ORIGINAL ARTICLE

Nucleotide variation in *Quercus crispula* Blume

ND Quang, S Ikeda and K Harada
Faculty of Agriculture, Ehime University, Tarumi, Matsuyama, Japan

Quercus is attractive for evolutionary studies, primarily for developing the concepts of the species, speciation and adaptation; however, remarkably little is known about levels of nucleotide polymorphism in the nuclear functional genes of this genus. This article provides the first characterization of levels of nucleotide polymorphism in 11 gene fragments in natural populations of a *Quercus* species, *Quercus crispula* Blume. Results show that the level of nucleotide variation in this oak is generally higher than that in conifers, as high as that in a European oak, but lower than that in an aspen. The level of population recombination is relatively high. Within-population inbreeding is negligible and between-population differentiation is modest. The decay of linkage disequilibrium is significantly faster in the species-wide samples and the three northernmost populations than in the other populations. Statistical tests support the hypothesis of a recent bottleneck for several populations in the southern part of Japan. The amounts and patterns of nucleotide variation, recombination and linkage disequilibrium, and genetic differentiation observed among populations of this species are contradictory to our expectations, given the recent colonization history of the northern Japan populations.

Heredity (2008) **101,** 166–174; doi:10.1038/hdy.2008.42; published online 28 May 2008

Keywords: oak; nucleotide variation; DNA; evolution; *Quercus*; genetics

Introduction

Nucleotide sequence data integrate information over a long time period (on average, 4*N* generations, where *N* is the effective population size), and thus can provide powerful insights into the important factors contributing to the evolutionary dynamics of a species. In any natural population, the amounts and patterns of nucleotide variation are influenced by both biological factors (for example, mutation, recombination and natural selection) and demographic factors (for example, range expansion and colonization, population subdivision and admixture) affecting *N*.

In recent years, nucleotide variation in nuclear functional genes has been extensively studied in herbaceous plants (Innan *et al*., 1996; Filatov and Charlesworth, 1999; Kuittinen and Aguage, 2000; Aguade, 2001; Zhang and Ge, 2006; Arunyawat *et al*., 2007). However, there has been only a very limited number of such published reports about tree species. The amounts and patterns of genetic variation in plants are strongly affected by life history traits and mating systems (Hamrick and Godt, 1996); hence, it will be problematic to generalize the patterns of nucleotide variation in herbaceous plants to woody plants. Thus far, most data of nucleotide variation in trees have been from *Pinus* (Dvornyk *et al*., 2002; Garcia-Gil *et al*., 2003; Neale and Savolainen, 2004; Pot *et al*., 2005), *Cryptomeria* (Kado *et al*., 2003), *Chamaecyparis* (Kado *et al*., 2008) and *Populus* (Ingvarsson, 2005).

The genus *Quercus* comprises 350–500 species distributed throughout the northern hemisphere (Kubitzki, 1993; Nixon, 1993). They are conspicuous members of temperate deciduous forests of North America, Europe, Asia, as well as the evergreen Mediterranean maquis. The members of this genus are attractive for research, primarily for developing concepts of the species, speciation and evolution (Van, 1976; Rieseberg and Wendel, 1993). Nonetheless, remarkably little is known about levels of nucleotide polymorphism in the functional genes of this genus. Hence, this article provides the first characterization of levels of nucleotide polymorphism in a species of *Quercus*, *Quercus crispula* Blume.

Q. crispula Blume (Fagaceae) is a common tree throughout the cool-temperate deciduous forests of southern Sakhalin, the Kuril Islands, Japan, and Korea (Watanabe, 1994). The populations are dense and continuous in northern Japan but sparse and commonly isolated in small patchy forests at elevations >800 m in the south (Figure 1). cpDNA data (Kanno *et al*., 2004; Okaura *et al*., 2007) are consistent with palynological data (Hibino and Takeuchi, 1998; Miyoshi, 1998; Morita *et al*., 1998), suggesting a recent colonization history (from 8000 to 12000 years ago) for the northern Japan (approximately more than 35° N) populations. The southern populations are shown to have remained *in situ* during the last glacial maximum (LGM) and lasted from 20 000 to 18 000 years BP. The former populations are likely to be adaptive to a colder climate and the latter to a warmer climate (Okaura *et al*., 2007). Our previous cpDNA study also revealed a similar colonization history for another deciduous tree, *Fagus crenata* (Fagaceae), usually forming a major vegetation class together with *Q. crispula* (Okaura and Harada, 2002).

The present study addresses the following questions: (1) What is the level of nucleotide diversity in this oak

Correspondence: Dr ND Quang, Faculty of Agriculture, Ehime University, 3-5-7 Tarumi, Matsuyama, Japan.
E-mail: quanghoa@agr.ehime-u.ac.jp.

Received 22 January 2008; revised 10 April 2008; accepted 15 April 2008; published online 28 May 2008

species as compared to pines and other broadleaved species? (2) What are the pattern and extent of linkage disequilibrium (LD) among populations of *Q. crispula*? (3) Do the patterns of genetic variation differ significantly between the northern and southern Japan populations? (4) Can we find sequence evidence favoring the hypothesis of climatic adaptation in populations of *Q. crispula*?

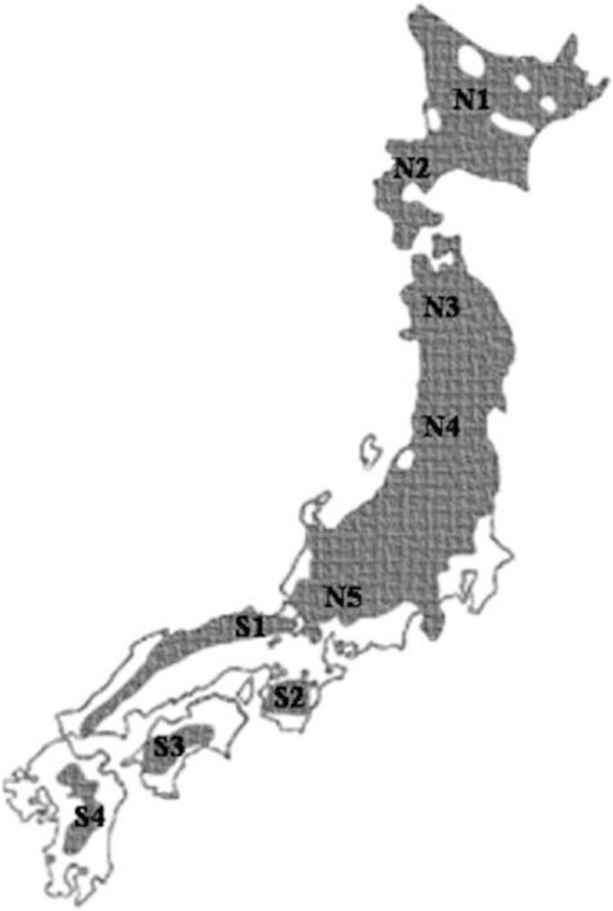

Figure 1 Approximate geographical distribution of the populations sampled and the current pattern of species distribution in Japan. The name codes of each population (N1–S4) are the same as those given in Supplementary Table S1. The shadings indicate the patterns of the current population distribution and density, based on Horikawa (1972).

Materials and methods

Samples

Fresh leaves were collected from trees belonging to nine populations of *Q. crispula* distributed throughout Japan (Figure 1; Supplementary Table S1). Leaves collected from another Japanese oak were used as outgroup species (*Quercus phillyraeoides*, no morphological evidence for the hybridization of this species with *Q. crispula*). We collected leaves from trees growing at least 30 m apart. Trees clearly distinguishable by morphological characteristics were chosen. The samples were immediately processed for DNA extraction or frozen at −20 °C until use.

Loci

We genotyped 11 loci, all of which are nuclear functional genes (Table 1). Except for *MetE*, the others are 10 expressed sequence tag (EST) loci extracted from oak EST resources reported in a previous study (Casasoli *et al*., 2006).

DNA extraction and PCR amplification

Total genomic DNA was extracted from fresh or frozen leaf tissues using the cetyltrimethylammonium bromide method. The primers for the amplification of *MetE* are identical to those specified by Chiang *et al*. (2002). Those for the amplification of the 10 EST loci are the same as those reported by Casasoli *et al*. (2006).

Subcloning and sequencing

The three long (>1000 bp) loci, *MetE*, Cons109 and Cons14, were cloned and sequenced to generate haplotype sequences. The PCR products were cloned into the TOPO vector using a Zero Blunt TOPO PCR Cloning kit for sequencing (Invitrogen). The clones were sequenced using the BigDye terminator sequencing kit (Applied Biosystems, Foster City, CA, USA) on an ABI310 automated sequencer (Applied Biosystems). At least

Table 1 Summary of loci studied

Locus	*Location*[a]	*Alignment length (bp)*			*Functional association*[b]
		Total	*Exon*	*Intron*	
MetE	NA	1385	611	774	Catalyst in the terminal step in the methionine biosynthesis of plants
Cons21	Q2	480	179	301	*Arabidopsis thaliana* ribosomal protein S27 ($4e^{-40}$)
01A03	Q7	539	212	327	*Fragaria × ananassa* putative 40S ribosomal protein (e^{-126})
Cons126	Q7	239	188	51	*Prunus avium* 60S ribosomal protein L24 ($3e^{-62}$)
Cons107	Q11	448	222	226	*Alnus firma* mRNA for nonsymbiotic hemoglobin ($2e^{-76}$)
Cons58	Q5	240	110	130	*Arabidopsis thaliana* dormancy-associated protein ($4e^{-22}$)
Cons105	Q12	440	90	350	*Malus domestica* transposon gene for putative DNA topoisomerase ($5e^{-45}$)
Cons41	NA	354	244	110	*Aquilegia formosa* heat shock protein 70 (HSP70; $3e^{-60}$)
Cons106	NA	237	189	48	Spinach gene for plastocyanin ($7e^{-27}$)
Cons14	Q5	1100	243	857	*Arabidopsis thaliana* At2g34480 gene ($4e^{-36}$)
Cons109	Q7	1015	194	821	*Fagus sylvatica* mRNA for metallothionein-like protein (e^{-100})
Total	⩾5	6477	2045	4432	

[a] Chromosome location of the 10 EST loci are the same as in Casasoli *et al*. (2006); NA, not yet mapped.
[b] The sequence of 10 EST loci of *Quercus crispula* was used to search for homologous gene by BLAST tool in DDBJ (http://www.ddbj.nig.ac.jp/); numbers in the parentheses are *E*-values.

8–10 different clones of each fragment were sequenced to detect the presence of multiple haplotypes within individuals and to control for *Taq* polymerase errors. For the other eight short (<600 bp) loci, we used a direct sequencing technique to obtain single nucleotide polymorphism (SNP) data. The frequency of each SNP was scored by direct sequencing of the gene fragments in which it was present. Heterozygous sites in a single individual were visible as double peaks in the chromatograms generated by the automated sequencer. The peaks were confirmed by repeating the experiment twice if necessary.

Data analysis

Sequences were verified manually. All indels (insertion/deletion) were excluded from analyses. Analyses of sequence data of the three long loci, *MetE*, Cons109 and Cons14, were performed using DnaSP version 4.10.8 (Rozas *et al*., 2003). For the eight short loci, nucleotide diversity (π) was estimated by dividing the average pairwise sequence difference (k) by the number of sites.

The alignments of all the sequences of *MetE*, Cons109 and Cons14 were used to perform analysis of recombination with DnaSP. The minimum number of recombination events (R_M) that have occurred along the sequence was estimated following the four-gamate test (Hudson and Kaplan, 1985). The two indices of LD, D' and r^2, were estimated between pairs of polymorphic sites for each locus with DnaSP. The statistical significance of each pairwise test of LD was evaluated using the χ^2-test. All pairwise analyses are not fully independent because of LD itself; thus, the proportion of significant ($P<0.05$) pairwise tests was compared for pairs of sites separated by different physical distances (>400 or <400 bp). Comparison of these ratios over the two separate distance ranges was carried out using Fisher's exact test (two-tailed). Although the nonlinear regression analysis (Remington *et al*., 2001) was useful as a representation of the overall LD rate (Ingvarsson, 2005; Zhu *et al*., 2007), it yields a least-squares estimate of the population recombination rate per base pair; this estimate may be imprecise and unrealistic because of several factors, for example, the nonindependence between linked sites and the nonequilibrium populations.

Tajima's *D* (Tajima, 1989a) statistic was used to test for deviations from neutrality of the three loci, *MetE*, Cons109 and Cons14. This test measures skews in the frequency spectrum; a negative *D* suggests an excess of low-frequency polymorphisms and a positive *D* indicates an excess of intermediate-frequency polymorphisms. Tajima's test is conservative for testing departures from neutral equilibrium conditions, particularly under the assumption of no recombination. In addition, the test statistics *D* and F proposed by Fu and Li (1993) were computed for the three loci. All these tests assume random mating and random sampling, and DnaSP was used to carry out the tests. We did not apply these tests for the eight short loci because of the small number of segregating sites per locus. Instead, we performed the multilocus Hudson–Kreitman–Aguade (HKA) test (Hudson *et al*., 1987; http://lifesci.rutgers.edu/~heylab/) to assess the neutral prediction of the ratio of polymorphism to divergence across the eight EST loci. *Q. phillyraeoides* homologous sequences were used as outgroup for this test.

Inbreeding indices (F_{IS}) and their statistical significance (with 1000 randomizations) were estimated using FSTAT ver. 2.9.3.2 (Goudet, 2001). The SNP data of all the 11 loci were used in this analysis for the three northernmost and three southernmost populations; SNP data of the three loci, *MetE*, Cons109 and Cons14, were used for estimating F_{IS} for the three central populations (N4-S1). We estimated the standard F_{ST} statistic for *MetE*, Cons109 and Cons14 using the method of Hudson *et al*. (1992); the probability of the statistical test of genetic differentiation was estimated using the method of Hudson (2000) (*S*nn statistic-based probabilities, evaluated by 10 000 permutations). For the eight short loci, we used their SNP data to estimate F_{ST}, 99% confidence intervals of F_{ST}, and its statistical significance with 1000 randomizations with FSTAT.

Results

We obtained the sequences of 11 loci from three northernmost (N1–N3) and three southernmost (S2–S4) populations; the sequences of *MetE*, Cons109 and Cons14 were additionally obtained from three populations (N4, N5 and S1) located approximately in the center of Japan (Figure 1). The length of the aligned sequence for each locus varied between 237 and 1385 bp, with a total of 6477 bp in length. The number of sequences sampled for each locus (n) in each population is shown in Tables 2 and 3. Only two variants were detected for all the polymorphic sites in this study. A total of 261 SNPs were detected, and thus an average of 1 SNP every ~25 bp was found in our sample set. High haplotype fractions (h/n) were found in every population at all loci (Table 2).

Levels of within-population inbreeding and between-population differentiation

The F_{IS} values did not vary significantly among different populations, ranging from -0.081 (N5) to 0.044 (S1). None of the populations examined had F_{IS} values significantly deviating from 0. The F_{ST} values are summarized in Table 3. There was slight variation across loci in the levels of genetic differentiation. Although all loci showed significant genetic differentiation across populations, the level of population differentiation was modest. The averages of F_{ST} across loci were low both in the pool and in the three northernmost and three southernmost populations (Table 3). Previous microsatellite studies (Ohsawa *et al*., 2007) also revealed very low levels of between-population differentiation and within-population inbreeding for all populations examined.

Nucleotide variation

Standard statistics of nucleotide variation for the three loci, *MetE*, Cons109 and Cons14, in each population and the pool (combined sequence data) are presented in Table 2. A total of 510 sequences was sampled for the three loci in nine populations, thus an average of ~19 sequences per locus per population were used to estimate nucleotide polymorphism (θ_W) as well as nucleotide diversity at total sites (π_{tot}) and at silent sites (π_{sil}). The levels of nucleotide diversity are heterogeneous among loci, with *MetE* being the least diverse locus ($\pi_{tot}=0.00667$, $\pi_{sil}=0.00711$ in the pool) and Cons109 being the highest ($\pi_{tot}=0.00721$, $\pi_{sil}=0.00768$

Table 2 Summary of nucleotide variation for *MetE*, Cons109 and Cons14

Population	*Locus*	n^a	S^b	h^c	$\pi_{tot}{}^d$	$\pi_{sil}{}^d$	$\theta_W{}^d$	*Tajima's D*^{e}	*Fu and Li's D*^{e}	*Fu and Li's F*^{e}
N1	*MetE*	14	29	13	587	621	662	−0.4843	−0.1927	−0.3123
	Cons109	26	28	16	699	710	722	−0.3180	−0.5670	−0.7680
	Cons14	20	29	20	678	722	743	−0.1120	−0.5980	−0.8650
N2	*MetE*	14	30	12	637	699	685	−0.2972	−0.4497	−0.4684
	Cons109	24	26	19	673	987	687	−0.6570	−0.8310	−0.4790
	Cons14	20	29	19	735	766	743	−0.2110	−0.8790	−0.1210
N3	*MetE*	14	28	9	679	632	639	0.2656	0.5876	0.5736
	Cons109	20	28	14	721	746	777	−0.6510	−0.8590	−0.4200
	Cons14	20	28	17	699	710	717	−1.1300	−1.2760	−1.3400
N4	*MetE*	12	26	8	792	799	625	1.1926	0.7400	0.9778
	Cons109	20	27	19	745	763	750	−0.0470	−0.5690	−0.7310
	Cons14	20	25	16	732	785	641	1.1230	1.1100	0.9710
N5	*MetE*	12	25	7	588	645	612	−1.2220	−1.1140	1.4680
	Cons109	20	26	11	697	699	744	−1.1210	−0.9860	−0.9920
	Cons14	20	24	14	677	756	633	1.1120	0.9770	0.9927
S1	*MetE*	14	26	8	620	699	594	0.1888	−0.0353	0.0295
	Cons109	26	27	17	700	743	697	0.2780	0.7860	0.8710
	Cons14	24	25	15	614	675	609	0.5430	0.7790	0.9880
S2	*MetE*	12	22	5	786	798	529	2.1532*	1.3665*	1.7890*
	Cons109	26	22	10	761	782	568	2.0490*	1.7200*	1.9800*
	Cons14	20	23	9	778	796	589	2.0080*	1.7700*	1.6900*
S3	*MetE*	14	21	6	309	444	456	−1.3535	−2.1183	−2.1900
	Cons109	24	25	12	598	652	660	−0.9770	−0.8210	−0.9110
	Cons14	20	28	10	710	734	717	−0.0120	−0.7880	−0.9760
S4	*MetE*	14	26	11	618	643	594	0.1774	−0.3968	−0.2759
	Cons109	20	24	13	691	678	666	0.7860	−0.1200	−0.2310
	Cons14	20	24	13	669	721	615	0.5980	0.7830	0.9180
Pool^f	*MetE*	120	66	56	667	711	889	−0.8623	−3.3344*	−2.7100*
	Cons109	206	79	66	721	768	1318	−0.9970	−0.9210	−0.8910
	Cons14	184	84	65	699	715	1319	−0.9699	−0.8870	−0.5230
	All^g	510	229		693	780	960			

[a] Number of sequences sampled.
[b] Number of segregating sites.
[c] Number of haplotypes.
[d] Parameters of nucleotide variation were multiplied by a factor of 10^5.
[e] An asterisk (*) is labeled when a statistic value is significant at a 95% confidence level ($P<0.05$).
[f] Combined sequence data of all 9 populations.
[g] Sequence of all 3 loci (3500 bp) was used for analysis.

Table 3 Summary of F_{ST} and tests of genetic differentiation

Locus	*Pool*	*Southernmost populations*	*Northernmost populations*
MetE	0.078*	0.097*	0.070*
Cons109	0.034*	0.046*	0.044*
Cons14	0.077*	0.072*	0.078*
Combined 3 loci	0.091*	0.087*	0.093*
Short EST loci	0.023 * (0.001–0.033)	0.044* (0.006–0.053)	0.037* (0.009–0.051)
Average of all loci	0.053	0.065	0.057

Wright's fixation index (F_{ST}) was estimated based on the sequence data for the 3 loci *MetE*, Cons109, and Cons14 and the SNP data for the 8 short loci across populations examined (9 populations for the 3 loci and 6 populations for the 8 short loci) by the method of Hudson *et al.* (1992) and with FSTAT, respectively. Probabilities of the statistical test of genetic differentiation (*P*) were estimated based on Hudson's (2000) *S*nn statistic, evaluated by 10 000 permutations for the 3 loci; significance of F_{ST} for the 8 short loci was evaluated with FSTAT (1000 randomizations). The numbers in parentheses indicate 99% confidence intervals of F_{ST}; an asterisk (*) indicates statistical significance at a 95% confidence level ($P<0.05$).

in the pool). When the sequence data of all three loci are combined, $\pi_{\text{tot}}=0.00693$, $\pi_{\text{sil}}=0.00780$ and $\theta_{\text{W}}=0.00961$ in the pool. As shown in Table 4, π was also estimated from the eight short loci based on their SNP data. Because of the short sequence lengths, we combined the data of all the eight loci to estimate π at total sites (2529 bp). A total of 20 sequences per locus per population were used for the estimation. In the pool of six populations, π_{tot} was estimated to be 0.00598 (±0.00100; Table 4). This value is quite similar to that estimated from the combined sequence data of the three long loci mentioned above.

Being more recently colonized, the northern Japan populations are expected to have a lower level of genetic diversity than the southern populations. However, our calculations showed that nucleotide diversity in the three northernmost populations (N1–N3) was not lower than that in the three southernmost populations (S2–S4) for all

Table 4 Summary of nucleotide variation in the eight short EST loci

Population	*N1*	*N2*	*N3*	*S1*	*S2*	*S3*	*Pool*
π^a (s.d.)	618.72 (1.73)	655.14 (0.49)	618.93 (2.38)	570.44 (3.28)	564.00 (3.82)	578.10 (3.84)	598.15 (1.00)
HKA^b	0.991	0.932	0.897	0.954	0.939	0.984	0.933

[a] π and its standard deviation (s.d.) were multiplied by a factor of 10^5.
[b] Probability of the multilocus HKA test (Hudson *et al*., 1987) for each population and the pool.

Table 5 Summary of levels of recombination

	N1	*N2*	*N3*	*N4*	*N5*	*S1*	*S2*	*S3*	*S4*	*Pool*
R_M^{a}	9	9	9	12	2	2	3	1	4	36
No. of informative sites	54	51	66	63	57	51	63	15	45	111
R_M/no. of informative sites	0.20	0.18	0.14	0.19	0.04	0.04	0.05	0.07	0.09	0.32

[a] R_M, the minimum number of recombination events (Hudson and Kaplan, 1985).

three loci, *MetE*, Cons109 and Cons14 as well as the short loci (Table 4); it is also clearly seen in Tables 2 and 3 that all genetic diversity parameters, that is, π_{tot}, π_{sil} and θ_W are generally not lower in the northern populations than in the southern populations.

Recombination and linkage disequilibrium

In this article, we calculated the minimum number of recombination events, R_M, per informative site estimated from the combined data of all the three loci, *MetE*, Cons109 and Cons14 (3500 bp), to compare the level of recombination among populations. As shown in Table 5, R_M per informative site for the three loci was 0.32 in the pool, and ranged from 0.04 to 0.20 among populations. The levels of recombination appeared higher in the four northernmost populations (N1–N4) than in the remaining populations, despite that the statistical test supported a significant or marginally significant difference with N5, S1 and S2 populations only ($P<0.10$, two-tailed Fisher's exact test).

In theory, when recombination drives allelic diversity, LD between polymorphic sites is expected to decline with increasing distance along the chromosome. To understand this relationship, we compared the number of significant comparisons among sites >400 bp apart and <400 bp apart for each locus in the pool. The results of all three loci showed that both the average values of r^2 and $|D'|$ were significantly larger among sites <400 bp apart, and the percentage of significant pairwise comparison was also significantly higher ($P<0.05$) in this group (Figure 2; Supplementary Tables S2, S3 and S4), suggesting that on average, LD significantly declined among sites >400 bp apart in species-wide samples. However, LD did not decline significantly among sites >400 bp apart within populations, except for the three northernmost populations (N1–N3; Figure 2; Supplementary Tables S2, S3 and S4). The patterns of LD decay were similar when the low-frequency alleles (<10%) were included or excluded from the analysis.

Neutrality tests

To examine the fit of nucleotide polymorphism data to the neutral equilibrium model, we performed the tests of Tajima's *D* (Tajima, 1989a) and Fu and Li's *D* and F (Fu and Li, 1993) for each locus in each population separately and in the pool of nine populations. No significant Tajima's *D* and Fu and Li's *D* and F were observed at any of the three loci except for *MetE* in the pool (Fu and Li's *D* and F were significantly negative) and all the three loci in population S2 (Tajima's *D* and Fu and Li's *D* and F were all significantly positive; Table 2). It is interesting to note that Tajima's *D* is negative at all the three loci in the pool but positive in three out of four populations in southern Japan (S1, S2 and S4). The multilocus HKA test was used to assess departures from the neutral model for the eight short EST loci. We observed no evidence for any significant deviations in any population and in the pool ($P>0.05$; Table 4).

Discussion

Nucleotide variation

The occurrence frequency of SNP in our samples of *Q. crispula* is 1 every ~25 bp, which is comparable to that of other plant species previously reported. For example, 1 SNP every ~50 bp was detected when the average of 1095 randomly selected gene fragments in maize was considered (Yamasaki *et al*., 2005); a frequency of 1 every 23 bp was scored when the average of 10 unlinked loci in rice was taken (Zhu *et al*., 2007); a frequency of 1 every 60 bp was obtained when the average of 5 loci in European aspen was considered (Ingvarsson, 2005). This is one of the first studies describing the level of nucleotide polymorphism in oak species. The LD mapping approach requires abundant SNPs in the target species. In this study, SNPs were detected in all gene fragments, even in those with only a couple of hundred base pairs in length. Thus, screening for adaptive variation in oaks can be efficient with short sequences by the genome-wide scan approach as performed in maize (Yamasaki *et al*., 2005).

The present study is an additional contribution to increase knowledge about nucleotide diversity in forest tree species. We use both π_{tol} and π_{sil} to compare the levels of nucleotide diversity among species for the following discussion because π_{sil} has not yet been reported for several tree species in previous studies (Supplementary Table S5). Nucleotide diversity is

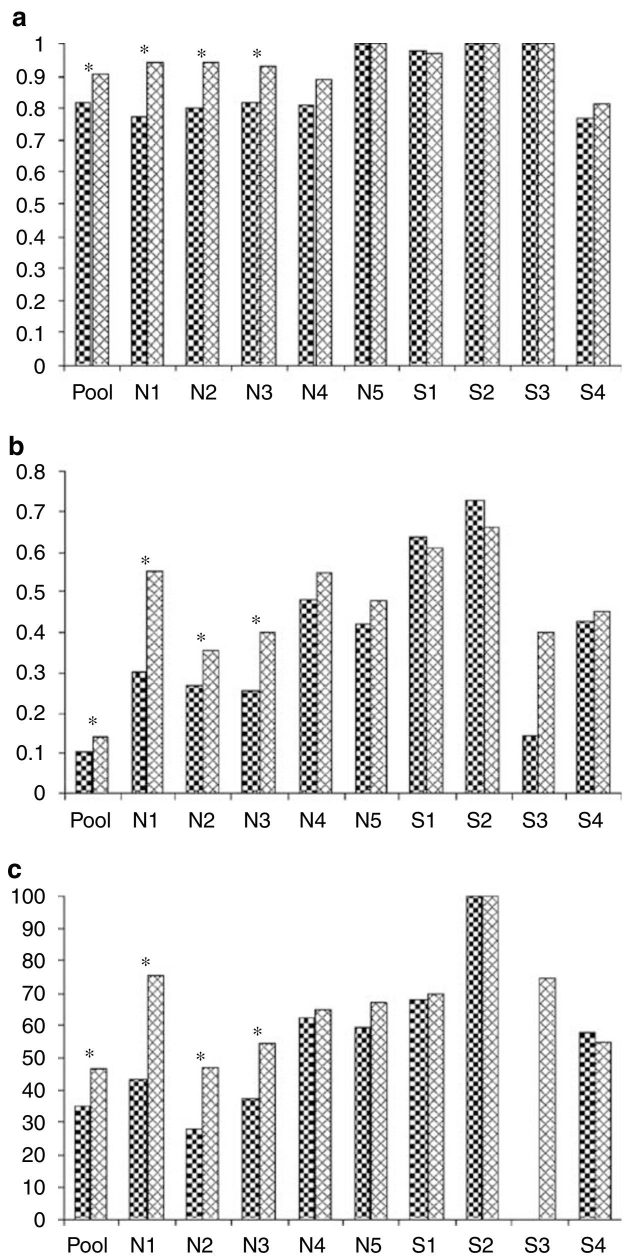

Figure 2 LD analysis of the species-wide sample (Pool) and different local populations (N1–S4) at the *MetE* locus. Average values estimated among sites >400 bp apart and <400 bp apart are indicated by the first and second columns, respectively, for the pool and for each population separately; (**a**) comparisons of $|D'|$, (**b**) comparisons of r^2 and (**c**) comparisons of percentage of significant pairwise tests; when the difference between the two groups is statistically significant ($P<0.05$), an asterisk (*) is labeled.

generally low in conifers such as *Pinus*, *Cryptomeria* and *Chamaecyparis* but relatively high in broadleaved trees such as *Populus* and *Quercus*. The nucleotide diversity of *Q. crispula* is generally higher than that of conifers, as high as that of a European oak, but lower than that of the European aspen (Supplementary Table S5).

Besides the mating system and life history traits, demographic history has a major impact on the level of within-population genetic diversity (Hamrick and Godt, 1996; Charlesworth, 2003). The results of this study imply that the northernmost populations do not harbor lower nucleotide diversity than the southernmost populations. This is interesting given the recent colonization history of the northern Japan populations. Theoretical models showed that the recently colonized populations are generally expected to show a decline in genetic diversity due to genetic bottlenecks and founder events (Slatekin, 1977; Pannell and Charlesworth, 1999; Pannell and Dorken, 2006). Possible alternative hypotheses will be discussed below together with population recombination and LD decay.

Mating system and population structures

Given the differences in the current patterns of population distribution and the density of the forests in northern and southern Japan, the level of inbreeding may be expected to be higher in the southern populations at some degree. Nevertheless, inbreeding indices were not significantly different from zero in all the populations sampled.

The average F_{ST} was estimated to be 0.053 in this study; the equivalent measurement of F_{ST} was 0.013 as estimated from microsatellite data in a previous study (Ohsawa *et al*., 2007). Even though these two F_{ST} estimates were substantially different, which was probably due to sampling strategies (latitudinal gradient along Japan vs altitudinal gradient in a narrow latitudinal range) and the genetic nature (biallelic vs multiple allelic polymorphism and evolutionary mechanisms) of the markers, both studies consistently revealed a modest level of genetic differentiation among populations. *Q. crispula* is an obligate outcrossing and wind-pollinated long-lived forest tree and has seeds that may be dispersed over a long distance by birds and rodents (Watanabe, 1994). Hence, the low levels of population differentiation observed among populations of this species are consistent with a general observation for outbreeding and wind-pollinated species (Hamrick and Godt, 1990). Nonetheless, low F_{ST} was unexpectedly observed among the recently colonized populations (northernmost populations). Several theoretical studies showed that colonization frequently leads to an excess of differentiation among populations compared with the equilibrium situation (Austerlitz *et al*., 1997; Le Corre and Kremer, 1998); however, with time, genetic differentiation is expected to decrease because of migration between extant populations (Pannell and Dorken, 2006).

Tajima's *D*-values of all the three loci were negative in the pool but positive or even significantly positive in several southern Japan populations. This suggests an excess of low-frequency polymorphisms in the species-wide samples and an excess of intermediate-frequency polymorphisms in the populations from southern Japan. The excess of low-frequency or intermediate-frequency polymorphisms is consistent across three loci, thus natural selection seems not to be the evolutionary force causing these skews in the frequency spectrum. The excess of intermediate-frequency polymorphisms in the populations from the south appears to be a genetic consequence of a recent reduction in population size (Tajima, 1989b), because of the preferential loss of low-frequency polymorphisms. The observation of more negative Tajima's *D* in the pooled sample as opposed to the within-populations of *Q. crispula* in this study was similar to that in *Populus* (Ingvarsson, 2005) and wild tomatoes (Arunyawat *et al*., 2007). This excess of low-frequency polymorphism can be explained by species range expansion in Japan, probably the recent northern colonization.

Furthermore, palynological studies have suggested that oak forests covered a broad area in southern Japan during LGM and have just recently reached the northern areas. In southern Japan, the frequency of oak pollen rapidly increased 12 000 years ago and then began to decrease from 8500 years ago (Miyoshi, 1998; Takahara, 1998). However, in a population from southern Japan (S3), we did not observe the excess of intermediate-frequency polymorphisms; this may suggest that either no such significant reduction in population size occurred in this population or singleton mutations have been reaccumulated in this population owing to gene flow from other populations within a time period of approximately 8500 years.

Recombination and linkage disequilibrium

We showed that the level of recombination was relatively high in *Q. crispula* as R_M per informative site was 0.32 for the species-wide samples. The values of R_M per informative site at seven loci in *Arabidopsis thaliana* were lower and ranged from 0.08 to 0.15 (Kuittinen and Aguage, 2000); however, these values were still much higher than those in *Cryptomeria japonica,* which ranged from 0.0000 to 0.0526 (Kado *et al*., 2003); R_M per informative site in the *Pal*1 locus of *Pinus sylvestris* was 0.25 (Dvornyk *et al*., 2002). Note that *A. thaliana* is mostly selfing and thus seems to have very low effective recombination rates (Gaut and Long, 2003). The recombination rate in *C. japonica* was about one-thirtieth of that in *A. thaliana* on average (Kado *et al*., 2003). *P. sylvestris* is among the tree species with the widest-known geographic distributions in the world, whereas *Q. crispula* is restricted to some East Asian areas; however, the levels of both nucleotide diversity and population recombination were generally higher in this oak than the conifer. This is attributable to the higher mutation and recombination rates in oaks than in conifers. Alternatively, past and/or ongoing hybridization among closely related oak species, such as between *Q. crispula* and *Q. dentata* (Ishida *et al*., 2003) and between *Q. crispula* and *Q. serrata* (Kanazashi *et al*., 1997) in Japan may be one factor maintaining the high levels of nucleotide variation and recombination in oaks.

This study together with a previous study (Ingvarsson, 2005) showed that broadleaved trees such as oaks and aspens may show generally higher levels of species-wide nucleotide diversity and recombination than conifers. Theoretical and empirical studies have shown that the mating system primarily affects diversity at the population level rather than at the species level (Savolainen *et al*., 2000; Ingvarsson, 2002; Charlesworth, 2003; Wright *et al*., 2003). Oaks, aspens and conifers are all mostly outcrossing long-lived trees. Thus, differences in life history traits might be an important factor responsible for such different levels of nucleotide diversity and recombination between broadleaved and coniferous species.

Perhaps, one of our most interesting findings is that the three northernmost populations harbor a relatively large amount of nucleotide diversity, higher levels of recombination and faster LD decay than the southern populations in Japan. This result is contradictory to theoretical expectations given the recent colonization history of the northern Japan populations. Several possibilities should be considered to explain this paradox. The first is that effective population size is larger in the recently colonized populations than in the southern populations due to a recent population expansion that occurred in the former and/or a recent population bottleneck in the latter. Statistical tests and palynological data appear to favor the hypothesis of a recent population bottleneck in southern Japan populations.

A second possibility could be natural selection. Purifying selection reduces genetic diversity and the southern populations are more intensively selected than the northern populations; the LD decay was shown to be significantly slower in the former than in the latter, thereby causing hitchhiking and background selection more effective (Kaplan *et al*., 1989; Begun and Aquadro, 1992; Charlesworth *et al*., 1993). However, the possibility of selection is not supported by statistical tests in this study. It should be noted that, because of the generally low statistical power of most tests (Wright and Gaut, 2005), the effects of selection cannot be excluded entirely for all loci.

The relatively high levels of nucleotide polymorphism and recombination in the recently colonized populations in northern Japan may also be explained by several possibilities. Firstly, gene flow subsequent to colonization might be sufficiently strong to increase genetic diversity in the populations when they reach demographic equilibrium (Pannell and Dorken, 2006). Secondly, the long lifespan and juvenile phase of trees together with long-distance seed dispersal were shown to strongly attenuate founder effects during colonization (Austerlitz *et al*., 1997; Austerlitz and Garnier-Gere, 2003). Lastly, multiple colonization events causing an admixture of different lineages may overcome bottlenecks in recently colonized populations, particularly when propagules have long residence time or retain high genetic diversity as do the acorns of oaks (Lavergne and Molofsky, 2007).

Although further studies are needed to test each hypothesis mentioned above, the current landscape of genetic diversity, recombination and LD in the natural populations of *Q. crispula* characterized in this study may provide primarily information for conservation genetics and LD mapping studies for this species and facilitate molecular evolution studies in oaks.

Acknowledgements

We thank Professor John Pannell (Oxford University), Dr NM Tue (CMES, Ehime University), the subject editor and three anonymous referees for their helpful discussion and critical reading of our article. This research was supported by a Grant-in-Aid (No. 16380104) from the Japan Society for the Promotion of Science.

References

Aguade M (2001). Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. *Mol Biol Evol* **18**: 1–9.

Arunyawat U, Stephant W, Städler T (2007). Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol Biol Evol* **24**: 2310–2322.

Austerlitz F, Garnier-Gere PH (2003). Modelling the impact of colonization on genetic diversity and differentiation of forest trees: interaction of life cycle, pollen flow and seed long-distance dispersal. *Heredity* **90**: 282–290.

Austerlitz F, Jung-Muller B, Godelle B, Gouyon PH (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theor Popul Biol* **51**: 148–164.

Begun D, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–529.

Casasoli M, Derory J, Morera-Dutrey C, Brendel O, Porth I, Guehl JM *et al.* (2006). Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. *Genetics* **172**: 533–546.

Charlesworth B, Morgan TH, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Charlesworth D (2003). Effects of inbreeding on the genetic diversity of populations. *Philos Trans R Soc Lond Biol B Biol Sci* **358**: 1051–1070.

Chiang Y, Ge CXJ, Chou CH, Wu WL, Chiang TY (2002). Nucleotide sequence diversity at the *methionine synthase* locus in endangered *Dunnia sinensis* (Rubiaceae): an evaluation of the positive selection hypothesis. *Mol Biol Evol* **19**: 1367–1375.

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002). Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. *Mol Biol Evol* **19**: 179–188.

Filatov DA, Charlesworth D (1999). DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. *Genetics* **147**: 915–925.

Fu YX, Li WH (1993). Statistical tests of neutrality of mutatons. *Genetics* **133**: 693–709.

Garcia-Gil MR, Mikkonen M, Savolainen O (2003). Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Mol Ecol* **12**: 1195–1206.

Gaut BS, Long AD (2003). The lowdown on linkage disequilibrium. *Plant Cell* **15**: 1502–1506.

Goudet J (2001). FSTAT (version 2.9.3): a program to estimate and tests gene diversities and fixation indices. http://www2.unil.ch/popgen/softwares/fstat.htm.

Hamrick JL, Godt MJW (1990). Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL and Weir BS (eds). *Plant Population Genetics, Breeding and Genetic Resources*. Sinauer Associates: Sunderland, MA.pp 43–63.

Hamrick JL, Godt MJW (1996). Effects of life history traits on genetic diversity in plant species. *Philos Trans R Soc Lond B Biol Sci* **351**: 1291–1298.

Hibino K, Takeuchi S (1998). Vegetation history of the Tohoku district. In: Yasuda Y and Miyoshi N (eds). *Illustration: Vegetation History of the Japanese Archipelago*. Asakura-shoten: Tokyo, Japan. (in Japanese) 62–72.

Horikawa Y (1972). *Atlas of the Japanese flora, an Introduction to Plant Sociology of East Asia*. Gakken Co. Ltd: Tokyo, Japan.

Hudson RR (2000). A new tests statistic for detecting genetics differentiation. *Genetics* **155**: 2011–2014.

Hudson RR, Boos DD, Kaplan NL (1992). A statistical test for detecting geographic subdivision. *Mol Biol Evol* **9**: 138–151.

Hudson RR, Kaplan NL (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hudson RR, Kreitman M, Aguade M (1987). A tests of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.

Ingvarsson P (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L. Salicaceae). *Genetics* **169**: 945–953.

Ingvarsson PK (2002). A metapopulation perspective of genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* **56**: 2368–2373.

Innan H, Tajima F, Terauchi R, Miyashita T (1996). Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* **143**: 1761–1770.

Ishida TA, Hattori K, Sato H, Kimura MA (2003). Differentiation and hybridization between *Quercus crispula* and *Q. dentate* (Fagaceae): insights from morphological traits, amplified fragment length polymorphism markers, and leafminer composition. *Am J Bot* **90**: 769–776.

Kado T, Matsumoto A, Ujino-Ihara T, Tsumura Y (2008). Amounts and patterns of nucleotide variation within and between two Japanese conifers, sugi (*Cryptomeria japonica*) and hinoki (*Chamaecyparis obtuse*) (Cupressaceae *sensu* lato). *Tree Genet Genomes* **4**: 133–141.

Kado T, Yoshimaru H, Tsumura Y, Tachida H (2003). DNA variation in a conifer, *Cryptomeria japonica* (Copressaceae *sensu* lato). *Genetics* **164**: 1547–1599.

Kanazashi A, Kanazashi T, Kawahara T (1997). Evidence of reproductive barriers between closely related species of *Quercus*. In: Steiner KC (ed). *Diversity and Adaptation in Oak Species*. Proc 2nd Meeting of Working Party 2.08.05. Genetics Quercus IUFRO.

Kanno M, Yokoyama J, Suyama S, Ohyama M, Itoh T, Suzuki M (2004). Geographical distribution of two haplotypes of chloroplast DNA in four oak species (*Quercus*) in Japan. *J Plant Res* **117**: 311–317.

Kaplan NL, Hudson RR, Langley CH (1989). The 'hitchhiking effect' revisited. *Genetics* **123**: 887–899.

Kubitzki K (1993). Fagaceae. In: Kubitzki K, Rohwer JG, Bittrich V (eds). *The Families and Genera of Vascular Plants*, Vol. 2. Springer: Berlin. pp 301–309.

Kuittinen H, Aguage M (2000). Nucleotide variation at the *chalcone isomerase* locus in *Arabidopsis thaliana*. *Genetics* **155**: 863–872.

Lavergne S, Molofsky J (2007). Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proc Natl Acad Sci USA* **10**: 3883–3888.

Le Corre V, Kremer A (1998). Cumulative effects of founding events during colonization on genetic diversity and differentiation in an island and stepping stone model. *J Ecol Biol* **11**: 495–512.

Miyoshi N (1998). Vegetation History of the Chugoku and Shikoku districts. In: Yasuda Y and Miyoshi N (eds). *Illustration: Vegetation History of the Japanese Archipelago*. Asakura-shoten: Tokyo, Japan. (in Japanese) 138–149.

Morita Y, Choi K, Hibino K (1998). Vegetation history of the Chubu and Tokai districts. In: Yasuda Y and Miyoshi N (eds). *Illustration: Vegetation History of the Japanese Archipelago*. Asakura-shoten: Tokyo, Japan. (in Japanese) 92–102.

Neale DB, Savolainen O (2004). Association genetics of complex traits in conifers. *Trends Plant Sci* **9**: 325–330.

Nixon KC (1993). Infrageneric classification of *Quercus* (Fagaceae) and typification of sectional names. *Ann Sci Forest* **50** (Suppl. 1): 25s–34s.

Ohsawa T, Tsuda Y, Saito Y, Sawada H, Ide Y (2007). Altitudinal genetic diversity and differentiation of *Quercus crispula* in the Chichibu mountains, central Japan. *Int J Plant Sci* **168**: 333–340.

Okaura T, Harada K (2002). Phylogeographical structure revealed by chloroplast DNA variation in Japanese Beech (*Fagus crenata* Blume). *Heredity* **88**: 322–329.

Okaura T, Quang ND, Ubukata M, Harada K (2007). Phylogeographic structure and late quaternary population history of the Japanese oak *Quercus mongolica* var. *crispula* and related species revealed by chloroplast DNA variation. *Genes Genet Syst* **82**: 465–477.

Pannell JR, Charlesworth B (1999). Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution* **53**: 664–676.

Pannell JR, Dorken ME (2006). Colonisation as a common denominator in plant metapopulations and range expansions: effects on genetic diversity and sexual systems. *Landscape Ecol* **21**: 837–848.

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Gere P, Cato S *et al.* (2005). Nucleotide variation in genes involved in wood formation in two pine species. *New Phytol* **167**: 101–112.

Remington DL, Thornsberry JM, Matsoura Y, Wilson LM, Whitt SR, Doebley J *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* **98**: 11479–11484.

Rieseberg LH, Wendel JF (1993). Introgression and its consequences in plants. In: Harrison RG (ed). *Hybrid Zones and the Evolutionary Process*. Oxford University Press: Oxford. pp 70–109.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Savolainen O, Langley CH, Lazzaro BP, Fr H (2000). Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Mol Biol Evol* **17**: 645–655.

Slatekin M (1977). Gene flow and genetic drift in a species subject to frequent local extinctions. *Theor Popul Biol* **12**: 253–262.

Tajima F (1989a). Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Tajima F (1989b). The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.

Takahara H (1998). Vegetation history of the Kinki district. In: Yasuda Y and Miyoshi N (eds). *Illustration: Vegetation History of the Japanese Archipelago*. Asakura-shoten: Tokyo, Japan. (in Japanese) 114–134.

Van VL (1976). Ecological species, multispecies, and oaks. *Taxon* **2–3**: 769–776.

Watanabe S (1994). *Specia of Trees*. Tokyo University Press: Tokyo. (in Japanese) 450 pp.

Wright SI, Gaut BS (2005). Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* **22**: 506–519.

Wright SI, Lauga B, Charlesworth D (2003). Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Mol Ecol* **12**: 1247–1263.

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF *et al.* (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**: 2859–2872.

Zhang L-B, Ge S (2006). Multiple analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. *Mol Biol Evol* **24**: 769–783.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007). Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: Severe bottleneck during domestication of rice. *Mol Biol Evol* **24**: 875–888.

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