Genome Scanning for Interspecific Differentiation Between Two Closely Related Oak Species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]

Caroline Scotti-Saintagne,* Stéphanie Mariette,*,1 Ilga Porth,† Pablo G. Goicoechea,‡ Teresa Barreneche,*,2 Catherine Bodénès,* Kornel Burg† and Antoine Kremer*,3

*INRA, UMR Biodiversité Gènes et Ecosystèmes, 33612 Cestas Cedex, France, †ARC Seibersdorf Research, Biotechnology Unit, A-2444 Seibersdorf, Austria and ‡NEIKER, Dpto Biotecnologia, 01080 Vitoria-Gasteiz, Spain

ABSTRACT

Interspecific differentiation values (G_{ST}) between two closely related oak species (*Quercus petraea* and *Q. robur*) were compiled across different studies with the aim to explore the distribution of differentiation at the genome level. The study was based on a total set of 389 markers (isozymes, AFLPs, SCARs, microsatellites, and SNPs) for which allelic frequencies were estimated in pairs of populations sampled throughout the sympatric distribution of the two species. The overall distribution of G_{ST} values followed an L-shaped curve with most markers exhibiting low species differentiation ($G_{ST} < 0.01$) and only a few loci reaching >10% levels. Twelve percent of the loci exhibited significant G_{ST} deviations to neutral expectations, suggesting that selection contributed to species divergence. Coding regions expressed higher differentiation than noncoding regions. Among the 389 markers, 158 could be mapped on the 12 linkage groups of the existing *Q. robur* genetic map. Outlier loci with large G_{ST} values were distributed over 9 linkage groups. One cluster of three outlier loci was found within 0.51 cM; but significant autocorrelation of G_{ST} was observed at distances <2 cM. The size and distribution of genomic regions involved in species divergence are discussed in reference to hitchhiking effects and disruptive selection.

U NDERSTANDING speciation remains one of the fundamental problems in biology. The predominant view is that new species arise most often in allopatry where geographically isolated populations of the same ancestral species diverge progressively (MAYR 1963). However, this view has been challenged by both empirical results and theoretical investigations. Indeed, sympatric speciation events have been observed in controlled experiments (RICE and HOSTERT 1993; RUNDLE 2002), under natural conditions (SCHLIEWEN *et al.* 2001), and also demonstrated by simulation models (DIECKMANN and DOEBELI 1999; KONDRASHOV and KONDRASHOV 1999). In the simplest scenario, sympatric speciation occurs when disruptive selection favors two extreme phenotypes. Accordingly, the intermediary individuals that are less adapted are eliminated and progressively reproductive isolation is established between the two extreme phenotypes.

With the big advances in genetic and molecular analysis in the last decade the main issue has now moved to the divergence between species at the genome level. In

his view of speciation, WU (2001) presents genes as the units of species differentiation. This opinion contradicts the biological species concept (MAYR 1963), which assumes a highly coadaptive genetic architecture leading to whole-genome isolation. In Wu's model, the maintenance of two sympatric and interfertile species will translate at the genome level to a mosaic of impermeable and permeable regions to gene flow. Impermeable regions accumulate divergence in response to selection whereas permeable regions share introgressed genes that decrease differentiation in these regions. Genomic differentiation between closely related species has been investigated by comparing positions of markers on genetic maps of the parental species and their hybrids (RIESEBERG *et al.* 2000) and analyzing the distribution of quantitative trait loci (QTL) of traits exhibiting interspecific phenotypic differentiation (ORR 2001).

The two predominant European oaks, pedunculate (*Quercus robur* L.) and sessile oak [*Q. petraea* (Matt.) Liebl.], are an interesting model to study interspecific differentiation at a genome level. The two species are interfertile and cohabit in most European forests despite their soil preferences. *Q. robur* is more frequent on soils with high nutrient availability and *Q. petraea* occupies drier sites (LÉVY *et al.* 1992). However, in most situations the two species cohabit in the same stands along a gradient of water availability and soil fertility. Thanks also to the extensive pollen flow (STREIFF *et al.* 1998), natural hybridization between both species is

quite frequent (BACILIERI *et al.* 1996). In spite of interspecific gene flow, strong phenotypic differences are maintained for leaf morphological and ecophysiological traits (KREMER *et al.* 2002). A recent study indicated that QTL controlling leaf morphology were distributed all over the genome with, however, two clusters on two linkage groups (SAINTAGNE *et al.* 2004). These phenotypic differences contradict earlier reports based on molecular data. With only a few exceptions (GÖMÖRY *et al.* 2001) all genetic surveys conducted with different markers indicated extremely low species differentiation (PETIT and KREMER 1993; BARRENECHE *et al.* 1996; COART *et al.* 2002; MARIETTE *et al.* 2002). The genetic homogeneity between the two species was well illustrated in the study of BODÉNÈS *et al.* (1997a), where among 2800 RAPD fragments only 2% displayed a significant allelic frequency difference. Rare outlier markers exhibiting large species differences were found for isozymes by GÖMÖRY (2000), for sequence-characterized amplified regions (SCARs; PARAN and MICHELMORE 1993) by BODÉNÈS *et al.* (1997b), and for amplified fragment length polymorphisms (AFLPs; VOS *et al.* 1995) by COART *et al.* (2002) and MARIETTE *et al.* (2002).

Scanning the genome for genetic diversity has been suggested as a method to detect molecular signatures of natural selection and has actually been implemented in Drosophila (HAMBLIN and AQUADRO 1999), in goatgrasses (DVORAK *et al.* 1998), in maize (TENAILLON *et al.* 2001), in sorghum (DRAYE *et al.* 2001), and in tomato (STEPHAN and LANGLEY 1998). The rationale of the method is based on the observation that a beneficial mutation in a coding region of the genomes leads to a selection sweep at the selected locus and the sweep extends to the flanking regions due to linkage (SCHLÖTTERER 2003). We extended here the concept of genomic scanning to genetic differentiation, as was recently implemented in humans (AKEY *et al.* 2002). The reason is that for an outcrossing species, gametic disequilibrium is more pronounced at an interpopulational level than at the within-population level, the latter being constantly eroded by recombination due to random mating (LE CORRE and KREMER 2003). As species differences between the two sympatric *Q. petraea* and *Q. robur* are suspected to be generated by divergent selection (PETIT *et al.* 2003), species differentiation can actually be analyzed in a similar way to population divergence due to diversifying selection toward different optima (LE CORRE and KREMER 2003). These authors found that population differentiation due to diversifying selection creates large between-population disequilibria between loci involved in traits submitted to selection, whether these loci are linked or not. Hence we would expect that a systematic scanning of the genome for interspecific differentiation would decipher the "molecular architecture" of species divergence. In this study we assembled results from previous genetic surveys conducted in these two oak species during the past 10 years, with various marker techniques. We completed these surveys by additional molecular screening to obtain a large data set of interspecific differentiation between the two species (interspecific G_{ST} values; NEI 1987). These G_{ST} values were then plotted along the linkage groups of an oak genetic map (BARRENECHE *et al.* 1998). The aim of this study was, by combining population and mapping studies, to describe the genomic arrangement of species differentiation between the two closely related oak species, *Q. petraea* and *Q. robur*.

MATERIALS AND METHODS

Mapping population: The mapping pedigree is a F_1 full-sib cross of *Q. robur* (3PxA4) composed of 278 full-sibs. The male parent originated from Arcachon (latitude 44.40N, longitude 1.11W) and the female parent was located at the Forestry Research Station of Pierroton (latitude 44.44N, longitude 0.46W).

A subset of 94 full-sibs was used for the construction of a *Q. robur* genetic map (BARRENECHE *et al.* 1998) and the entire set (278) was used to build a framework map for the QTL mapping (SAINTAGNE *et al.* 2004).

Sampling of natural populations and markers: Genetic differentiation between *Q. petraea* and *Q. robur* was estimated in three genetic surveys that were conducted with different genetic markers during the past 10 years (Table 2). Allele frequencies were assessed in populations of each species that were clustered in pairs throughout the natural range of the two species. The sampling procedure (the location and composition of pairs) and the markers used varied among the three surveys.

1. The first survey was based on isozymes (ZANETTO *et al.* 1994) and comprised seven pairs of populations.
2. The second survey was conducted with SCARs and comprised 8 pairs of populations (BODÉNÈS *et al.* 1997b). Five pairs of populations of the second survey were also used for a microsatellite diversity analysis (MUIR *et al.* 2000).
3. The third survey was conducted with microsatellites, AFLPs, and single-nucleotide polymorphisms (SNPs) and comprised 10 pairs of populations. Seven pairs of populations were already analyzed by MARIETTE *et al.* (2002) to estimate genetic diversity and differentiation for six microsatellites and for four *Pst*I-*Mse*I primer-enzyme AFLP combinations. For this study the sampling was extended to the 10 pairs and the molecular analyses were done with AFLPs (eight *Eco*RI-*Mse*I and three *Pst*I-*Mse*I primer enzyme combinations), additional microsatellites, and SNPs (Tables 1 and 3).

In the first two surveys, acorns were collected in reportedly pure stands of *Q. petraea* and *Q. robur* that were separated by <150 km within each pair (except one pair in Scandinavia, where the two stands were ~500 km apart). Acorns were sown in the nursery, and isozymes or DNA were extracted from tissues collected on the seedlings. In the third survey, the data originated from adult trees sampled in the forest and not from their offspring raised in the nursery. Populations were also sampled in pairs and a pair consisted of a continuous stand comprising the two species (MARIETTE *et al.* 2002). However, for one pair (Scandinavia) the two stands were geographically distant (Table 1). A multivariate analysis of leaf morphology permitted us to assign the species name to each tree and trees with intermediate morphology were excluded from the sample (KREMER *et al.* 2002).

Molecular analysis: Details of the protocols for the molecu-

TABLE 1

Location and composition of stands used in the third genetic survey

Country	Name	Composition	No. of trees genotyped/species
France	Petite Charnie	*Q. petraea* and *Q. robur*	10
The Netherlands	De Meinweg	*Q. petraea* and *Q. robur*	10
Spain	Arlaban	*Q. petraea* and *Q. robur*	10
United Kingdom	Dalkeith	*Q. petraea* and *Q. robur*	10
Austria	Sigmundsherdberg	*Q. petraea* and *Q. robur*	10
United Kingdom	Roudsea Wood	*Q. petraea* and *Q. robur*	10
Switzerland	Büren	*Q. petraea* and *Q. robur*	10
Germany	Escherode	*Q. petraea* and *Q. robur*	10
Hungary	Sopron	*Q. petraea* and *Q. robur*	10
Denmark/Sweden	Hald Ege/Gysinge[a]	*Q. petraea*/*Q. robur*	10/10

[a] Hald Ege is a pure *Q. petraea* stand and Gysinge is a pure *Q. robur* stand.

lar analysis are given in the different references of the previous articles (Table 2). We give here only the protocols for the molecular analyses realized to obtain the unpublished third survey data. The protocol to analyze *Pst*I/*Mse*I combinations of AFLPs in oaks was described in GERBER *et al.* (2000). For this study we extended the protocol to *Eco*RI/*Mse*I combinations with no major modification. The 50- to 700-bp sizing standard marker (LI-COR, Biotechnology Division) was employed to determine the size of fragments and the Saga LI-COR software was used for scoring the AFLP fragments. The development of the 38 microsatellites was described in STEIN-KELLNER *et al.* (1997) and in KAMPFER *et al.* (1998). Protocols for amplification and separation of microsatellites were described in STREIFF *et al.* (1998). Thirty-eight SNPs were developed from 14 expressed sequence tags (ESTs) differentially expressed between the two species in response to an osmotic stress as described by I. PORTH, C. SCOTTI-SAINTAGNE, A. KREMER, P. SCHUSTER, E. HEBERLE-BORS and K. BURG (unpublished results). Among the 38 SNPs, 23 were scored in the natural populations for estimating G_{ST}, among which 3 were also mapped in the segregating full-sib family (Table 3).

Estimation of the interspecific genetic differentiation: For all types of markers (dominant and codominant) the genetic differentiation between sessile and pedunculate oak populations was estimated by Nei's coefficient of genetic differentiation, G_{ST} (NEI 1987). The data were bulked over all populations within a given species to obtain allelic frequencies for each species. The species sample size on which the allele frequencies were calculated varied between 50 and 1190 per species (Table 2).

Dominant markers: For estimating genetic differentiation at a single locus for AFLP markers, the allele frequency of the null allele was derived from the frequency of phenotypes that did not exhibit a band by using a second-order Taylor expansion (MARIETTE *et al.* 2001). To avoid biases due to the low frequency of null alleles and low sample sizes, the calculation of G_{ST} was restricted to the fragments with an observed frequency $<1–3/n_i$ [n_i is the sample size of species i following Lynch and Milligan's (LM) recommendation (LYNCH AND MILLIGAN 1994)]. Chi-square tests based on the presence and absence of bands and not on the alleles controlling the expression of bands were used to test for frequency differences between the two species. The analysis was performed by using the HAPDOM computer program (Antoine Kremer, INRA-UMR BIOGECO, Cestas, France). In total G_{ST} values were computed on 167 *Eco*RI/*Mse*I-AFLP markers and 136 *Pst*I/*Mse*I-AFLP markers (Table 3).

Codominant markers: For all isozymes (ZANETTO *et al.* 1994), SCARs (BODÉNÈS *et al.* 1997b), microsatellites (MUIR *et al.*

TABLE 2

Sample sizes and distribution of populations and species

Marker type (no. of loci)	No. of pairs	Average sample size (no. trees/species/pair)	No. of trees/ species	Geographic distribution of pairs	Reference
Isozymes (12)	7	120	840	Range wide	ZANETTO *et al.* (1994)
SCARs (13)	8	45	360	Range wide	BODÉNÈS *et al.* (1997b)
Microsatellites (2)	5	16	80	Western part	MUIR *et al.* (2000)
Microsatellites (6)	7	170	1190	Western part	MARIETTE *et al.* (2002)
Microsatellites (30)	10	5	50	Western part	This study
SNPs (23)	10	10	100	Western part	I. PORTH, C. SCOTTI-SAINTAGNE, A. KREMER, P. SCHUSTER, E. HEBERLE-BORS and K. BURG (unpublished results)
AFLPs (107)	7	45	345	Western part	MARIETTE *et al.* (2002)
AFLPs (196)	10	10	100	Western part	This study

TABLE 3

Number of markers used for estimating G_{ST} and for mapping purposes

Markers	No. of markers with G_{ST} value	No. of outlier markers (%)[a]	No. of markers with G_{ST} values and mapped
*Eco*RI/*Mse*I-AFLPs	167	13 (8)	77
*Pst*I/*Mse*I-AFLPs	136	14 (10)	29
Microsatellites	38	5 (13)	37
SNPs	23	5 (22)	3
SCARs	13	7 (53)	6
Isozymes	12	3 (25)	6
Total no.	389	47 (12)	158

[a] Outlier loci were identified according to BEAUMONT and NICHOLS (1996; see text).

2000; MARIETTE *et al.* 2002; this study), and SNPs (I. PORTH, C. SCOTTI-SAINTAGNE, A. KREMER, P. SCHUSTER, E. HEBERLE-BORS and K. BURG, unpublished results), interspecific G_{ST} values were estimated following NEI (1987, p. 191).

Statistical tests for species differentiation were done by using two methods depending on the markers and the data sets. For large sample sizes of codominant markers (isozymes, SCARs, and microsatellites analyzed by MARIETTE *et al.* 2002) chi-square and *G*-tests were used for testing allele frequency differences. For the remaining microsatellites (MUIR *et al.* 2000 and this study) and SNPs, the species differentiation was tested by permuting genotypes among the two species.

Finally, the overall survey of interspecific differentiation was based on 389 markers, combining dominant and codominant markers (Table 3). We compared the distribution of the G_{ST} values over all loci to their expectation under the neutral assumption. BEAUMONT and NICHOLS (1996) have shown that the distribution F_{ST} as a function of heterozygosity in the context of an island model is quite robust to a wide range of conditions (population structure, demographic structure, mutations level). We applied this method (with the infinite allele model) to identify markers deviating from the null hypothesis of neutral evolution. All G_{ST}'s were first transformed to F_{ST} values by using the COCKERHAM and WEIR (1987) transformation [$F_{ST} = nG_{ST}/(G_{ST} + n - 1)$, where n is the number of populations] and F_{ST}'s were plotted as a function of expected heterozygosities. The analysis was done in a two-step procedure. The first envelope of neutral expectation was based on the overall mean value of F_{ST}. Markers with F_{ST} values outside the 95% envelope (corresponding to the null hypothesis) were then removed and a new analysis was done on the basis of the mean value of F_{ST}. Markers with F_{ST} values outside the 95% envelope after the second analysis were considered as outliers. Calculations were done using the Fdist2 program (BEAUMONT and NICHOLS 1996).

Linkage map for G_{ST} mapping and distribution of markers along the linkage groups: We constructed a particular genetic map dedicated to our objective. This map, called "G_{ST} map," was assembled by using marker locations from two previous maps built with the same pedigree: (1) the saturated map of *Q. robur* (BARRENECHE *et al.* 1998) that was constructed with a sample of 94 individuals from the full-sib family, which in its last updated version comprised >600 markers, and (2) the "QTL map" (SAINTAGNE *et al.* 2004) based on all 278 offspring, which comprised only 128 markers. The density of markers was higher in the former than in the latter map; however, the precision of marker location was better in the latter than in the former because of the differences in sample size. And last, markers that were polymorphic in natural populations, for

which G_{ST} values were available, could not be mapped when they were not variable between the two parents of the mapping pedigree. To cope with these constraints, we constructed the so-called G_{ST} map with the main aim to locate as precisely as possible the different markers and to map as many as possible markers for which G_{ST} values were available.

The G_{ST} map was constructed in two steps. The first step consisted of constructing a map comprising the markers scored on the 278 full-sibs, *e.g.*, having the most precise location. The construction was done according to the pseudo-testcross strategy (GRATTAPAGLIA *et al.* 1995). Analysis of linkage among loci was carried out with JOINMAP version 3.0 (VAN OOIJEN and VOORRIPS 1993) using the LOD grouping command (LOD $\geq$ 4) and the calculate map command (LOD > 3, REC < 0.4) by performing a ripple each time after adding one locus. The consensus map between the two parents was built by using codominant markers and the dominant markers displaying a 3:1 segregation type as bridge markers. Before applying the map integration command (LOD > 3, REC < 0.4) differences in recombination rates between linked loci were tested using a standard G^2-statistic. When the test was significant ($P < 0.01$), markers were not used as bridge markers. The second step consisted of adding all other markers, which were scored on only 94 offspring. The addition of new markers was done by maintaining the order of the markers genotyped on 278 individuals (fixed-order option of Joinmap). The map integration was performed with the same previous options (LOD > 3, REC < 0.4). Markers in conflict with the fixed order were removed.

The final G_{ST} map comprised in total 527 markers distributed on 980 cM, with one marker on average every 1.8 cM. Among the 527 markers, 158 were characterized for their G_{ST} values (Table 3), and only these markers are represented in Figure 4. Before plotting the G_{ST} values along the linkage groups, we checked the distribution of the markers used for the construction of the map. The distribution of markers was compared to the null hypothesis of random distribution. If the genome is subdivided in N intervals, in the case of random distribution of markers, the number of markers per interval would follow a Poisson distribution of mean μ. If the average number of random occurrences per interval is μ, then the probability that x markers are within a given interval is

$$P(x) = \frac{e^{-\mu}(\mu)^x}{x!}.$$

We compared the distribution of markers to the Poisson distribution by using a *G*-test of goodness of fit. The comparison to the Poisson distribution was done by subdividing the

linkage group in intervals of 2 cM. The size of the interval was a compromise between the density of markers available on the map and the precision of the marker position.

Distribution of interspecific G_{ST} values along the linkage groups: As one of the goals of this study was to investigate the distribution of interspecific differentiation we plotted G_{ST} values along the linkage groups.

We used Moran's index, or the spatial autocorrelation (SOKAL and ODEN 1978), to check for correlation between G_{ST} values between two markers separated by a given genetic distance on the linkage groups. The correlation between G_{ST} values of two markers separated by distance q can be written as

$$ Iq = \frac{n\sum_{i=1}^{n}\sum_{j\neq1}^{n}w_{ij}(a_i - \overline{a})(a_j - \overline{a})}{W\sum_{i=1}^{n}(a_i - \overline{a})^2} \quad \text{and} \quad W = \sum_{i=1}^{n}\sum_{j\neq1}^{n}W_{ij}, $$

with n the total number of markers; $W_{ij} = 1$ if markers i and j are within distance class q and is set to 0 when the two markers are not within the distance class q.

The a values represent the level of interspecific genetic differentiation of the marker (G_{ST}) and $\overline{a}$ is the mean of the a_i when all the n markers are considered.

Observed I values were compared to the null hypothesis of random distribution of G_{ST} values by using a permutation test. G_{ST} values were reshuffled among markers by keeping the marker position constant. Ten thousand permutations were used to construct the distribution corresponding to the null hypothesis using the SPAGeDI program (HARDY and VEKEMANS 2002).

RESULTS

Overall distribution of markers and G_{ST} values in the genome: We compared the distribution of G_{ST} values estimated for AFLPs and microsatellites. In addition, the AFLP markers could be subdivided in two different classes: fragments digested by *Eco*RI/*Mse*I and by *Pst*I/*Mse*I enzymes. *Eco*RI and *Pst*I differed in their ability to cut restriction sites containing methylated cytosine. *Pst*I (5′ CTGCAG 3′) is greatly inhibited by C methylation whereas *Eco*RI (5′ GAATTC 3′) is relatively insensitive to C methylation. The distribution of distances to the linkage group ends was different for both types of AFLPs. *Pst*I/*Mse*I markers were uniformly distributed (chi-square test nonsignificant), while *Eco*RI/*Mse*I markers were preferentially located in the internal parts of the linkage groups (Figure 1; chi-square test significant, $P = 0.0002$).

The interspecific G_{ST} values of the *Eco*RI/*Mse*I and *Pst*I/*Mse*I-AFLP markers followed an L-curve distribution (Figure 2), with numerous loci displaying a low G_{ST} value and a few markers displaying larger values. The two distributions are not significantly different when compared by Fisher's exact test ($P = 0.245$). However, the mean G_{ST} value of *Pst*I/*Mse*I-AFLP markers is twofold larger than the mean value of *Eco*RI/*Mse*I markers (Figure 2, 0.024 for *Pst*I/*Mse*I vs. 0.016 for *Eco*RI/*Mse*I). Medians of the G_{ST} values of the two distributions were significantly different when compared with the nonparametric Wilcoxon test ($W = 22,913$, $P = 0.0031$).

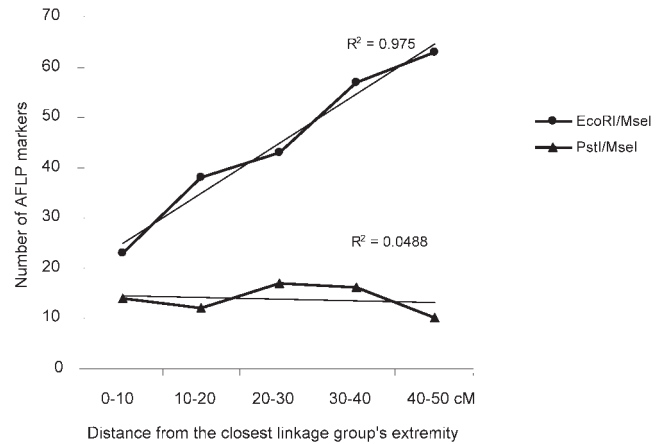The distribution of the microsatellite marker G_{ST} values differed from the distribution of the AFLP markers

FIGURE 1.—Variation of the number of *Eco*RI/*Mse*I and *Pst*I/*Mse*I markers according to their position within the linkage groups. The x-axis indicates the distance from the markers to their closest linkage group extremity. Markers were grouped into 10-cM distance classes. The y-axis indicates the number of markers corresponding to each distance class. The length of all linkage groups was standardized to 100 cM and marker positions were recalculated by interpolation.

as shown by Fisher's exact test ($P < 0.0002$). Compared to the AFLP markers, microsatellites displayed larger differentiation: 55% of the G_{ST} values were >1%, whereas only 29% of the AFLP markers were in this category. The mean of the G_{ST} values for microsatellites was closer to values obtained for *Pst*I/*Mse*I-AFLP markers, with, however, different medians as shown by the Wilcoxon test ($W = 4106.5$; $P = 0.0044$).

The distribution of G_{ST} values was compared to the neutral expectation by using the method of BEAUMONT and NICHOLS (1996). The overall mean F_{ST} value (0.0357) over the 389 markers was used to construct the expected distribution of F_{ST} in an infinite allele model. Thirty-six markers exhibited F_{ST} exceeding 95% of the null distribution. These markers were removed and a second analysis was done (mean F_{ST} value = 0.0176). In total, 47 markers fell outside the 0.95 envelope corresponding to the neutral expectation (Figure 3 and Table 3). There were differences in the proportion of outlier markers according to marker types (Table 3), with SNPs, SCARs, and isozymes exhibiting a larger proportion of outlier loci than AFLPs and microsatellites.

Distribution of G_{ST} values along linkage groups: The consensus map contained 527 markers from which 158 were characterized for their G_{ST} values (Table 3 and Figure 4). The map covered 980 cM with on average one marker every 1.8 cM, which corresponded to 82% of the estimated genome size (BARRENECHE *et al.* 1998).

The distribution of the 527 markers used for constructing the G_{ST} map was random as shown by a nonsignificant G-test ($G = 5.54$, 3 d.f., $P = 0.13$; see Figure 5a), when the distribution of the number of markers per interval was compared to the Poisson distribution (see MATERIALS AND METHODS). Similarly, when we con-
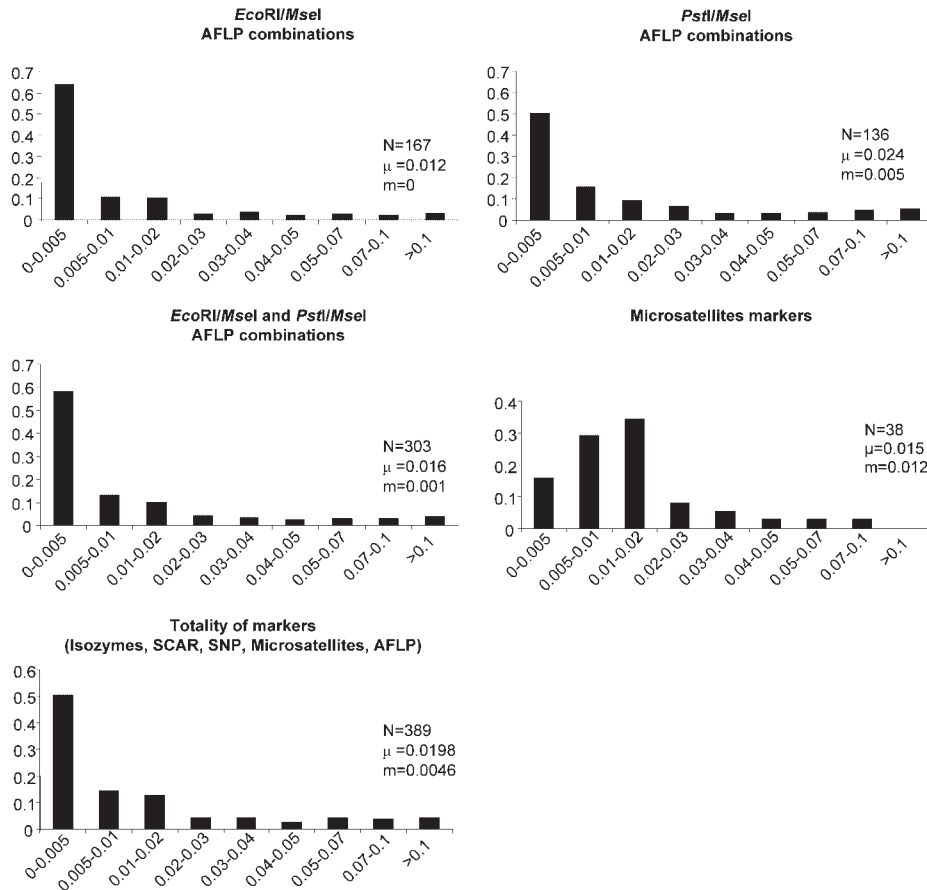
FIGURE 2.—Distribution of the G_{ST} values according to the marker types. G_{ST} values were arranged in different classes on the *x*-axis, and proportion of markers is on the *y*-axis. *N*, number of markers; μ, mean of the G_{ST} values; *m*, median of the G_{ST} values.

sidered the subset of 158 markers with a known G_{ST} value that were mapped, the distribution was random as shown by a nonsignificant *G*-test ($G = 0.52$, 1 d.f., $P = 0.47$; see Figure 5b).

Unfortunately only 20 outlier loci among the 47 identified by the Beaumont and Nichols test were polymorphic in the mapping pedigree and could be mapped. They were distributed over nine linkage groups (Figure 4). G_{ST} values were not randomly distributed along the linkage groups as shown by the spatial autocorrelation (Figure 6). Significant autocorrelations were observed for the first distance class. Indeed, pairs of markers separated by <2 cM exhibited significant correlation of their G_{ST} values ($P = 0.003$). The distance of 2 cM should be considered as an upper limit of the width for correlated differentiation as the number of markers (with known G_{ST} values) was too low to estimate autocorrelation in smaller intervals. The significant autocorrelation at distances <2 cM was mostly generated by one cluster of outlier loci located on linkage group 12 (Figure 7). Fifty pairs of markers were separated by <2 cM (Figure 7), but 3 pairs assembled markers with outlier loci that were all located on one cluster (linkage group 12), within <0.51 cM (Figures 4 and 7). Interestingly this cluster comprises two different marker types (one microsatellite ssrQrZAG112 and two *Pst*I/*Mse*I-AFLP markers, P-CCA/M-CAA-181 and P-CCA/M-ATA-335). Groupings

of markers with large G_{ST} values occurred also on linkage groups 2 and 4 (Figures 4 and 7).

DISCUSSION

Genome-wide distribution of interspecific differentiation: Our results clearly confirmed earlier reports that the genomes of these two closely related oak species are extremely permeable. An important body of results shows that various genetic markers exhibit low species differentiation (BODÉNÈS *et al.* 1997a; MARIETTE *et al.* 2002). However, most of these studies were restricted to a low number of loci, and the limited sampling in the genome did not allow us to identify those genomic regions that are less permeable. Our study based on 389 markers clearly showed that the distribution of species differentiation follows an L-shaped curve, with only a few markers exhibiting large species differentiation. Earlier random amplification of the oak genome suggested that markers that differentiate the two species are likely to be present in extremely small numbers (BODÉNÈS *et al.* 1997a). To our knowledge there are only two reported studies on genome-wide distribution of G_{ST} values. MARIETTE *et al.* (2002) in addition to interspecific values also provided distributions for intraspecific G_{ST} values in both oak species separately. The L distribution was confirmed at both levels, but the distribution of inter-
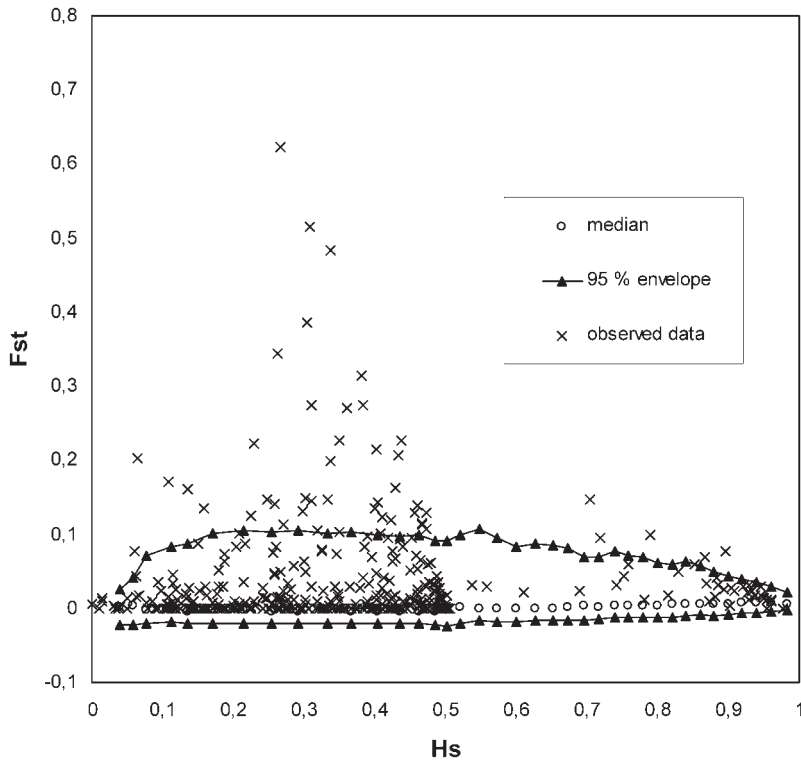
FIGURE 3.—Distribution of F_{ST} values as a function of heterozygosity (H_S). The envelope of values corresponding to neutral expectations (with $F_{ST} = 0.0176$) with the infinite allele model was constructed according to BEAUMONT and NICHOLS (1996).

specific G_{ST} was much more skewed than the distribution of the intraspecific G_{ST}. In humans, a clear L-shaped distribution was also found in a large-scale study based on 26,530 SNPs, but again skewness was less pronounced than in our case at the interspecific level (AKEY *et al.* 2002). These differences may indicate that the genomic regions contributing to interspecific divergence are fewer than those for intraspecific divergence. Despite the evolutionary stochastic variance of G_{ST} of neutral markers (ROBERTSON 1975), we found that 12% of the markers exhibited G_{ST} values that were not compatible with the neutral expectation according to the Beaumont and Nichols test (BEAUMONT and NICHOLS 1996). This number may be even larger because in multilocus systems genes showing low G_{ST} values may also be responding to selection. In a recent article, LE CORRE and KREMER (2003) used simulations to monitor the evolution of G_{ST} of genes contributing to a quantitative trait undergoing diversifying selection in a set of populations. The results indicated that only a reduced number of genes contributing to a trait will actually behave as outliers, whereas others will behave as neutral markers despite their contribution to the trait submitted to selection. As a conclusion, the overall distribution of G_{ST} values (Figure 2) is most likely composed of a mixture of two partially overlapping distributions corresponding to markers undergoing selection and neutral markers. Because of their partial overlap, inferences about their response to selection can be made only for markers with extreme G_{ST} values (outliers). The distribution of microsatellites does not fit to this general L shape, but

this may be attributed merely to the way microsatellites were developed. Only those microsatellite motifs that exhibited allelic polymorphism were actually used as genetic markers, whereas many others were discarded when molecular libraries were screened simply because they were not polymorphic (STEINKELLNER *et al.* 1997). Hence the distribution in Figure 2 is truncated due to the molecular screening procedure that was applied during the development of microsatellite markers. AFLPs did not undergo this screening procedure although they were partly pruned by applying the LYNCH and MILLIGAN (1994) restriction. However, there was a difference in the level of interspecific differentiation detected by two different AFLP markers. The mean G_{ST} value for *Pst*I/*Mse*I-AFLP markers was twofold larger than that for *Eco*RI/*Mse*I-AFLP markers. These differences could be due to the sensitivity to cytosine methylation of the *Pst*I restriction enzyme. Different studies show that markers generated by *Eco*RI/*Mse*I and *Pst*I/*Mse*I enzyme combinations are differently distributed in the genome. *Eco*RI/*Mse*I markers are preferentially localized in centromeric regions whereas *Pst*I/*Mse*I markers are localized in the hypomethylated noncentromeric regions of the chromosome (CASTIGLIONI *et al.* 1999; YOUNG *et al.* 1999), as confirmed also in our study (Figure 1). Yet, the DNA methylation has an essential regulatory gene expression function. It provides a mechanism to turn off permanently the transcription of genes whose activity is not required in a particular cell type. This stable silencing of a large fraction of the genome would allow the transcriptional machinery to
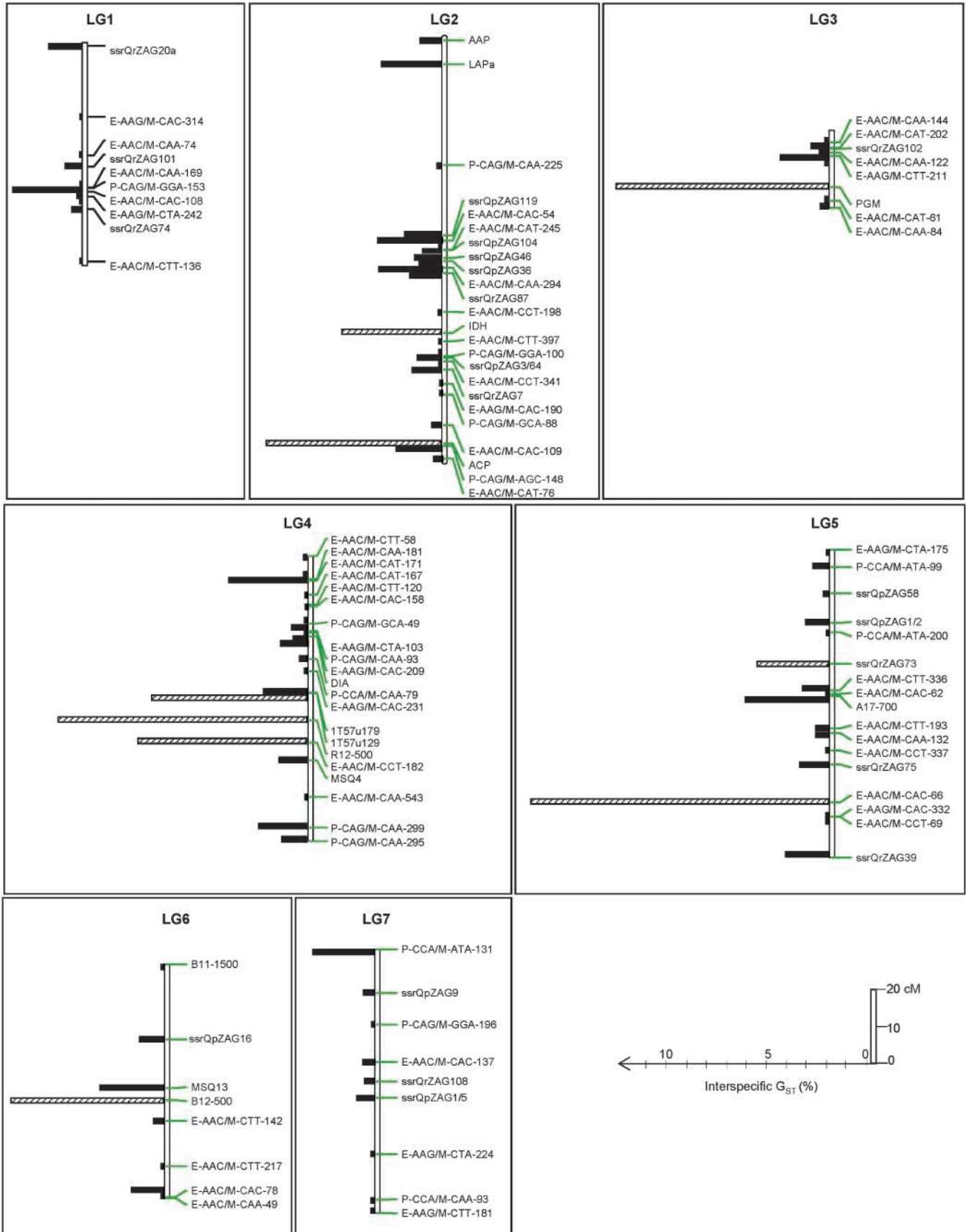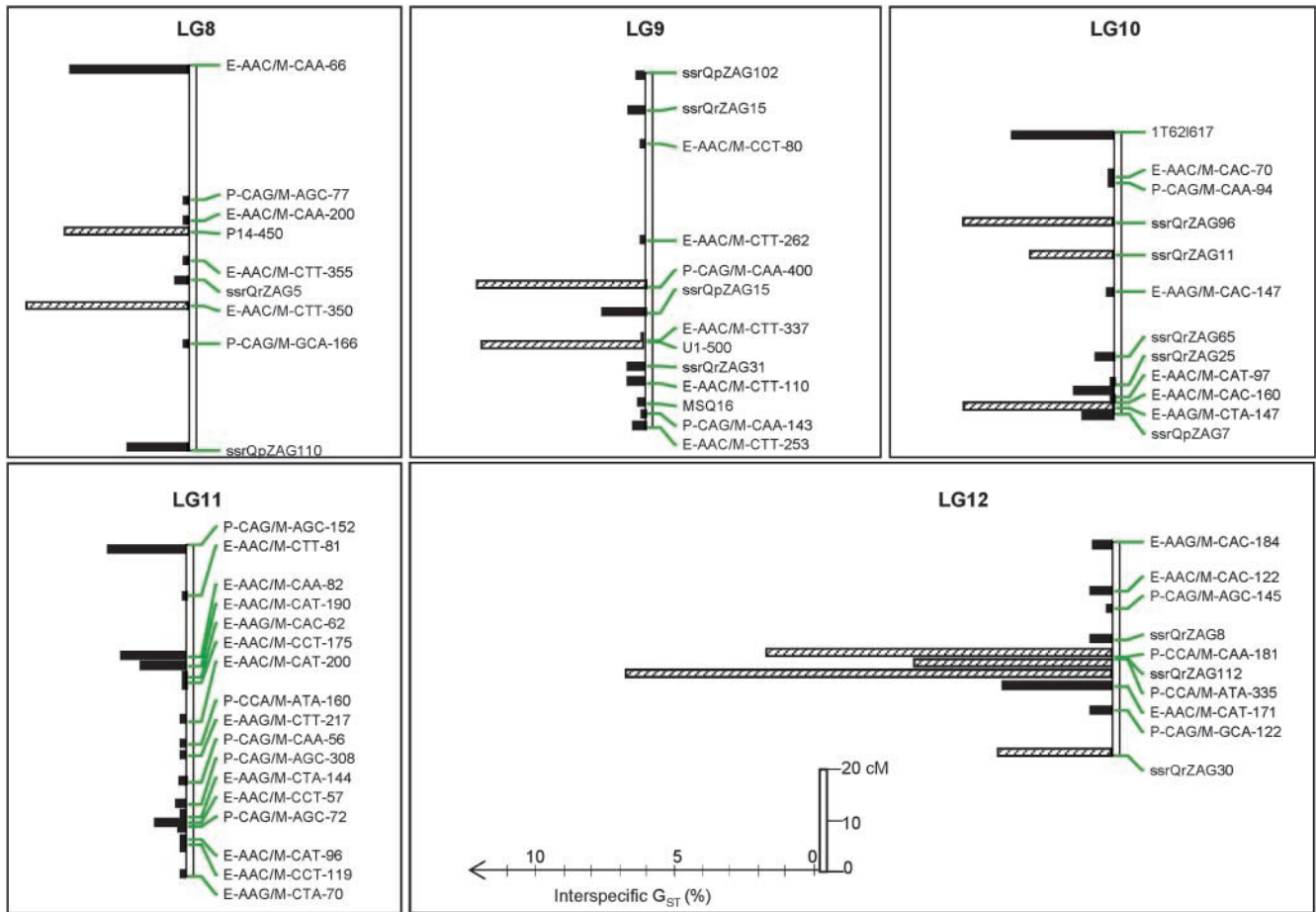
FIGURE 4.—*Continued.*

focus on those genes that are essential for the expression and the maintenance of the differentiated phenotypes (KASS *et al.* 1997). Hence *Pst*I cuts preferentially in the coding regions that are expressed, whereas the *Eco*RI restriction enzyme cuts rather randomly in the genome. Interestingly our results indicated that fragments digested by *Pst*I exhibited larger species differentiation than fragments digested by *Eco*RI, suggesting that species differentiation would preferentially be located in nonneutral regions of the genome. The proportion of outlier loci, *e.g.*, markers with large G_{ST} values, was also different according to marker types (Table 3). It was much lower in markers located in anonymous regions (AFLPs and microsatellites) than in markers located in genes or nonanonymous regions (isozymes, SNPs, and SCARs). These results also contribute to the conclusion

that species divergence between *Q. petraea* and *Q. robur* resides mostly in functional regions of the genome.

Distribution of interspecific differentiation along linkage groups: Interspecific differentiation between the two species was widely distributed throughout the genome as markers with large G_{ST} values were present at various locations on all linkage groups (Figure 4). Here we identified 47 outlier loci (12%) among which only 20 could be positioned on the genetic map. Their location over nine different linkage groups indicated that selection acting toward species divergence is widespread in the genome. To our knowledge this is the first systematic genome scan available for species differentiation. Interestingly our results confirmed those observed by a different approach, where QTL involved in phenotypic discriminant characters were also distributed over all the

FIGURE 4.—Distribution of G_{ST} values along the linkage groups. Markers without G_{ST} values are not indicated on the map, but were used for the map construction. Markers with hatched bars are outlier loci with large G_{ST} values (according to BEAUMONT and NICHOLS 1996) P, *Pst*I/*Mse*I-AFLP marker; E, *Eco*RI/*Mse*I-AFLP marker; MSQ, microsatellite Quercus; ssr, single-sequence repeat; microsatellite markers; 1T, 2T, SNPs; A17-700, B11-1500, B12-500, P14-450, R12-500, U1-500, SCARs; IDH, ACP, PGM, DIA, LAPa, AAP, isozymes.
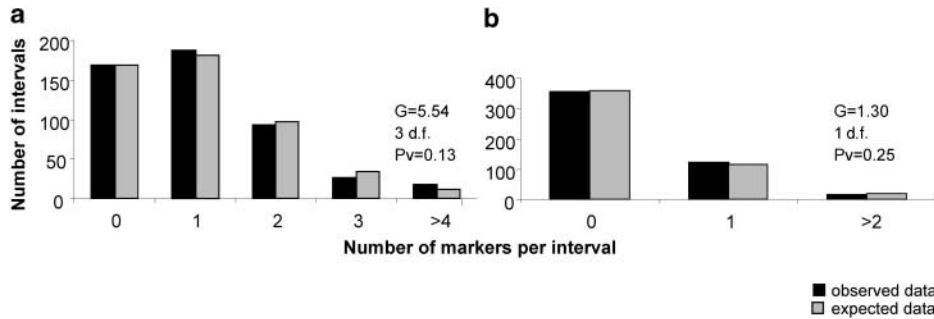
FIGURE 5.—Observed and expected distribution of the markers in the genome. (a) Distribution of the total number of mapped markers (527) in the genome. (b) Distribution of the mapped markers with a known G_{ST} value (158) in the genome.

linkage groups (SAINTAGNE *et al.* 2004). Wide distribution of large G_{ST} values was also observed in humans, when intraspecific differentiation was calculated among African-American, East Asian, and European-American populations (AKEY *et al.* 2002). Besides their broad distribution, high G_{ST} values were clustered at short genetic distances in a few hot spots, which are well illustrated on three linkage groups (LG12, LG2, and LG4). In one of these spots (LG12) three outlier loci concentrated within <0.51 cM. As shown by the BEAUMONT and NICHOLS (1996) test, differentiation within this spot deviates from neutral expectation and the size of the spot is most likely caused by hitchhiking effects due to selection. Hitchhiking effects result from the linkage disequilibrium near the selected locus (ANDOLFATTO 2001). Our results suggested that selection is most likely involved in species divergence. Directional selection within population (within species in our case) is known to create a local reduction of within-population diversity, called selective sweep (KAPLAN *et al.* 1989). When acting in opposing directions in two different species, disruptive selection toward different optima will contribute to decreasing the within-species diversity at the target locus and to increasing species divergence. Hence disruptive selection will actually increase differentiation at the target locus and in adjacent regions as a result of hitchhiking. In oak species, the physical size of the hitchhiked

region for interspecific differentiation is <2 cM on the basis of autocorrelation analysis (Figure 6). This is much larger than in humans. AKEY *et al.* (2002) found that correlations between G_{ST} values of adjacent SNPs were significant when SNPs were separated by <200 kb, which translates into a genetic distance close to 0.2 cM. Discrepancy between these values may be due to the difference in the strength of selection responsible for species and population differentiation. The size of a hitchhiked region depends on the ratio of the selective advantage of the favored alleles to the recombination fraction (ANDOLFATTO 2001). As already indicated the skewness of the overall L distribution of G_{ST} is much more pronounced, even in oaks (MARIETTE *et al.* 2002), for interthan for intraspecific differentiation, suggesting a difference in the strength of selection. Interestingly, in humans the size of the hitchhiked region for population differentiation (200 kb; AKEY *et al.* 2002) is much larger than the size of linkage disequilibrium (LD) within populations. Consensual reported values of LD vary between 10 and 30 kb (ARDLIE *et al.* 2002), suggesting that differentiation may be a more powerful tool than within-population diversity for detecting selection signatures
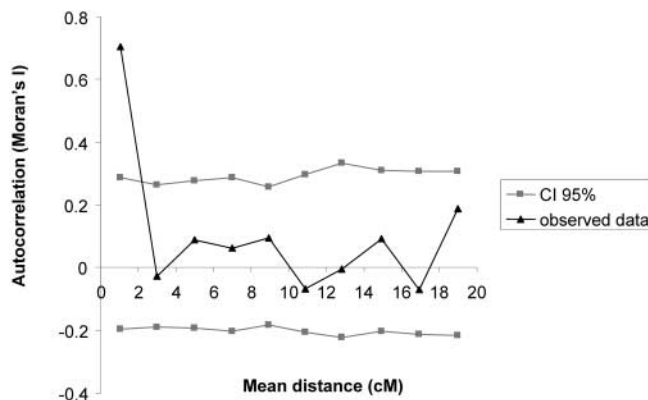

FIGURE 6.—Autocorrelation (*I*) of G_{ST} values as a function of genetic distance. The *x*-axis indicates the classes of distance between pairs of markers (in centimorgans). The *y*-axis indicates Moran's index (*I*).
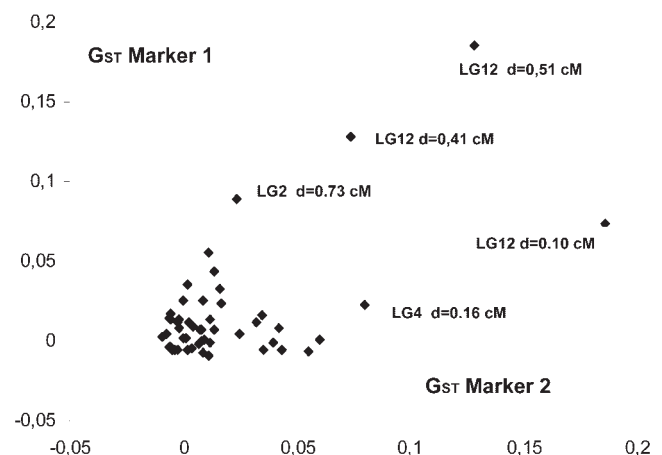

FIGURE 7.—Covariation of G_{ST} values for pairs of markers separated by <2 cM. Annotations are given for all pairs composing at least one outlier locus (according to BEAUMONT and NICHOLS 1996). The annotation includes the linkage group (LG) and the distance (*d*, in centimorgans) separating the two markers. The three pairs on LG12 comprise three markers located within <0.51 cM (see text and Figure 4).

in populations. LDs in outcrossing plants, mainly maize, are significant at much lower distances (up to a few kilobases) and would indicate a much narrower hitchhiking window (Flint-Garcia *et al.* 2003). The 2-cM size of the hitchhiked region in our study is an overall estimate that may vary from region to region. Autocorrelation analysis assumes that the strength of the processes leading to the spatial distribution is stationary, meaning in our case that the strength of selection would have the same magnitude at all hitchhiked regions. No argument supports this assumption. The results as shown in Figure 4 actually indicate different sizes. Linkage groups 2 and 4 exhibit regions extending over 4 cM, whereas LG12 exhibits narrow regions. The 2-cM size of the hitchhiked region is also an upper limit. Because the overall number of markers was not large enough we were not able to calculate correlations between adjacent G_{ST}'s at smaller distances. It is highly likely that these correlations would actually increase at smaller distances. At this stage, with the existing sampling of markers, we were not able to monitor the decay of differentiation within the hitchhiked region. In combination with the QTL detection of discriminant phenotypic traits, this study will help to locate genes that may be responsible for species differentiation. Ongoing investigations based on gene expression studies will help to identify candidate genes that will be further mapped on the oak genetic maps. Their colocalization with high-G_{ST} regions will provide a piece of evidence for their involvement in species differentiation.

LITERATURE CITED

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

Andolfatto, P., 2001 Adaptive hitchhiking effects on genome variability. Curr. Plant Biol. **11:** 635–641.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. **3:** 299–309.

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

Barreneche, T., N. Bahrman and A. Kremer, 1996 Two dimensional gel electrophoresis confirms the low level of genetic differentiation between *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. For. Genet. **3:** 89–92.

Barreneche, T., C. Bodénès, C. Lexer, J. F. Trontin, S. Fluch *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.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B **263:** 1619–1626.

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

Bodénès, C., T. Labbé, S. Pradère and A. Kremer, 1997b General vs. local differentiation between two closely related white oak species. Mol. Ecol. **6:** 713–724.

Castiglioni, P., P. Ajmone-Marsan, R. Van Wijk and M. Motto, 1999 AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. Theor. Appl. Genet. **99:** 425–431.

Coart, E., V. Lamote, M. De Loose, E. Van Bockstaele, P. Lootens *et al.*, 2002 AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl.] in Flemish populations. Theor. Appl. Genet. **105:** 431–439.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

Dieckmann, U., and M. Doebeli, 1999 On the origin of species by sympatric speciation. Nature **400:** 354–357.

Draye, X., Y. R. Lin, X. Y. Qian, J. E. Bowers, G. B. Burow *et al.*, 2001 Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiol. **125:** 1325–1341.

Dvorak, J., M. C. Luo and Z. L. Yang, 1998 Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing Aegilops species. Genetics **148:** 423–434.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler, IV, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54:** 357–374.

Gerber, S., S. Mariette, R. Streiff, C. Bodénès and A. Kremer, 2000 Comparison of microsatellites and AFLP markers for parentage analysis. Mol. Ecol. **9:** 1037–1048.

Gömöry, D., 2000 Gene coding for a non-specific NAD-dependent dehydrogenase shows a strong differentiation between *Quercus robur* and *Quercus petraea*. For. Genet. **7:** 167–170.

Gömöry, D., I. Yakovlev, P. Zhelev, J. Jedináková and L. Paule, 2001 Genetic differentiation of oak populations within the *Quercus robur/Quercus petraea* complex in central and eastern Europe. Heredity **86:** 557–563.

Grattapaglia, D., F. L. Bertolucci and R. R. Sederoff, 1995 Genetic mapping of quantitative trait loci (QTLs) controlling vegetative propagation in *Eucalyptus grandis* and *E. urophylla*, using the pseudo-testcross mapping strategy and RAPD markers. Theor. Appl. Genet. **90:** 933–947.

Hamblin, M. T., and C. F. Aquadro, 1999 DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859–869.

Hardy, O. J., and X. Vekemans, 2002 SPAGeDI A versatile computer program to analyse spatial genetic structure at the individual or population level. Mol. Ecol. Notes **2:** 618–620.

Kampfer, S., C. Lexer, J. Glössl and H. Steinkellner, 1998 Characterization of (GA)n microsatellite loci from *Quercus robur*. Hereditas **129:** 183–186.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kass, S. U., N. Landsberger and A. P. Wolffe, 1997 DNA methylation directs a time-dependent repression of transcription initiation. Curr. Biol. **7:** 157–165.

Kondrashov, A. S., and F. A. Kondrashov, 1999 Interactions among quantitative traits in the course of sympatric speciation. Nature **400:** 351–354.

Kremer, A., J. L. Dupouey, J. D. Deans, J. Cottrell, U. Csaikl *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.

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

Lévy, G., M. Becker and D. Duhamel, 1992 A comparison of the ecology of pedunculate oak and sessile oak: radial growth in the centre and northeast of France. For. Ecol. Manage. **55:** 51–63.

Lynch, M., and B. Milligan, 1994 Analysis of population-genetic structure using RAPD markers. Mol. Ecol. **3:** 91–99.

Mariette, S., D. Chagne, C. Lezier, P. Pastuszka, A. Raffin *et al.*, 2001 Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. Heredity **86:** 469–479.

Mariette, S., J. Cottrell, U. M. Csaikl, P. Goikoechea, A. König *et al.*, 2002 Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Q. petraea* (Matt.) Liebl. and *Q. robur* L. stands. Silvae Genet. **51:** 72–79.

Mayr, E., 1963 *Animal Species and Evolution.* Belknap Press, Cambridge, MA.

Muir, G., C. C. Fleming and C. Schlötterer, 2000 Species status of hybridizing oaks. Nature **405:** 1016.

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

Orr, A. H., 2001 The genetics of species differences. Trends Ecol. Evol. **16:** 343–350.

Paran, I., and R. W. Michelmore, 1993 Development of reliable PCR-based markers linked to downy mildew in resistance genes in lettuce. Theor. Appl. Genet. **85:** 985–993.

Petit, R. J., and A. Kremer, 1993 Ribosomal DNA and chloroplast DNA polymorphisms in a mixed stand of *Quercus robur* and *Q. petraea.* Ann. Sci. For. **50** (Suppl. 1): 41s–47s.

Petit, R. J., C. Bodénès, A. Ducousso, G. Roussel and A. Kremer, 2003 Hybridization as a mechanism of invasion in oaks. New Phytol. **161:** 151–164.

Rice, W. R., and E. E. Hostert, 1993 Laboratory experiments on speciation: What have we learned in 40 years? Evolution **47:** 1637–1653.

Rieseberg, L. H., S. J. E. Baird and K. A. Gardner, 2000 Hybridization, introgression, and linkage evolution. Plant Mol. Biol. **42:** 205–224.

Roberstson, A., 1975 Gene frequency distributions as a test of selective neutrality. Genetics **81:** 775–785.

Rundle, H., 2002 A test of ecologically dependent postmating isolation between sympatric sticklebacks. Evolution **56** (2): 322–329.

Saintagne, C., C. Bodénès, T. Barreneche, D. Pot, C. Plomion *et al.*, 2004 Distribution of genomic regions differentiating oak species assessed by QTL detection. Heredity **92:** 20–30.

Schliewen, U., K. Rassmann, M. Markmann, J. Markert, T. Kochers *et al.*, 2001 Genetic and ecological divergence of a monophyletic cichlid species pair under fully sympatric conditions in Lake Ejagham, Cameroon. Mol. Ecol. **10:** 1471–1488.

Schlötterer, C., 2003 Hitchhiking mapping—functional genomics from the population genetics perspective. Trends Genet. **19:** 32–38.

Sokal, R. R., and N. L. Oden, 1978 Spatial autocorrelation in biology. 1. Methodology. Biol. J. Linn. Soc. **10:** 199–228.

Steinkellner, H., S. Fluch, E. Turetschek, C. Lexer, R. Streiff *et al.*, 1997 Identification and characterization of (GA/CT)n microsatellite loci from *Quercus petraea.* Plant Mol. Biol. **33:** 1093–1096.

Stephan, W., and C. H. Langley, 1998 DNA polymorphism in Lycopersicum and crossing-over per physical length. Genetics **150:** 1585–1593.

Streiff, R., T. Labbe, R. Bacilieri, H. Steinkellner, J. Glossl *et al.*, 1998 Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. Assessed with isozymes and microsatellites. Mol. Ecol. **7:** 317–328.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.) Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Van Ooijen, J. W., and R. E. Voorrips, 1993 Plant Research International, Biometris, Wageningen. Plant J. **3:** 739–744.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Wu, C. I., 2001 The genic view of the process of speciation. J. Evol. Biol. **14:** 851–865.

Young, W. P., J. M. Schupp and P. Keim, 1999 DNA methylation and AFLP marker distribution in the soybean genome. Theor. Appl. Genet. **99:** 785–792.

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

Communicating editor: L. Excoffier