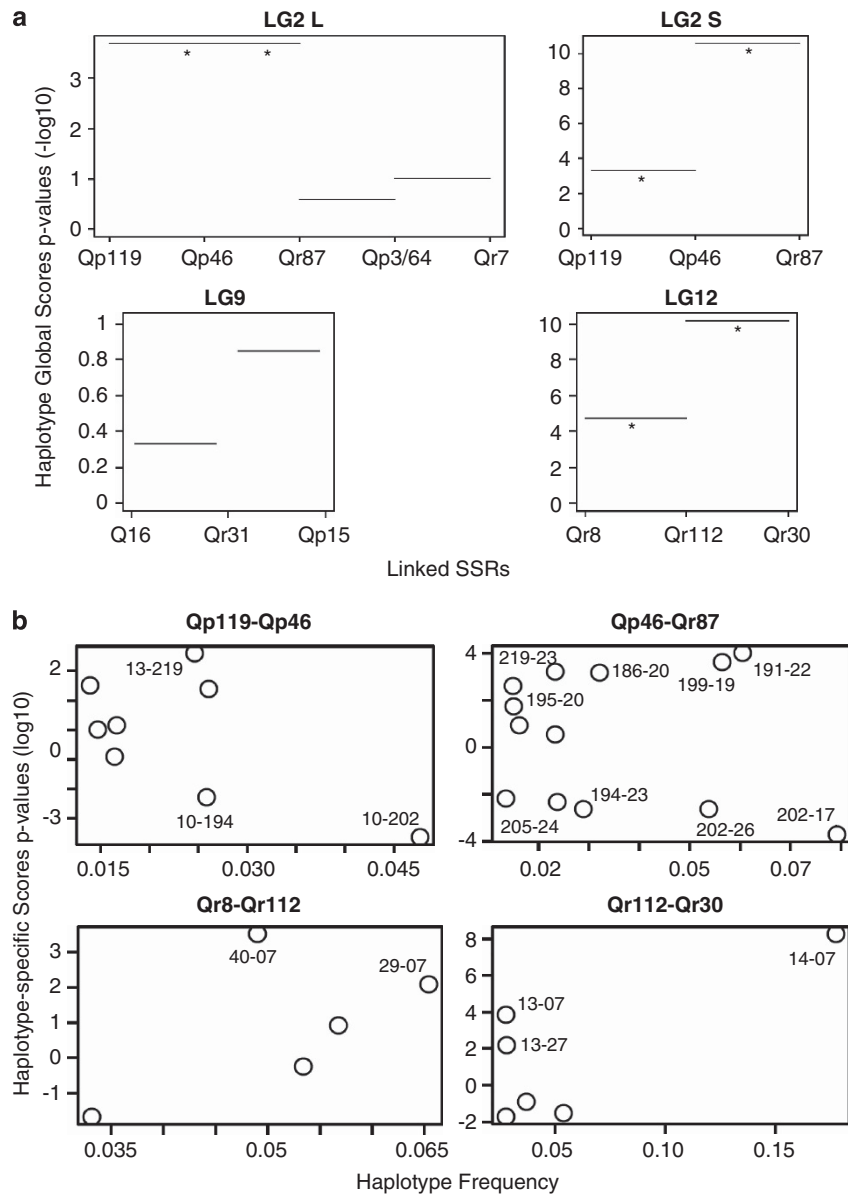


Figure 4 Haplotype score tests. (a) Global haplotype score P -values, on a logarithmic scale, are represented by horizontal bars for contiguous two-marker haplotypes. Significant scores are indicated by asterisks. (b) Logarithmic haplotype-specific scores P -values (represented by circles) are plotted against the haplotype frequencies. Significant scores are labeled with the corresponding haplotype codes. Positive scores indicate association with *Q. petraea* and negative scores with *Q. robur*.

The first of these analyses is essentially a simple long-range LD test. These tests have been used extensively in fine-scale LD mapping and association studies (McPeck and Strahs, 1999; Sabeti *et al.*, 2002; Butty *et al.*, 2007; Curtis *et al.*, 2008), investigations of genome-wide epistatic interactions (Zhang and Liu, 2007) and in gene flow and introgression studies (Koopman *et al.*, 2007). A direct comparison of LD among segments with and without outliers was not possible due to large Hardy–Weinberg disequilibrium at two of the three linked loci in the control region. Nevertheless, our recombination estimates indicate that the LD should be lower in the control region than in regions of similar length containing outlier loci. One limitation of our analyses is the potential bias in LD estimates due to uncertainties in haplotype reconstructions. However, the agreement between the two haplotype reconstruction methods, the large probabilities for the two-marker haplotypes inferred using reconstructions based on the

coalescent and the uniqueness of most low-probability haplotypes should considerably reduce this risk. On the other hand, LD could be increased by several biological and historical factors besides divergent selection, including admixture (Gaut and Long, 2003). However, we do not believe that the sampling of different populations from each species presents a problem in our study because: first, high-frequency haplotypes were not exclusive to one population (data not shown), and second, the two temperate European oaks exhibit little differentiation among populations (Zanetto *et al.*, 1993).

The second test estimates the population genetics recombination parameter (ρ), which is expected to decrease under divergence hitchhiking (Via, 2009). We are not aware of any previous use of this parameter in such a context, but it seems to be a very promising tool for detecting reduced rates of recombination around the genes that are responding to divergent selection. Although there seems to be

a need for further fine-tuning before this approach can be generalized, our estimates indicate that regions around putatively selected genes do indeed exhibit reduced effective recombination relative to control regions.

The third approach used score tests to identify associations between frequent haplotypes and species. Score tests have been widely used for haplotype association studies in humans (North *et al.*, 2006; Li and Schaid, 2009). In principle, these tests could be sensitive to existing interspecific genetic structure. However, the association score tests were in good agreement with both the haplotypic LD estimates (LG2, LG10) and the outlier tests (LG12). Moreover, the differences revealed by the sliding-window analyses of LG2 suggest that these tests have enough sensitivity to differentiate between putatively selected and neutral regions located in the same linkage group. However, it remains to be determined whether H–W disequilibrium could affect the conclusions obtained with these tests (Schaid *et al.*, 2002).

Advantages of haplotypic approaches for studying divergent selection

As noted previously, the search for ecological speciation footprints is one of the main causes for the increase in the number of studies dealing with divergent selection and outlier tests. Ascertaining the actual environmental causes leading to divergent selection is one of the major challenges in such studies. This has been partly solved with the application of several indirect methods, such as co-location with QTLs (Rogers and Bernatchez, 2005) or the search of parallel outliers in replicate pairs of populations (Bonin *et al.*, 2006). However, these methods do not solve another problem arising from allele frequencies-based outlier methods, that is, whether outliers are caused by directional or divergent (bi-directional) selection. One major advantage of haplotypic LD and score tests is that they can detect selective pressures affecting either both species in question or only one of them (Figure 5). Conversely, allele frequencies-based differentiation outlier methods can only detect selection in the species with the lower heterozygosity (F_{DIST2} , $\ln RH$), variance in number of repeats ($\ln RV$) or migration rate (BAYESFST). Our results based on haplotype reconstruction point to simultaneous divergent (that is,

bi-directional) selection in at least two out of the three linkage groups harboring outlier markers, suggesting that both oak species are evolving according to their own trajectories, rather than there being a situation in which only one of the two species is actively diverging from the other. Such insights into the selection type (that is, divergent vs directional) can only be obtained by the use of haplotype-based methods. If used more broadly, these methods could greatly improve the characterization of divergent selection in nature.

Overall, our results indicate that haplotypic approaches may be uniquely powerful in studies of divergent selection. As such, they should now be adapted and scaled-up to analyze larger data sets (for example, Galindo *et al.*, 2010; Baxter *et al.*, 2011; Hohenlohe *et al.*, 2011). Higher marker density will improve the probabilities of the inferred haplotypes, thereby further increasing the confidence in haplotypic analyses.

DATA ARCHIVING

Full genotypes and metadata, together with Supplementary Information files, are stored at the Dryad repository (provisional 10.5061/dryad.099s2).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was conducted with the financial support from the Commission of the European Communities, via the ‘Quality of life and Management of Living resources’ research Program (Project OAKFLOW QLK5-2000-00960). RJP was supported by the LINKTREE project from the Eragnet Biodiversa programme (ANR-08-BDVA-006). We are indebted to project partners who provided samples (J Buiteveld, K Burg, J Cottrell, L Gil, F Gugerli, A König, M Lascoux, A Lowe), to EM Gillet and F Raspail for compiling updated versions of GSED and PERMUT, and to SC González-Martínez and P Garnier-Géré for critical reading and helpful discussions. Three anonymous referees are also acknowledged for criticism that improved the manuscript. PGG is thankful to the Pierroton group for their warm welcome during a half-sabbatical year that triggered the ideas developed in this study. PGG, RJP and AK are supported by the EU network of excellence EVOLTREE.

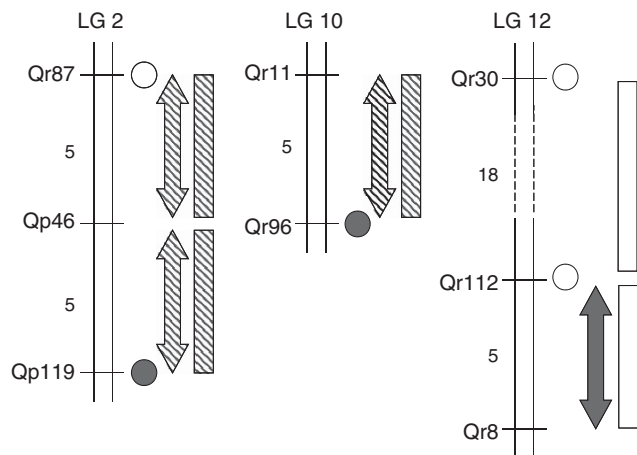


Figure 5 Comparison among selection signals detected by single-locus methods (differentiation outlier tests, circles) and methods that account for haplotypic information (LD, arrows; Score Tests, rectangles). The symbol patterns indicate whether selection signals were detected in both species (hatched) or in only one of them (empty, *Q. petraea*; filled, *Q. robur*). Approximate distances (cM) between markers are given to the left of the linkage groups.

Abramoff MD, Magelhae PJ, Ram SJ (2004). Image processing with ImageJ. *Biophoton Int* **11**: 36–42.

Akey J, Lin J, Xiong M (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9**: 291–300.

Apple JL, Grace T, Joern A, Amand PS, Wisely SM (2010). Comparative genome scan detects host-related divergent selection in the grasshopper *Hesperotettix viridis*. *Mol Ecol* **19**: 4012–4028.

Armitage P (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375–386.

Bader JS (2001). The relative power of SNPs and haplotypes as genetic markers for association tests. *Pharmacogenomics* **2**: 11–24.

Barreneche T, Bodénès C, Lexer C, Trontin JF, Fluch S, Streiff R *et al.* (1998). A genetic linkage map of *Quercus robur* L (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5S rDNA markers. *Theor Appl Genet* **97**: 1090–1103.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD *et al.* (2011). Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**: e19315.

Beaumont MA, Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**: 969–980.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond B Biol Sci* **263**: 1619–1626.

Bierne N (2010). The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution* **64**: 3254–3272.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006). Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol* **23**: 773–783.

- Bowcock AM, Kidd JR, Mountain JL, Herbert JM, Carotenuto L, Kidd KK *et al.* (1991). Drift, admixture and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci USA* **88**: 839–843.
- Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783–796.
- Butty V, Roy M, Sabeti P, Besse W, Benoist C, Mathis D (2007). Signatures of strong population differentiation shape extended haplotypes across the human *CD28*, *CTLA4*, and *ICOS* costimulatory genes. *Proc Natl Acad Sci USA* **104**: 570–575.
- Clark AG (2004). The role of haplotypes in candidate gene studies. *Genet Epidemiol* **27**: 321–333.
- Crawford D, Bhargava T, Li N, Hellenthal G, Rieder M, Nickerson D *et al.* (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* **36**: 700–706.
- Crawford NG (2010). SMOGD: a software for the measurement of genetic diversity. *Mol Ecol Res* **10**: 556–557.
- Curtis D, Vine AE, Knight J (2008). Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann Hum Genet* **72**: 261–278.
- Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**: 997–1004.
- Dow BD, Ashley MV, Howe HF (1995). Characterization of highly variable (GA/CT)_n microsatellites in the bur oak, *Quercus macrocarpa*. *Theor Appl Genet* **91**: 137–141.
- Egan SP, Nosil P, Funk DJ (2008). Selection and genomic differentiation during ecological speciation: isolating the contributions of hos-association via a comparative genome scan of *Neochlamisus bebbianae* leaf beetles. *Evolution* **62**: 1162–1181.
- El Mousadik A, Petit RJ (1996). High level of genetic differentiation for allelic richness among populations of the argan tree (*Argania spinosa* L. Skeels) endemic of Morocco. *Theor Appl Genet* **92**: 832–839.
- Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* **14**: 2611–2620.
- Excoffier L, Laval G, Balding D (2003). Gametic phase estimation over large genomic regions using and adaptive window approach. *Hum Genomics* **1**: 7–19.
- Excoffier L, Lischer HEL (2010). Arlequin suite ver. 3.5: a new series of programs to perform population genetic analysis under Linux and Windows. *Mol Ecol Res* **10**: 564–567.
- Excoffier L, Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921–927.
- 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.
- Galindo J, Grahame JW, Butlin RK (2010). An EST based genome scan using 454 sequencing in the marine snail *Littorina saxatiles*. *J Evol Biol* **23**: 2004–2016.
- Gaut BS, Long AD (2003). The lowdown on linkage disequilibrium. *Plant Cell* **15**: 1502–1506.
- Goudet J (2001). Fstat, a program to estimate and test gene diversities and fixation indices (v. 2.9.3). Available from <http://www2.unil.ch/popgen/softwares/fstat.htm>. Updated from Goudet (1995).
- Goudet J, Raymond M, de Meeüs T, Rousset F (1996). Testing differentiation in diploid populations. *Genetics* **144**: 1933–1940.
- Guo S, Thompson E (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**: 361–372.
- Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011). Next-generation RAD sequencing identifies thousand of SNPs for assessing hybridization between rainbow and wastlope cutthroat trout. *Mol Ecol Res* **11**: 117–122.
- Hudson RR, Bailey K, Skarecky D, Kwiatkowski J, Ayala FJ (1994). Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**: 1329–1340.
- Jakobsson M, Rosenberg NA (2006). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801–1806.
- Jaramillo-Correa JP, Verdú M, González-Martínez SC (2010). The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evol Biol* **10**: 22.
- Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G *et al.* (2007). A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol Ecol* **16**: 3955–3969.
- Jost L (2008). G_{ST} and its relatives do not measure differentiation. *Mol Ecol* **17**: 4015–4026.
- Kampfer S, Lexer C, Glössl J, Steinkellner H (1998). Characterization of (GA)_n microsatellite loci from *Quercus robur*. *Hereditas* **129**: 183–186.
- Kauer M, Dieringer D, Schlotterer C (2003). A microsatellite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*. *Genetics* **165**: 1137–1148.
- Kim Y (2006). Allele frequency distribution after recurrent selective sweeps. *Genetics* **172**: 1967–1978.
- Kim Y, Nielsen R (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1512–1524.
- Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
- Kohn M, Pelz HJ, Wayne RK (2000). Natural selection mapping of the warfarin-resistance gene. *Proc Natl Acad Sci USA* **97**: 7911–7915.
- Koopman WJM, Li S, Coart E, Van de Weg E, Vosman B, Roldán-Ruiz I *et al.* (2007). Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. *Mol Ecol* **16**: 243–256.
- Kremer A, Dupouey JL, Deans JD, Cottrell J, Csaikl U, Finkeldey R *et al.* (2002). Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Ann For Sci* **59**: 777–787.
- Lepais O, Gerber S (2011). Reproductive patterns shape introgression dynamics and species succession within the European white oaks complex. *Evolution* **65**: 156–170.
- Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, Kremer A *et al.* (2009). Species relative abundance and direction of introgression in oaks. *Mol Ecol* **18**: 2228–2242.
- Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–195.
- Li N, Stephens M (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.
- Li WY, Schaid DJ (2009). Power comparison between similarity-based multilocus association methods, logistic regression and score tests for haplotypes. *Genet Epidemiol* **33**: 183–197.
- Liu N, Zhang K, Zhao H (2008). Haplotype-association analysis. *Adv Genet* **60**: 335–405.
- Mariette S, Cottrell J, Csaikl UM, Goicoechea PG, König A, Lowe AJ *et al.* (2002). Comparison of levels of 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.
- Maynard-Smith JM, Haigh J (1974). The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23–35.
- McPeck MS, Strahs A (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* **65**: 858–875.
- McVean G (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**: 1395–1406.
- Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010). Widespread genomic divergence during sympatric speciation. *Proc Natl Acad Sci USA* **107**: 9724–9729.
- Muir G, Schlotterer C (2005). Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus spp.*). *Mol Ecol* **14**: 549–561.
- Muller CH (1952). Ecological control of hybridization in *Quercus*: a factor in the mechanism of evolution. *Evolution* **6**: 147–161.
- Murray MC, Hare MP (2006). A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. *Mol Ecol* **15**: 4229–4242.
- Narum S, Hess JE (2011). Comparison of F_{ST} outlier tests for SNP loci under selection. *Mol Ecol Res* **11**: 184–194.
- Nei M (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* **70**: 3321–3323.
- Nei M, Maruyama T, Chakraborty R (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**: 1–10.
- Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H *et al.* (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**: 1289–1299.
- North BV, Sham PC, Knight J, Martin ER, Curtis D (2006). Investigation of the ability of haplotype association and logistic regression to identify associated susceptibility loci. *Ann Hum Genet* **70**: 893–906.
- Nosil P, Egan SP, Funk DJ (2007). Heterogeneous genomic differentiation between walking-stick ecotypes: 'isolation by adaptation' and multiple roles for divergent selection. *Evolution* **62**: 316–336.
- Nosil P, Feder JL (2012). Genomic divergence during speciation: causes and consequences. *Philos Trans R Soc Lond B Biol Sci* **367**: 332–342.
- Nosil P, Funk DJ, Ortiz-Barrientos D (2009). Divergent selection and heterogeneous genomic divergence. *Mol Ecol* **18**: 375–402.
- Ohta T, Kimura M (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res* **22**: 201–204.
- Parelle J, Brendel O, Jolivet Y, Dreyer E (2007). Intra- and interspecific diversity in the response to waterlogging of two co-occurring white oak species (*Quercus robur* and *Q. petraea*). *Tree Physiol* **27**: 1027–1034.
- Petit RJ, Bodénès C, Ducousso A, Rousset G, Kremer A (2004). Hybridization as a mechanism of invasion in oaks. *New Phytol* **161**: 151–164.
- Poncet BN, Herrmann D, Gugerli F, Taberlet P, Holderegger R, Gielly L *et al.* (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Mol Ecol* **19**: 2896–2907.
- Ponton S, Dupouey JL, Bréda N, Dreyer E (2002). Comparison of water-use efficiency of seedlings from two sympatric oak species: genotype x environment interactions. *Tree Physiol* **22**: 413–422.
- Prida A, Ducousso A, Puech J-L, Petit RJ, Nepveu G (2007). Variation in wood volatile compounds in a mixed oak stand: strong species and spatial differentiation in whisky-lactone content. *Ann For Sci* **64**: 313–320.
- Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Przeworski M (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160**: 1179–1189.
- Przeworski M, Coop G, Wall JD (2005). The signature of positive selection on standing genetic variation. *Evolution* **59**: 2312–2323.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0., <http://www.R-project.org>.

- Raymond M, Rousset F (1995). An exact test for population differentiation. *Evolution* **49**: 1280–1283.
- Rogers SM, Bernatchez L (2005). Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Mol Ecol* **14**: 351–361.
- Rosenberg NA (2004). Distruct: a program for the graphical display of population structure. *Mol Ecol Notes* **4**: 137–138.
- Sabeti PC, Reich DE, Higgins JM, Haninah Z, Levine P, Richter DJ *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.
- Savolainen V, Anstett M-C, Lexer C, Hutton I, Clarkson JJ, Norup MV *et al.* (2006). Sympatric speciation in palms in an oceanic island. *Nature* **441**: 2010–2013.
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**: 425–434.
- Schlötterer C (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**: 753–763.
- Schluter D (2000). *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford, UK.
- Scotti-Saintagne C, Mariette S, Porth I, Goicoechea PG, Barreneche T, Bodénès C *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.
- Sinnwell JP, Schaid DJ (2009). haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.44. http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm.
- Smadja G, Galindo J, Butlin R (2008). Hitching a lift on the road to speciation. *Mol Ecol* **17**: 4177–4180.
- Steinkellner H, Fluch S, Turetschek E, Lexer C, Streiff R, Kremer A *et al.* (1997). Identification and characterization of (GA/CT)_n microsatellite loci from *Quercus petraea*. *Plant Mol Biol* **33**: 1093–1096.
- Stephens M, Donnelly P (2003). A comparison of Bayesian methods for haplotype reconstruction. *Am J Hum Genet* **73**: 1162–1169.
- Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978–989.
- Streiff R, Ducouso A, Lexer C, Steinkellner H, Glössl J, Kremer A (1999). Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L and *Quercus petraea* (Matt.) Liebl. *Mol Ecol* **8**: 831–841.
- Via S (2009). Natural selection in action during speciation. *Proc Natl Acad Sci USA* **106**:suppl. 1 9939–9946.
- Via S (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation with gene flow. *Philos Trans R Soc Lond B Biol Sci* **367**: 451–460.
- Via S, West J (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol Ecol* **17**: 4334–4345.
- Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
- Zanetto A, Kremer A, Labbé T (1993). Differences of genetic variation based on isozymes of primary and secondary metabolism in *Quercus petraea*. *Ann For Sci* **50**: 245–252.
- Zeng YF, Liao WJ, Petit RJ, Zhang D-Y (2010). Exploring species limits in two closely related Chinese oaks. *PLoS One* **5**: e15529.
- Zhang Y, Liu JS (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* **39**: 1167–1173.
- Zoldo V, Pape D, Brown SC, Panaud O, Iljak-Yakovlev S (1988). Genome size and base composition of several *Quercus* species: inter- and intra-population variation. *Genome* **41**: 162–168.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)