SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Demographic history and adaptive synonymous and nonsynonymous variants of nuclear genes in *Rhododendron oldhamii* (Ericaceae)

Yi-Chiang Hsieh[1], Chung-Te Chang[2], Jeng-Der Chung[3] & Shih-Ying Hwang[1✉]

Demographic events are important in shaping the population genetic structure and exon variation can play roles in adaptive divergence. Twelve nuclear genes were used to investigate the species-level phylogeography of *Rhododendron oldhamii,* test the difference in the average GC content of coding sites and of third codon positions with that of surrounding non-coding regions, and test exon variants associated with environmental variables. Spatial expansion was suggested by R_2 index of the aligned intron sequences of all genes of the regional samples and sum of squared deviations statistic of the aligned intron sequences of all genes individually and of all genes of the regional and pooled samples. The level of genetic differentiation was significantly different between regional samples. Significantly lower and higher average GC contents across 94 sequences of the 12 genes at third codon positions of coding sequences than that of surrounding non-coding regions were found. We found seven exon variants associated strongly with environmental variables. Our results demonstrated spatial expansion of *R. oldhamii* in the late Pleistocene and the optimal third codon position could end in A or T rather than G or C as frequent alleles and could have been important for adaptive divergence in *R. oldhamii.*

Spatial and temporal patterns underlie population demographic processes of plant species and the phylogenetic relationship between and within species can be revealed by using chloroplast and nuclear DNA sequence data[1–3]. Molecular techniques, such as amplified fragment length polymorphisms (AFLPs), expressed sequence tag simple sequence repeats (EST-SSRs), and methylation-sensitive amplification polymorphisms (MSAPs) have been commonly employed in investigation testing for environmentally dependent local adaptation[4–6]. Nonetheless, nuclear gene sequences can be amplified and sequenced spanning coding and non-coding regions. The variation in non-coding and coding sequences in nuclear genes can be used, respectively, in investigation of phylogeny and phylogeography and in testing for correlation with environmental variables contributing to adaptive evolution[4,7,8]. Within the coding region of a gene, the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site is commonly used for the inference of adaptive evolution of genes driven by natural selection[9–11], and synonymous substitutions are thought to be inconsequential because of the conservative nature of amino acids. However, synonymous substitutions can have significant effects on gene expression, protein folding, and protein cellular function[12,13], and hence synonymous substitutions may not be "silent"[14].

The difference in the relative frequency of synonymous codons for individual amino acids in protein coding sequences is coined as codon usage bias. Codon bias can vary among species and/or among genes of a genome and may be derived via mutational bias processes, GC biased conversion, or driven by selection co-adapting with

[1]School of Life Science, National Taiwan Normal University, 88 Tingchow Road, Section 4, Taipei 11677, Taiwan. [2]Department of Life Science, Tunghai University, 1727 Taiwan Boulevard, Section 4, Taichung 40704, Taiwan. [3]Division of Silviculture, Taiwan Forestry Research Institute, 53 Nanhai Road, Taipei 10066, Taiwan. ✉email: hsy9347@ntnu.edu.tw

tRNAs in optimizing the efficiency and accuracy of translation[15–17]. Theory suggests that the strength of natural selection on synonymous sites may be weak and effective population size is thought to be the determining factor for natural selection to be effective on codon usage pattern[18,19], and synonymous substitution is subjected to strong purifying selection[20,21]. Nonetheless, selection acting on codon bias has been found in prokaryotes with large effective population sizes[22,23] and in eukaryotic species with low effective population sizes[24–26].

In genes with high codon bias, the "preferred codons" often end in either C or G according to major codon preference model[27,28]. The level of GC content at third codon positions is considered as an indicator reflecting codon usage pattern, and the level of gene expression has been found to be positively correlated with the level of GC content at third codon positions[29,30]. Studies have shown that overall codon bias is more phenomenal in monocots than in dicots and, respectively, tended to use C/G and A/T at third codon positions[31,32]. However, optimal codons tend to end in G or C in dicots, resulting in higher average GC content at coding sites and at third codon positions compared with surrounding non-coding regions[32,33].

Rhododendron oldhamii Maximowicz belongs to the subgenus *Tsutsusi* of Ericaceae is an endemic dicot species widely, but fragmented distributed in the lowlands and up to 2,500 m in the humid understory of broadleaf forests in Taiwan. Heterogeneity in flowering times of *R. oldhamii* populations located in different geographic areas, across the species' distribution range, have been shown either via examination of herbarium specimen records[34] or by field studies[35]. Differential flowering times of species distributed in different geographic areas can result in reproductive isolation or reproductive incompatibility[36], and in consequence a limited gene flow between populations[36,37]. Hsieh et al.[4] showed that gene dispersal was limited within geographic regions of *R. oldhamii* and inferred that the discontinuities of population distribution resulted from recent population bottlenecks during the Holocene based on EST-SSR data. Although organisms in isolated, small populations may be restricted in developing local adaptation in response to changing environments[38], study based on EST-SSR found population divergence at regional level in association with environmental variables underlying local adaptation in *R. oldhamii*[4]. Moreover, genetic and epigenetic variations based on AFLP and MSAP also found environmentally dependent adaptive divergence in populations of *R. oldhamii*[5]. *R. oldhamii* discontinuously distributed in a wide geographic range makes it an excellent exemplar endemic subtropical forest tree species occurring in Taiwan for investigating not only the population divergence associated with environmental differences, but also the phylogeographic history related to the current genetic structure of this species.

Since it has been shown that adaptive divergence driven by natural selection at local scales in *R. oldhamii* based on EST-SSRs[4], it is worthwhile to study further from the evolutionary perspective of natural selection on codon usage bias using DNA coding sequences. Additionally, the far-distant past demographic history shaping the genetic structure and distribution of *R. oldhamii* can be inferred using nuclear non-coding sequences in contrast to using frequency data of EST-SSRs reflecting more recent demography in the previous study[4]. Strong correlations of synonymous and nonsynonymous substitutions with environmental variables using regression approaches[7,8,39,40] may provide as evidence suggesting that they would have experienced the effects of selection. We hypothesized that both synonymous and nonsynonymous substitutions of nuclear coding sequences, particularly synonymous substitutions at third codon positions, in natural populations of *R. oldhamii* (Fig. 1) might have driven by natural selective forces in association with environmental heterogeneity. Using partial genomic DNA sequences of 12 haphazardly selected nuclear genes (Supplementary Table S1), we aimed to (1) investigate the species-level phylogeographic history of *R. oldhamii* based on intron sequences of multiple nuclear gene loci, (2) test the significant differences in the average GC content of coding sites and of third codon positions of nuclear coding sequences with that of surrounding non-coding regions, and (3) test the associations of synonymous and nonsynonymous variants in coding sequences with environmental variables.

Results

Genetic diversity. The sequences of 12 nuclear DNA loci (Supplementary Table S1) were obtained from 47 individuals across 18 populations (four geographic regions[4]) (Table 1). The length of aligned sequences of the pooled samples, including exon and intron sequences, for each locus ranged from 509 bp (*GAPC1*) to 908 bp (*PCFS4*) (Supplementary Table S2), and the total aligned length was 8871 bp. The total aligned length for exon and intron, respectively, was 2082 bp and 6798 bp. Lengths of exons and introns of the 12 genes, respectively, ranged from 36 bp (*GRP7*) to 372 bp (*LACS8*) and from 280 bp (*LHCA1*) to 859 bp (*GRP7*). The minimum number of recombination events (R_m) within gene ranged from 0 (*CPD*) to 10 (*GAPC1*) (Table 2). The number of haplotypes ranged from 2 to 16 and from 16 to 30, respectively, for the 18 populations and the four geographic regions. The number of haplotypes for the total aligned sequences was 94 (Table 1). Nucleotide diversity π ranged from 0.00163 (population CJ) to 0.00785 (population EGS) and from 0.00586 (southeast) to 0.00659 (central), respectively, for the 18 populations and the four geographic regions. Waterson's nucleotide diversity measure, θ_w, based on segregating sites ranged from 0.00163 (population CJ) to 0.00785 (population EGS) and from 0.00552 (south) to 0.00695 (central) for the 18 populations and the four geographic regions, respectively. The nucleotide polymorphisms (π and θ_w) for individual gene in each population were reported in Supplementary Tables S3 and S4. Friedman test revealed no significant difference in regional comparisons neither in π nor in θ_w ($\chi^2 = 0.7$, $P = 0.873$ in both cases). However, significant differences of π and θ_w among populations were found ($\chi^2 = 35.10$, $P = 0.006$; $\chi^2 = 35.33$, $P = 0.0056$, respectively). Moreover, significant differences of π and θ_w were found in pairwise population comparisons after correction for multiple comparisons at false discovery rate (FDR) of 5%, but no significant difference was found in between region comparisons (Supplementary Table S5). Significant positive inbreeding coefficient (F_{IS}), indicative of departure from Hardy–Weinberg equilibrium, representing homozygote excess, was found for all populations that were estimable (sample size > 1) except the population LLK (Table 1).
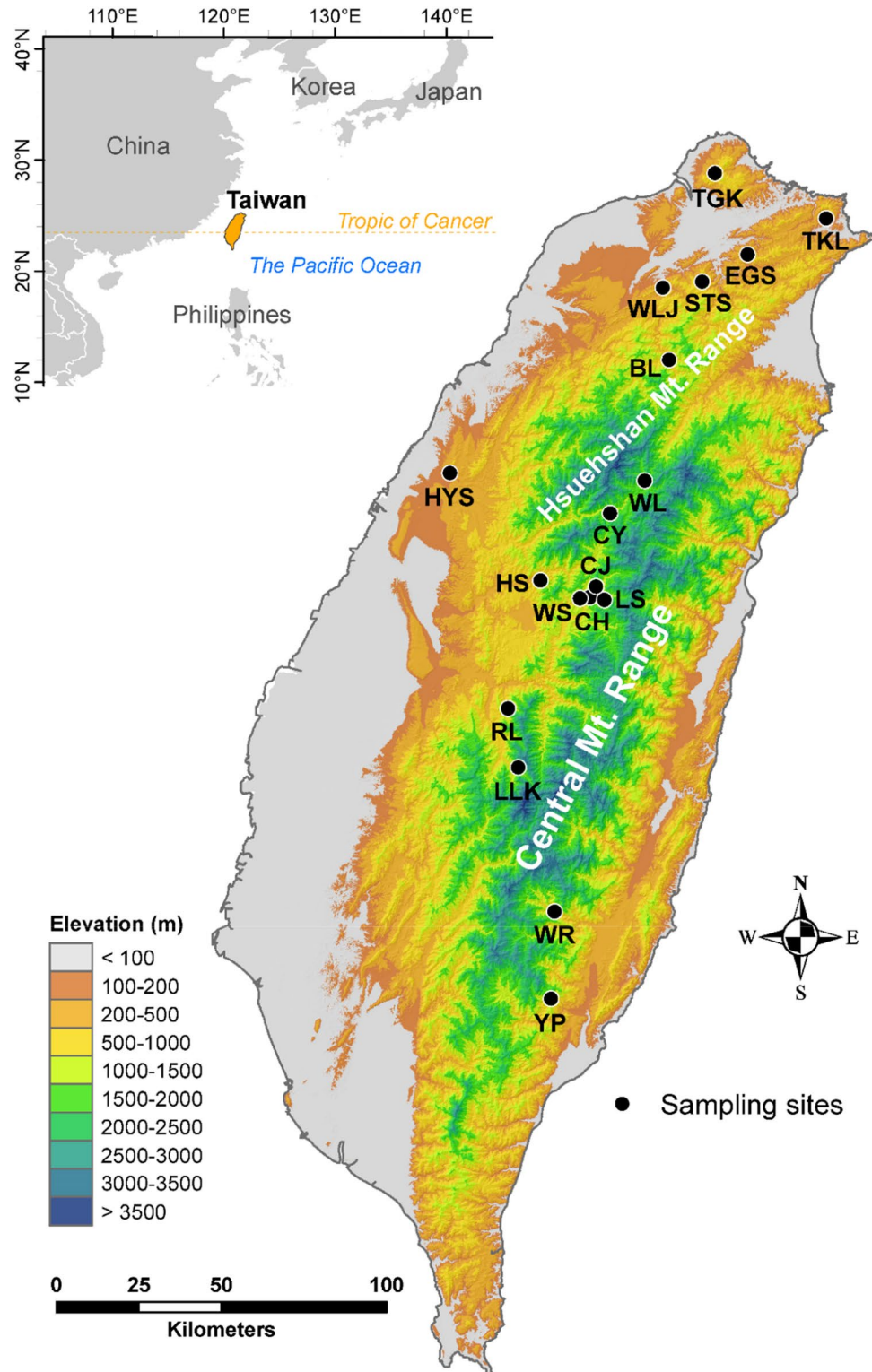
Figure 1. Sample locations of the 18 populations of *Rhododendron oldhamii* distributed in Taiwan. The 18 populations were assigned to four geographic regions according to EST-SSR clustering (Table 1)[4]. The four geographic groups were north group (populations BL, EGS, HYS, STS, TGK, TKL, and WLJ), central group (population WL), south group (populations CH, CJ, CY, HS, LLK, LS, RL, and WS), and southeast group (populations WR and YP). See Table 1 for population code abbreviations. We generated the map using ArcGIS v.10.6. The ASTER GDEM (Global Digital Elevation Model; https://asterweb.jpl.nasa.gov/gdem.asp) is used for elevational background.

Population (code)	Region	N	Locality (°E/°N)	N_h	π (SD)	θ_w (SD)	F_{IS} (95% CI)
Population							
Baling (BL)	North	4	121.38/24.68	8	0.00596 (0.00067)	0.00554 (0.00243)	**0.097** (0.025, 0.169)
Ergirshan (EGS)	North	1	121.62/24.97	2	0.00785 (0.00392)	0.00785 (0.0056)	–
Huoyanshan (HYS)	North	5	120.73/24.37	10	0.00549 (0.00056)	0.00530 (0.00218)	**0.423** (0.339, 0.508)
Shihtoushan (STS)	North	1	121.48/24.89	2	0.00563 (0.00281)	0.00563 (0.00403)	–
Tsaigongkeng (TGK)	North	1	121.52/25.19	2	0.00370 (0.00185)	0.00370 (0.00267)	–
Tsanguanliao (TKL)	North	1	121.86/25.06	2	0.00399 (0.002)	0.00399 (0.00288)	–
Wuliaojian (WLJ)	North	1	121.37/24.88	2	0.00429 (0.00215)	0.00429 (0.00309)	–
Wuling (WL)	Central	8	121.31/24.35	16	0.00668 (0.001)	0.00699 (0.00253)	**0.670** (0.619, 0.721)
Chungheng (CH)	South	1	121.15/24.03	2	0.00341 (0.00171)	0.00341 (0.00246)	–
Chingjing (CJ)	South	1	121.16/24.06	2	0.00163 (0.00081)	0.00163 (0.0012)	–
Chiayang (CY)	South	3	121.21/24.26	6	0.00681 (0.00096)	0.00650 (0.0031)	**0.340** (0.267, 0.414)
Hueisun (HS)	South	2	121.00/24.08	4	0.00655 (0.00174)	0.00671 (0.00366)	**0.133** (0.033, 0.234)
Leleku (LLK)	South	4	120.93/23.56	8	0.00613 (0.00069)	0.00555 (0.00243)	– 0.026 (– 0.095, 0.044)
Lushan (LS)	South	2	121.19/24.02	4	0.00733 (0.00185)	0.00729 (0.00397)	**0.116** (0.012, 0.220)
Renluen (RL)	South	1	120.90/23.73	2	0.00237 (0.00119)	0.00237 (0.00173)	–
Wushe (WS)	South	1	121.12/24.03	2	0.00756 (0.00378)	0.00756 (0.0054)	–
Wuru (WR)	Southeast	5	121.04/23.17	10	0.00506 (0.00071)	0.00507 (0.00209)	**0.191** (0.115, 0.266)
Yeinping (YP)	Southeast	5	121.03/22.93	10	0.00578 (0.00074)	0.00561 (0.00231)	**0.265** (0.176, 0.354)
Region							
North		14		28	0.00594 (0.00032)	0.00626 (0.002)	
Central		8		16	0.00659 (0.001)	0.00695 (0.00252)	
South		15		30	0.00605 (0.00034)	0.00552 (0.00174)	
Southeast		10		20	0.00586 (0.00043)	0.00606 (0.00209)	
Total		47		94	0.00659 (0.00027)	0.00945 (0.00235)	

Table 1. Population information and number of haplotypes, nucleotide diversity, and inbreeding coefficients (F_{IS}) based on the total aligned intron sequences of 12 nuclear genes for the 18 *Rhododendron oldhamii* populations. *N*, sample size; N_h, number of haplotypes; π, the average number of pairwise nucleotide differences per site; θ_w, the average nucleotide diversity of segregating site; F_{IS}, inbreeding coefficients. F_{IS} values do not bracket zero are in bold. Classification of populations into different geographic regions was based on the results of a previous study[6].

Demography and genetic structure.   Neutrality test statistics including Tajima's *D* and Fu and Li's *D** and *F** were mostly negative, albeit not significant, based on the aligned intron sequences of the 12 loci individually and the total aligned intron sequences of the pooled samples (Table 2). Consistent significant negative values of these statistics were only found for *ATMYB33*. Significant small R_2 values were found for the aligned intron sequences of the *AMP1*, *ATMYB33*, *CPD*, *GRP7*, and *PCFS4* genes of the pooled samples and for the total aligned intron sequences of the pooled samples. However, spatial expansion model was rejected by neutrality test statistics including *D*, *D**, *F**, and R_2 based on the total aligned intron sequences of the regional samples. Nonetheless, non-significant sum of square deviations (SSD) estimation revealed that the spatial expansion model could not be rejected based on the aligned intron sequences of the pooled samples of the 12 loci individually. Spatial expansion was also suggested by SSD when analyzed using the total aligned intron sequences of the regional and the pooled samples. Comparable sample sizes among regions and approximately equal migration rates among populations within regions were found based on the goodness-of-fit test for mismatch distribution under spatial expansion model (Supplementary Table S6).

The level of genetic differentiation was found to be significant when compared among regions (F_{ST} = 0.074, *P* = 0.001), but not significant when compared among populations (F_{ST} = -0.0078, *P* = 0.331). Significant pairwise F_{ST} was also found for between region comparisons (Supplementary Table S7). Additionally, genetic clustering using discriminant analysis of principal components (DAPC) showed no clear population or regional distinction (Supplementary Fig. S1).

The average GC content.   The average GC content at coding sites, at third codon positions, and in surrounding non-coding regions across 94 sequences of the 12 genes were 0.453, 0.419, and 0.356, respectively (Table 3). Paired Wilcoxon tests found that most of the average GC content at coding sites across 94 sequences of the 12 genes, except *GRP7*, was significantly (*P* < 0001) higher than the average GC content of surrounding non-coding regions (Table 3). Moreover, five (*AMP1, GRP7, LACS8, PCFS4*, and *SPA1*) and seven (*ATMYB33, CPD, GAPC1, HEME2, LHCA1, PMDH2*, and *SUI1*) of the 12 genes, respectively, showed significantly lower and higher average GC content across 94 sequences at third codon positions than that of surrounding non-coding regions.

Locus	Intron aligned length (bp)	Rm	S	π	θ_w	Hd	Nh	Neutrality test				
								D	D*	F*	R₂	SSD
AMP1	486	1	51	0.01059	0.03925	0.669	14	− **2.377**	1.348	− 0.191	**0.026**	0.00209
ATMYB33	695	2	24	0.00137	0.00689	0.427	16	− **2.394**	− **3.566**	− **3.732**	**0.023**	0.00143
CPD	727	0	13	0.00117	0.00351	0.641	14	− **1.810**	− 1.430	− 1.863	**0.033**	0.00207
GAPC1	356	10	26	0.01656	0.01456	0.732	22	0.413	0.777	0.763	0.110	0.01913
GRP7	859	1	26	0.00123	0.00593	0.578	21	− **2.427**	− 1.648	− **2.329**	**0.023**	0.00141
HEME2	633	4	41	0.01673	0.01270	0.830	28	0.999	− 1.256	− 0.432	0.126	0.02930
LACS8	332	1	11	0.00507	0.00672	0.713	14	− 0.645	− 1.197	− 1.192	0.072	0.02205
LHCA1	280	2	12	0.00372	0.00850	0.452	10	− 1.506	0.230	− 0.457	0.043	0.00329
PCFS4	830	5	32	0.00378	0.00763	0.875	27	− 1.558	− 1.454	− 1.790	**0.047**	0.00111
PMDH2	538	8	27	0.01046	0.01021	0.905	29	− 0.140	− 0.090	− 0.131	0.098	0.00823
SPA1	503	1	20	0.00527	0.00779	0.782	15	− 0.943	0.418	− 0.107	0.065	0.01036
SUI1	550	7	30	0.01287	0.01076	0.712	26	0.172	− 0.084	0.020	0.115	0.01506
Region												
North		24	164			1.000	28	− 0.265	− 0.280	− 0.325	0.112	0.00330
Central		11	150			1.000	16	− 0.359	0.203	0.049	0.127	0.01023
South		34	147			1.000	30	0.343	0.183	0.280	0.131	0.00381
Southeast		14	145			1.000	20	− 0.134	− 0.041	− 0.081	0.122	0.00616
Total	6798	51	313			1.000	94	− 1.094	− 0.750	− 1.079	**0.066**	0.00123

Table 2. Summary of nucleotide polymorphism and neutrality tests based on the aligned intron sequences of individual genes and the total aligned intron sequences for *Rhododendron oldhamii*. *Rm*, minimum number of recombination events; *S*, number of segregating sites; *π*, the average number of pairwise nucleotide diversity per site; *θ_w*, the average nucleotide diversity of segregating site; *Hd*, haplotype diversity; *Nh*, number of haplotypes. *D*, Tajima's *D*; *D**, Fu & Li's *D**; *F**, Fu & Li's *F**; *SSD*, sum of square deviations. P values of neutrality tests < 0.05 are in bold.

Locus	Mean GC content		
	GC$_I$	GC$_E$	GC3$_S$
AMP1	0.374	0.434*	0.315⁺
ATMYB33	0.308	0.519*	0.385*
CPD	0.353	0.429*	0.354*
GAPC1	0.351	0.479*	0.632*
GRP7	0.384	0.361⁺	0.364⁺
HEME2	0.357	0.404*	0.358*
LACS8	0.351	0.444*	0.319⁺
LHCA1	0.357	0.554*	0.648*
PCFS4	0.349	0.500*	0.308⁺
PMDH2	0.340	0.457*	0.429*
SPA1	0.350	0.396*	0.337⁺
SUI1	0.401	0.457*	0.574*
Average	0.356	0.453	0.419

Table 3. Mean GC contents for GC$_I$, GC$_E$, and GC3$_S$ across 94 sequences of the 12 nuclear genes. GC$_I$, the average GC content of non-coding region. GC$_E$, the average GC content of coding sites. GC3$_S$, the average GC content at third positions of codons. * and ⁺ represent, respectively, significantly higher and lower GC content based on Wilcoxon paired test ($P < 0.001$). Comparisons between GC$_E$ and GC$_I$ and between GC3$_S$ and GC$_I$ were performed.

Environmental heterogeneity and exon variation explained by environment and geography. We found no environmental heterogeneity among populations based on the eight retained environmental variables using permutational multivariate analysis of variance (PERMANOVA) ($P = 1$). However, significant environmental heterogeneity was found when compared among regions ($F = 21.52$, $R^2 = 0.6002$, $P = 0.001$). Significant environmental heterogeneity was also found in all pairwise regional comparisons except between the central-south regional group comparison ($P = 0.334$) (Supplementary Table S8).

	Variation (adjusted R^2)	F	P
Environment [a]	0.05648	1.3536	0.030
Environment + Geography [b]	0.07694	–	–
Geography [c]	− 0.01211	0.7381	0.804
[a + b + c]	0.12132	1.6351	0.003
Residuals [d]	0.87868	–	–

Table 4. The percentage of exon variation explained by non-geographically-structured environmental variables [a], shared (geographically-structured) environmental variables [b], pure geographic factors [c], and undetermined component [d] analyzed based on the eight retained environmental variables. Proportions of explained variation were obtained from variation partitioning by redundant analysis. F and P values are specified wherever applicable.

In the total aligned exon sequences of the pooled samples, 31 exon variable sites were found in nine out of the 12 nuclear genes examined. Using a variation partitioning, the total amount of explainable exon variation (12.13%) was significantly explained by a combinatorial effect of environment and geography ($F = 1.635$, $P = 0.003$), albeit large amount of exon variation was unaccountable (fraction [d]: 87.87%) (Table 4). However, essentially no exon variation was explained purely by geography independent of environment (fraction[c]: adjusted $R^2 = − 0.0121$, $F = 0.738$, $P = 0.804$). Nevertheless, exon variation was significantly explained by pure environment (fraction[a]: adjusted $R^2 = 0.0565$, $F = 1.254$, $P = 0.030$), albeit a larger portion of explainable exon variation was attributed to geographically-structured environmental differences (fraction [b]: adjusted $R^2 = 0.0769$).

Associations of environmental variables with exon variant alleles.   Seven out of the 31 exon variant alleles, including synonymous and nonsynonymous variations, were found to be strongly correlated with various combinations of environmental variables revealed either by generalized linear model (GLM) or generalized linear mixed effect model (GLMM), or by both GLM and GLMM (Table 5). The frequent to rare allele mutations in nonsynonymous substitution, including **G**CT→**T**CT in *LACS8_4*, C**G**A→C**A**A in *PCFS4_1*, C**A**C→C**G**C in *SPA1_1*, and A**A**T→A**G**T in *SPA1_6*, were found to be strongly associated either positively or negatively with various environmental variables (Table 5). Three synonymous variants, including GC**T**→GC**C** in *LACS8_3*, CT**C**→CT**G** in *SPA1_4*, and GT**A**→GT**C** in *SPA1_5*, were also found to be highly associated with environmental variables. Additionally, frequent alleles of these seven exon variants were found to be fixed in many populations across geographic regions (Fig. 2). For those exon variant alleles strongly correlated with environmental variables found by both GLM and GLMM, logistic regression plots were depicted (Fig. 3).

Discussion
Demographic history, genetic structure, and genetic diversity.
Historical and contemporary demographic events played important roles in shaping the genetic structure of natural populations of species and traces in patterns of genetic diversity can be used to reveal population demographic history[41,42]. Climatic oscillations during the Pleistocene have been widely recognized as the main historical factor shaping current population genetic structure and distributions of species or populations[42]. Limited gene flow between geographic regions resulted from bottleneck events in the Holocene approximately 9168–13,092 years ago was revealed based on EST-SSRs in *R. oldhamii*[4]. This study inferred that *R. oldhamii* has experienced a process of transition from historical connectivity toward contemporary regional isolation. In the present study, we found no consistent evidence of spatial expansion based on Tajima's D and Fu and Li's $D*$ and $F*$ using DNA sequences of 12 nuclear gene loci (Table 2). However, spatial expansion cannot be rejected because significant small R_2 values were found using the total aligned intron sequences of the pooled samples and the aligned intron sequences of the pooled samples of five of the 12 genes examined (Table 2). The discrepancy of the estimates of neutrality test statistics for different genes can be caused by a combination of factors, such as selection, demographic history, and differences in mutation rate[43,44]. Estimation of Tajima's D and Fu and Li's $D*$ and $F*$ was known to be influenced either by population reduction, population subdivision, a recent bottleneck, or migration which resulted in secondary contact among previously differentiated lineages[44–46]. Additionally, the power of statistical tests using Tajima's D and Fu and Li's $D*$ and $F*$ may be weak for small sample size, but R_2 statistic is superior for small sample size[47]. Nonetheless, a coherent pattern of spatial expansion was also suggested by the SSD statistic considering population subdivision[43,44,48] using the aligned intron sequences of the pooled samples of the 12 individual genes and the total aligned intron sequences of the pooled and the regional samples (Table 2).

Since nuclear DNA is the fastest evolving among the three genomes plants harbored[49] and nuclear intron sequences evolving faster than coding sequences[50], the nuclear intron sequences can reveal far-distant past demographic history in contrast to EST-SSRs located within protein-coding sequences. The estimation using formula $t = \tau/2\mu k$ suggests that the date of spatial expansion in *R. oldhamii* occurred during the late Pleistocene beginning approximately 68,784–119,685 years ago (Supplementary Table S6). In conjunction with the results of Hsieh et al.[4], *R. oldhamii* could have experienced spatial expansion in the late Pleistocene followed by bottleneck events occurred in the Holocene. Historical spatial expansion might have resulted in the lack of clear genetic distinction among populations and the extremely low across population differentiation ($F_{ST} = -0.0078$, $P = 0.331$), due to the retention of ancestral polymorphisms, based on nuclear DNA intron sequence data in the present study. Hence, no clear genetic structuring was also observed based on DAPC (Supplementary Fig. S1).

Exon variation	Frequent to rare allele change	Associated environmental variables	GLM		GLMM	
			Z	Estimate	Z	Estimate
LACS8_3	GC**T**→GC**C**[70] (Ala→Ala) (S)	BIO1	− 2.661	− 0.037*,**,***	− 2.662	− 0.037*,**
		BIO7	− 2.449	− 0.071*,**		
		Slope	2.37	0.089*	2.284	0.092*
		WSmean	− 2.305	− 1.427*,**	− 2.31	− 1.513*
LACS8_4	**G**CT→**T**CT[122] (Ala→Ser) (N_s)	BIO1	− 2.063	− 0.026*		
		BIO7	− 2.826	− 0.087*,**,***		
		WSmean	− 2.085	− 1.184*	− 2.047	− 1.314*
PCFS4_1	C**G**A→C**A**A[134] (Arg→Gln) (N_s)	EVI	− 2.692	− 20.846*,**,***		
		NDVI	− 1.631	− 111.91*,**,***		
		RH			− 41.34	− 0.171*,**,***
SPA1_1	C**A**C→C**G**C[98] (His→Arg) (N_s)	Aspect			− 3.116	− 0.009*,**,***
		BIO7	− 2.541	− 0.208*,**,***		
		RH	− 2.903	− 2.095*,**,***	− 2.847	− 2.095*,**,***
		Slope	2.395	0.257*,**,***	2.172	0.269*
		WSmean	− 2.131	− 3.308*,**		
SPA1_4	CT**C**→CT**G**[104] (Leu→Leu) (S)	WSmean	− 1.971	− 1.488*	− 1.97	− 1.488*
SPA1_5	GT**A**→GT**C**[113] (Val→Val) (S)	Aspect			− 3.116	− 0.009*,**,***
		BIO7	− 2.541	− 0.208*,**,***		
		RH	− 2.903	− 2.095*,**,***	− 2.847	− 2.095*,**,***
		Slope	2.395	0.257*,**,***	2.172	0.269*
		WSmean	− 2.131	− 3.308*,**		
SPA1_6	A**A**T→A**G**T[115] (Asn→Ser) (N_s)	Aspect			− 3.116	− 0.009*,**,***
		BIO7	− 2.541	− 0.208*,**,***		
		RH	− 2.903	− 2.095*,**,***	− 2.847	− 2.095*,**,***
		Slope	2.395	0.257*,**,***	2.172	0.269*
		WSmean	− 2.131	− 3.308*,**		

Table 5. Exon variable alleles strongly correlated with environmental variables based on generalized linear model (GLM) and generalized linear mixed-effects model (GLMM). *BIO1* annual mean temperature, *BIO7* temperature annual range, *EVI* enhanced vegetation index, *NDVI* normalized difference vegetation index, *RH* relative humidity, *WSmean* mean wind speed. The superscript numbers on the second column represent amino acid position of the respective protein in *Rhododendron catawbiense*. *S* synonymous substitution, N_s nonsynonymous substation. *Values do not bracket zero in 95% confidence intervals. **Values do not bracket zero in 99% confidence intervals. ***Values do not bracket zero in 99.5% confidence intervals. Exon variable sites were coded as allelic presence ("1") and absence ("0") of the rare alleles and implemented in a generalized linear model (GLM) and a generalized linear mixed effect model (GLMM) as response variables to assess the correlations of exon variant alleles with environmental variables, with binomially distributed residuals. The superscript numbers represent aligned exon sites for the nucleotide substitutions.

However, significant regional differentiation ($F_{ST} = 0.074$, $P = 0.001$) was observed which is consistent with the regional population divergence inferred based on EST-SSR and AFLP[4,5].

Higher or comparable level of nucleotide diversity ($\pi = 0.00659$) was found based on the total aligned intron sequences of the 12 nuclear genes examined (Table 2) compared to the level of nucleotide diversity of other species endemic to Taiwan, such as *Cinnamomum kanehirae* (chalcone synthase: $\pi = 0.00716$ and leafy gene: $\pi = 0.00479$)[3] and *Quercus glauca* (glyceraldehyde-3-phosphate dehydrogenase: $\pi = 0.0050$)[51]. Moreover, the level of nucleotide diversity of *R. oldhamii* was found to be higher than the level of nucleotide diversity ($\pi = 0.00134$) of a mainland *Rhododendron* species, *R. delavayi*, based on sequences of a major RNA Polymerase II subunit[52] and the nucleotide diversity ($\pi = 0.0039$) based on eight nuclear loci of a *Rhododendron* species, *R. weyrichii*, distributed in Japan and South Korea[53]. Nonetheless, the comparisons may not be appropriate because DNA sequences used in the calculation of nucleotide diversity were derived from different genes. Although there are no nucleotide diversity estimates based on DNA sequences available for comparison, the level of genetic diversity can be compared between congeneric species occurring in Taiwan based on EST-SSRs genotyped using the same set of amplification primer pairs. Comparable levels of *R. oldhamii* EST-SSR genetic diversity across populations (average $H_E = 0.284$)[4] were found when compared with other *Rhododendron* species belonging to the same subgenus *Tsutsusi* occurring in Taiwan (average $H_E = 0.293$)[54].

Forests in Taiwan were known to have a 1,500–1,600 m upward migration since the last glacial maximum[55]. In parallel with rising temperatures due to climate changes, upper altitudinal limits of mountain plants have been found to increase at a rate of ca. 3.6 m per year during the past century on the subtropical island of Taiwan and survival of plant species has been greatly affected[56]. It is probable that suitable ecological niches for
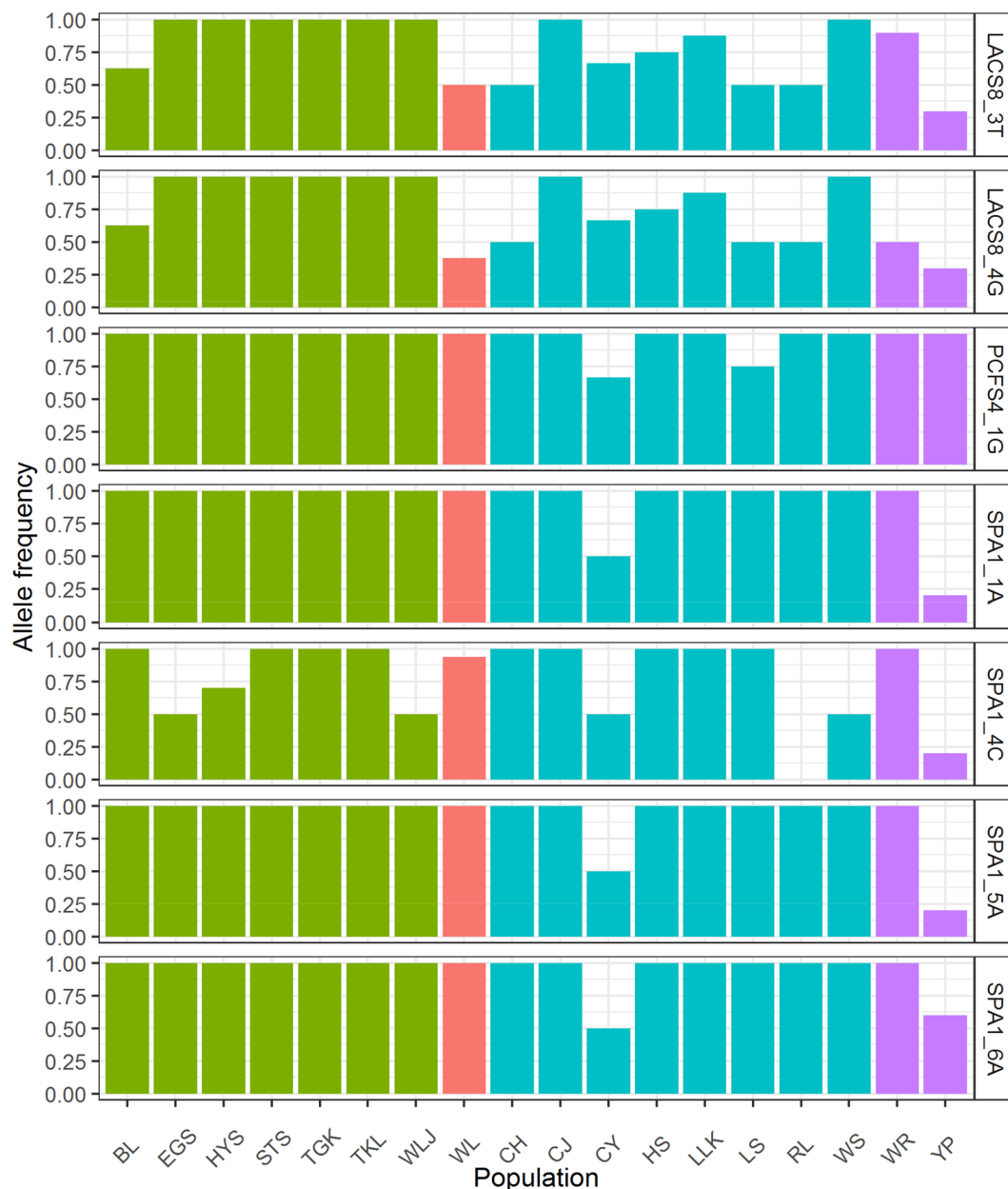
Figure 2. Distributions of frequent allele frequencies of the seven exon variants strongly associated with environmental variables across the 18 *Rhododendron oldhamii* populations.

the warmth-loving *R. oldhamii* living in the humid understory of forests could be reduced because of range retractions[57]. However, the response to forest fragmentation may differ in congeneric species adapted to different habitat types[58]. *R. oldhamii* harbors lower level of EST-SSR genetic diversity compared with other endemic species of the *R. pseudochrysanthum* complex belonging to the subgenus *Hymenanthesis* (average $H_E = 0.424$)[59]. The level of EST-SSR genetic diversity may not only reflect the outcome of a long evolutionary history, but also influenced by recent demographic events[2,3,42,60]. The lower level of EST-SSR genetic diversity in *R. oldhamii* compared with species of the *R. pseudochrysanthum* complex might have been resulted from inbreeding (Table 1) due to bottlenecks caused by habitat fragmentation in the recent past[4,61] in contrast to congeneric species of the *R. pseudochrysanthum* complex that experienced no bottlenecks[59].

Environmental variables strongly associated with synonymous and nonsynonymous variants of *LACS8* and *SPA1* nuclear genes. Elucidating the potential role of natural genetic variation in association with ecological factors has been important in evolutionary biology[62]. Environmental heterogeneity due to landscape complexity can have great influence on distribution of mountain species[63], and rugged topography and steep altitudinal environmental gradients, ranging from deep valleys to 3,000 m peaks, are common in the mountainous regions of Taiwan[64,65]. *R. oldhamii* distributed in an elevational range from 136 to 1868 m span-
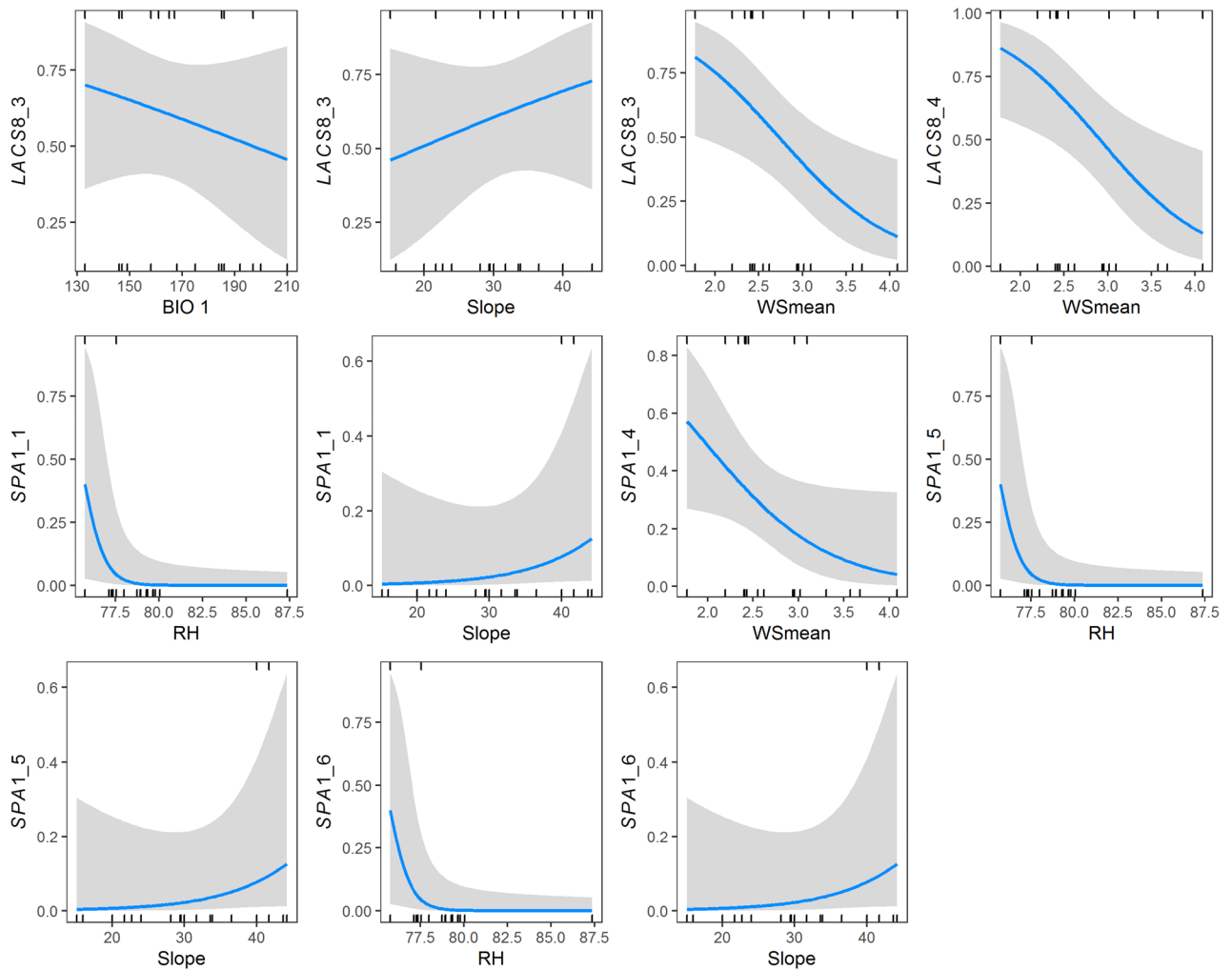
Figure 3. Logistic regression plots of the exon variants strongly correlated with environmental variables identified by both generalized linear and generalized linear mixed effect models presented in Table 5. Values of the y-axis represent the predicted probabilities of rare alleles of exon variants in *LACS8* and *SPA1* genes and numbers of the x-axis represent the values of environmental variables.

ning wide environmental gradients (Supplementary Table S9) that may have played important roles in shaping population adaptive evolution. In the present study, we used logistic regression approaches, including GLM and GLMM, to test for the most influential environmental variables strongly correlated with synonymous and non-synonymous variants in *LACS8* and *SPA1* genes (Table 5 and Fig. 3).

Our results found contrasting trends of changes in environmental variables between slope and other environmental variables associated with changes in the predicted probabilities of exon rare variant alleles of *LACS8* and *SPA1* (Fig. 3). The frequent alleles of the exon variants of these two genes were highly associated with higher values of environmental variables including BIO1, WSmean, and RH, but strongly associated with lower slope values (Table 5, Fig. 3). These results suggest that there were exon variants, particularly the frequent synonymous variants, played important roles in adapting to higher values of BIO1, WSmean, and RH, and individuals possessed these exon variations inhabiting the flatter, with smaller environmental variance, rather than the steep mountain slopes. Evidence of temperature plays an important role as an ecological driver either for adaptive genetic or epigenetic variation has been widely detected for diverse plant species distributed in different parts of the world[5,66–70]. Slope as a topological factor can act as a heat source in the day time and the surface temperature become warmer than the free atmosphere[71,72]. Warmer ambient temperature can hold moisture in the air resulting in higher relative humidity. Mean wind speed influenced by monsoon could be an important factor determining vegetation occurring in Taiwan apart from climatic factors among different altitudinal and geographic regions[73]. Our results in the present study found that environmental factors such as BIO1, BIO7, WSmean, RH, and slope strongly associated with coding sequence variation in natural populations of *R. oldhamii* suggestive of local adaptation in consistence with the findings of previous studies[4,5].

LACS8 is one member of long-chain acyl-coenzyme A synthetase gene family found in *Arabidopsis* and *LACS* mutants were found to have a damping effect on endoplasmic reticulum to plastid lipid trafficking causing lethality[74]. *SPA* encoding suppressor of phyA-105 proteins which are involved in regulating light dependent developmental processes, including photoperiodic flowering[75]. It is likely that BIO1, WSmean, and slope played

important roles in driving adaptive variation of *LACS8* gene. The protein encoded by *LACS8* was known to play an important role in signaling that governs biotic and abiotic stress responses, including temperature-induced stress that provokes changes in plasma membrane physico-chemical properties[76]. Moreover, adaptive variation in *SPA1* could be driven mainly by RH, WSmean, and slope, and might have been important to flowering response[75] of *R. oldhamii* individuals grown in different geographic regions[34,35].

Synonymous substitutions at third codon positions may have been the targets favored by selection.   Selection was demonstrated to play an important role in driving codon usage pattern[18,26,30], albeit codon bias has been attributed mostly to neutral forces, such as mutational bias and GC bias conversion[30]. In flowering plants, relationship between GC content of coding and non-coding sequences are heterogeneous among genes[77], and we found no significant positive correlations in the average GC content of coding sites and of third codon positions, respectively, with the average GC content of surrounding non-coding regions across all genes based on Spearman's rank correlation test ($\rho = -0.434$, $P = 0.161$; $\rho = 0.147$, $P = 0.651$, respectively). In addition, background compositional bias may also play an important role in the synonymous codon selection in a gene[30,78]. Although, intron length may have a positive relationship with the level of intron GC content[79], we found no significant correlation between length of intron and its GC content of the 12 nuclear genes examined (Spearman's $\rho = 0.007$, $P = 0.991$). Additionally, the average GC content of coding sites and of third codon positions were significantly different from the average GC content of surrounding non-coding regions (Table 3). Exon length may not be the influential factor causing the difference because no positive correlations in the average GC content of coding sites and of third codon positions, respectively, with exon length were found across all genes (Spearman's $\rho = 0.074$, $P = 0.820$ and Spearman's $\rho = 0.056$, $P = 0.863$, respectively). Nonetheless, selection might have played an important role in the levels of GC content at both synonymous and nonsynonymous substitution sites[13,14].

LACS8 and *SPA1* were the two genes that had lesser amounts of average GC content at third codon positions compared to that of surrounding non-coding regions (Table 3). Although the level of GC content at third codon positions was found to be positively correlated with the level of gene expression[29,30], our results found that synonymous variants at third codon positions of *LACS8* and *SPA1,* including variants of T/C (*LACS8_3*) and A/C (*SPA1_5*), possessed lesser amounts of GC content compared to surrounding non-coding regions (Table 3). The frequent T and A alleles in these two genes, respectively, were found to be closely associated with environmental variables (Table 5). Additionally, fixation of these two synonymous frequent alleles in many populations across geographic regions was found (Fig. 2). Although only partial coding sequences were examined and optimal codons may be more frequently end in G or C in dicots[14,32,33], our results suggest that optimal codons may not use G or C at third codon positions in *LACS8* and *SPA1* as expected, and they might have correlated with the optimal expression of these genes favored by selection[25,80,81].

Conclusions

Understanding the phylogeographic pattern of species and population adaptive evolution is important in evolutionary biology. In the present study, we haphazardly selected 12 nuclear loci for sequencing of natural population individuals and used in phylogeographic study and testing for adaptive evolution. The results of the present study in conjunction with the results of the previous study[4] suggest that *R. oldhamii* experienced spatial expansion in the far-distant past during the late Pleistocene followed by the recent bottlenecks in the Holocene resulting in population differentiation at regional scale. Exon variation was found to be significantly explained by environmental variables. Environmental variables that might have invoked strong selection on the seven adaptive exon variants were BIO1, WSmean, RH, and slope. Our results found causal associations of *LACS8* and *SPA1* genes, including synonymous and nonsynonymous variations, with environments in *R. oldhamii*. Our study suggests that synonymous variants, particularly those codons end in either T or A rather than G or C as expected in dicots[32,33], of nuclear genes may act as optimal codons with high frequency involved in adaptive divergence related to stress and flowering response of natural *R. oldhamii* populations located in different geographic regions.

Methods

Samples and nuclear loci.   Previous studies demonstrated that *R. oldhamii* populations can be classified geographically into four and three regional groups based on genotypic data of EST-SSR[6] and AFLP[7], respectively. The four regional groups based on EST-SSR were north, central, south, and southeast groups[4] (Table 1). The population genetic structuring of the 18 *R. oldhamii* populations analyzed based on EST-SSR and AFLP agreed with each other, except that the EST-SSR central group contains population WL (Table 1, Fig. 1), but population WL was clustered into the north group based on AFLP. We adopted the EST-SSR clustering for the present study. The number of samples collected for each geographic region ranged from 8 to 15 (north, $n = 14$; central, $n = 8$; south, $n = 15$; and southeast, $n = 10$). These samples were used for DNA extraction[82] and in direct sequencing of polymerase chain reaction (PCR)-amplified DNA products of 12 nuclear loci (Supplementary Methods and Supplementary Table S1).

PCR and sequencing.   PCR primers for the 12 genes (Supplementary Methods and Supplementary Table S10) were designed using PRIMER 3 (https://bioinfo.ut.ee/primer3-0.4.0/) based on EST sequences of *R. catawbiense*[83]. PCR amplifications were performed in a PTC-100 DNA programmable thermal cycler (MJ Research, Watertown, MA, USA) and done by initial denaturation (98 °C, 3 min), 40 cycles of denaturation (98 °C, 1 min), annealing (53.5–61.9 °C, 1 min) (Supplementary Table S10) and extension (72 °C, 1 min), and final extension (72 °C, 5 min) in a total of 40 μL PCR buffer. The PCR buffer contains 40 ng template DNA, 1X Phusion HF buffer, 1.5 mM $MgCl_2$, 0.2 mM deoxyribonucleotide triphosphate mix, 0.5 μM primer, and 2 U of

Phusion Hot Start DNA polymerase (Finnzymes Oy, Espoo, Finland). The amplification products were electrophoresed on a 1% agarose gel and the corresponding bands of the 12 genes under study were purified with Viogene Gel Extraction Kit (Viogene, Taipei, Taiwan) and directly sequenced using an ABI 3730 DNA sequencer (Applied Biosystems, Foster City, CA). Heterozygous site resolution, haplotype phasing, and functional annotation (Supplementary Table S11) were described in Supplementary Methods.

Sequence alignment, summary statistics, and neutrality tests.

Sequence alignment was performed using the *msa* function of R msa package[84] based on the ClustalW algorithm[85] in the R environment[86]. Summary statistics including the indices of the average number of pairwise nucleotide differences per site (π)[87], the average nucleotide diversity of segregating sites (θ_w)[88], and haplotype diversity (H_d) were computed using DNASP v.6[89]. Neutrality test statistics including Tajima's D[90], Fu and Li's D^* and F^*[91], and R_2[47] were also estimated using DNASP and tested for deviation from neutral expectation using 10,000 coalescent simulations. Summary statistics were estimated based on the total aligned intron sequences of the population, regional and pooled samples, and computed for the aligned intron sequences of the pooled samples for each gene separately. Neutrality test statistics were computed based on the total aligned intron sequences of the regional samples and the pooled samples and the aligned intron sequences of the pooled samples for each locus individually. Significant negative values of D, D^*, and F^* and significant small positive values of R_2 represent an excess of low frequency mutations, indicating unimodal mismatch distributions, representative of sudden expansion relative to a null model of demographic stability with multimodal mismatch distributions. The minimum number of R_m following the four-gamete test[92] and the number of segregating sites were estimated for each gene based on the aligned intron sequences of the pooled samples using DNASP. A goodness-of-fit test based on the SSD statistic was calculated using ARLEQUIN v.3.5[93] for the total aligned intron sequences of the regional and the pooled samples and the aligned intron sequences of the pooled samples of each gene separately considering population subdivision. A significant SSD value represents departure from the estimated demographic model of spatial expansion. In the goodness-of-fit test, though the error estimate is generally high, the time of spatial expansion was calculated using the formula $t = \tau/2\mu k$, where t is the time since the expansion, τ is the estimated number of generations since the expansion, μ is the mutation rate per site per generation, and k is the sequence length. We adopted a generation time of 15 years[53,94,95] and the mutation rate of 1.581×10^{-9} per site per year[53] used in the study of population demographic history of *R. weyrichii*[53], which is also belongs to the subgenus *Tsutsusi*, for calculation of the expansion time. Friedman test was used to assess the overall difference of nucleotide diversity (π and θ_w) at the population and regional levels using the *friedman* function of R agricolae package[96]. In Freidman test, nuclear locus was used as a blocking effect. Pairwise population and regional comparisons were performed and P values adjusted using Fisher's least significant difference.

The average GC content.

The GC content of 94 sequences derived from 47 individuals at coding sites, at third codon positions, and in surrounding non-coding regions of the 12 genes were calculated using CodonW (https://codonw.sourceforge.net//culong.html). Differences in the mean GC content of coding sites and of third codon positions compared with the average GC content of surrounding non-coding regions of each gene were assessed using paired Wilcoxon test (the R *wilcox.test* function of R).

Inbreeding coefficient and genetic structure.

The total aligned intron sequences of the samples of each population were used in estimating 95% confidence intervals (CIs) of F_{IS} using the *boot.ppfis* function of R hierfstat package[97] with 999 bootstrap resampling, and means and P values were calculated based on Z distribution. Across region/population F_{ST} and pairwise F_{ST} comparisons were estimated based on the aligned intron sequences using the *popStructTest* function of R package strataG[98] based on 999 permutations. Population structure was evaluated with DAPC[99] based on the total aligned intron sequences of the pooled samples using the *find.clusters* and *dapc* functions of R adegenet package[100].

Environmental variables.

Environmental variables with variance inflation factor (VIF) > 5 and highly correlated with other variables (|r| > 0.8) were removed (Supplementary Methods and Supplementary Table S12). Eight environmental variables: BIO1, BIO7, aspect, slope, EVI, NDVI, RH, and WSmean were retained as explanatory variables (Supplementary Table S9).

PERMANOVA was used to assess environmental heterogeneity, based on the eight retained environmental variables, among populations and among regions using the *adonis* function of R package vegan[101]. In PERMANOVA, environmental Euclidean distance matrix was used as response variable to test the differences among populations and among regions. Significance was determined with 999 permutations. The *pairwise.perm.manova* function of R package RVAideMemoire[102] was used in pairwise comparisons, and significance determined by 999 permutations and an FDR of 5%.

Associations of exon variant alleles with environmental variables.

Exon variable sites were coded as allelic presence ("1") and absence ("0") of the rare alleles and implemented in a GLM and a GLMM as response variables to assess the correlations of exon variant alleles with environmental variables, with binomially distributed residuals, and significance assessed with 95%, 99%, and 99.5% CIs. GLMs were performed using the R *glm* function. In GLMMs, environmental variables were used as fixed effects and geographic region as a random effect and analyzed using the *glmer* function of R lme4 package[103]. Exon variant alleles found to be significantly correlated with environmental variables detected by both GLM and GLMM were used in the visualization of the probability estimates against the associated environmental gradients using the *visreg* function of R visreg package[104].

Disentangling the effects of environment and geography explaining exon variation. The frequencies of the frequent alleles of exon variants were used in a variation partitioning analysis to disentangle the effects of environment and geography explaining exon variation. The *varpart* and *anova.cca* functions of R package vegan were used, respectively, for variation partitioning and testing for significance with 999 permutations. Exon variation was partitioned into four fractions explained by pure environmental variables (fraction [a]), geographically-structured environmental variables (fraction [b]), pure geographic variables (fraction [c]), and residual effects (fraction [d])[105,106], based on adjusted R^2 values[107]. Sample site coordinates were used as geographic variables in variation partitioning.

References

1. Small, R. L., Cronn, R. C. & Wendel, J. F. Use of nuclear genes for phylogeny reconstruction in plants. *Aust. Syst. Bot.* **17**, 145–170. https://doi.org/10.1071/SB03015 (2004).
2. Wu, S.-H. *et al.* Contrasting phylogeographical patterns of two closely related species, *Machilus thunbergii* and *Machilus kusanoi* (Lauraceae), Taiwan. *J. Biogeogr.* **33**, 936–947. https://doi.org/10.1111/j.1365-2699.2006.01431.x (2006).
3. Liao, P.-C. *et al.* Historical spatial range expansion and a very recent bottleneck of *Cinnamomum kanehirae* Hay. (Lauraceae) in Taiwan inferred from nuclear genes. *BMC Evol. Biol.* **10**, 124. https://doi.org/10.1186/1471-2148-10-124 (2010).
4. Hsieh, Y.-C. *et al.* Historical connectivity, contemporary isolation and local adaptation in a widespread but discontinuously distributed species endemic to Taiwan, *Rhododendron oldhamii* (Ericaceae). *Heredity* **111**, 147–156. https://doi.org/10.1038/hdy.2013.31 (2013).
5. Huang, C.-L. *et al.* Influences of environmental and spatial factors on genetic and epigenetic variations in *Rhododendron oldhamii* (Ericaceae). *Tree Genet. Genomes* **11**, 823. https://doi.org/10.1007/s11295-014-0823-0 (2015).
6. Li, Y.-S. *et al.* The contribution of neutral and environmentally dependent processes in driving population and lineage divergence in Taiwania (*Taiwania cryptomerioides*), . *Front. Plant Sci.* **9**, 1148. https://doi.org/10.3389/fpls.2018.01148 (2018).
7. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411–1423. https://doi.org/10.1534/genetics.110.114819 (2010).
8. Ren, J. *et al.* SNP-revealed genetic diversity in wild emmer wheat correlates with ecological factors. *BMC Evol. Biol.* **13**, 169. https://doi.org/10.1186/1471-2148-13-169 (2013).
9. Eyre-Walker, A. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21**, 569–575. https://doi.org/10.1016/j.tree.2006.06.015 (2006).
10. Wolf, J. B., Künstner, A., Nam, K., Jakobsson, M. & Ellegren, H. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biol. Evol.* **1**, 308–319. https://doi.org/10.1093/gbe/evp030 (2009).
11. Fay, J. C. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* **27**, 343–349. https://doi.org/10.1016/j.tig.2011.06.003 (2011).
12. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108. https://doi.org/10.1038/nrg1770 (2006).
13. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42. https://doi.org/10.1038/nrg2899 (2011).
14. Ingvarsson, P. K. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol. Biol. Evol.* **27**, 650–660. https://doi.org/10.1093/molbev/msp255 (2010).
15. Tiffin, P. & Hahn, M. W. Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. pekinensis. *J. Mol. Evol.* **54**, 746–753. https://doi.org/10.1007/s00239-001-0074-1 (2002).
16. Bailey, S. F., Hinz, A. & Kassen, R. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat. Commun.* **5**, 4076. https://doi.org/10.1038/ncomms5076 (2014).
17. Clément, Y. *et al.* Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet.* **13**, e1006799. https://doi.org/10.1371/journal.pgen.1006799 (2017).
18. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**, 1067–1076 (1995).
19. Subramanian, S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics* **178**, 2429–2432. https://doi.org/10.1534/genetics.107.086405 (2008).
20. McVean, G. A. & Vieira, J. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**, 245–257 (2001).
21. Haddrill, P. R., Zeng, K. & Charlesworth, B. Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. *Mol. Biol. Evol.* **28**, 1731–1743. https://doi.org/10.1093/molbev/msq354 (2011).
22. Kashiwagi, A., Sugawara, R., Tsushima, F. S., Kumagai, T. & Yomo, T. Contribution of silent mutations to thermal adaptation of RNA bacteriophage Qβ. *J. Virol.* **88**, 11459–11468. https://doi.org/10.1128/JVI.01127-14 (2014).
23. Agashe, D. *et al.* Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Mol. Bio. Evol.* **33**, 1542–1553. https://doi.org/10.1093/molbev/msw035 (2016).
24. Ingvarsson, P. K. Molecular evolution of synonymous codon usage in *Populus*. *BMC Evol. Biol.* **8**, 307. https://doi.org/10.1186/1471-2148-8-307 (2008).
25. He, B. *et al.* Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending. *Sci. Rep.* **6**, 35927. https://doi.org/10.1038/srep35927 (2016).
26. Szövényi, P. *et al.* Selfing in haploid plants and efficacy of selection: codon usage bias in the model moss Physcomitrella patens. *Genome Biol. Evol.* **9**, 1528–1546. https://doi.org/10.1093/gbe/evx098 (2017).
27. Kreitman, M. & Antezana, M. The population and evolutionary genetics of codon bias. In *Evolutionary Genetics: From Molecules to Morphology* (eds Singh, R. S. & Krimbas, C. B.) 82–101 (Cambridge University Press, Cambridge, 2000).
28. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649. https://doi.org/10.1016/S0959-437X(02)00353-2 (2002).
29. Yao, Z., Hanmei, L. & Yong, G. Analysis of characteristic of codon usage in waxy gene of *Zea mays*. *J. Maize Sci.* **16**, 16–21 (2008).
30. Camiolo, S., Melito, S. & Porceddu, A. New insights into the interplay between codon bias determinants in plants. *DNA Res.* **22**, 461–470. https://doi.org/10.1093/dnares/dsv027 (2015).
31. Kawabe, A. & Miyashita, N. T. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Syst.* **78**, 343–352. https://doi.org/10.1266/ggs.78.343 (2003).

32. Chiapello, H., Lisacek, F., Caboche, M. & Hénaut, A. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**, GC1–GC38. https://doi.org/10.1016/S0378-1119(97)00671-9 (1998).

33. Ingvarsson, P. K. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. *Mol. Biol. Evol.* **24**, 836–844. https://doi.org/10.1093/molbev/msl212 (2007).

34. Chang, Y.-M. *The Investigation of the Flowering Pattern of Rhododendron oldhamii Maxim* (Providence University, Taiwan, 2006).

35. Chi, W.-T. *The analysis of flowering rhythm and its relationship with the distribution of populations of Rhododendron oldhamii Maxim. in western Taiwan* (National Taiwan University, Taiwan, 2009).

36. Gavrilets, S. & Vose, A. Case studies and mathematical models of ecological speciation. 2. Palms on an oceanic island. *Mol. Ecol.* **16**, 2910–2921. https://doi.org/10.1111/j.1365-294X.2007.03304.x (2007).

37. Keller, S. R., Levsen, N., Ingvarsson, P. K., Olson, M. S. & Tiffin, P. Local selection across a latitudinal gradient shapes nucleotide diversity in balsam Poplar, *Populus balsamifera* L.. *Genetics* **188**, 941–952. https://doi.org/10.1534/genetics.111.128041 (2011).

38. Leimu, R. & Fischer, M. A meta-analysis of local adaptation in plants. *PLoS ONE* **3**, e4010. https://doi.org/10.1371/journal.pone.0004010 (2008).

39. Lobréaux, S. & Melodelima, C. Detection of genomic loci associated with environmental variables using generalized linear mixed models. *Genomics* **105**, 69–75. https://doi.org/10.1016/j.ygeno.2014.12.001 (2015).

40. Stucki, S. *et al.* High performance computation of landscape genomic models integrating local indices of spatial association. *Mol. Ecol. Res.* **17**(1072–1089), 2017. https://doi.org/10.1111/1755-0998.12629 (2017).

41. Dumolin-Lapègue, S., Demesure, B., Fineshi, S., Le Corre, V. & Petit, R. J. Phylogeographic structure of white oaks throughout the European continent. *Genetics* **146**, 1475–1487 (1997).

42. Avise, J. C. *Phylogeography: The History and Formation of Species* (Harvard University Press, Cambridge, 2000).

43. Schneider, S. & Excoffier, L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089 (1999).

44. Excoffier, L. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol. Ecol.* **13**, 853–864. https://doi.org/10.1046/j.1365-294X.2003.02004.x (2004).

45. Maruyama, T. & Fuerst, P. A. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* **108**, 745–763 (1984).

46. Maruyama, T. & Fuerst, P. A. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111**, 675–689 (1985).

47. Ramos-Onsins, S. E. & Rozas, J. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19**, 2092–2100. https://doi.org/10.1093/oxfordjournals.molbev.a004034 (2002).

48. Ray, N., Currat, M. & Excoffier, L. Intra-deme molecular diversity in spatially expanding populations. *Mol. Biol. Evol.* **20**, 76–86. https://doi.org/10.1093/molbev/msg009 (2003).

49. Wolfe, K. H., Li, W.-H. & Sharp, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* **84**, 9054–9058. https://doi.org/10.1073/pnas.84.24.9054 (1987).

50. Zhang, D.-X. & Hewitt, G. M. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol. Ecol.* **12**, 563–584. https://doi.org/10.1046/j.1365-294X.2003.01773.x (2003).

51. Shih, F.-L., Cheng, Y.-P., Hwang, S.-Y. & Lin, T.-P. Partial concordance between nuclear and organelle DNA in revealing the genetic divergence among *Quercus glauca* (Fagaceae) populations in Taiwan. *Int. J. Plant Sci.* **167**, 863–872. https://doi.org/10.1086/504923 (2006).

52. Sharma, A., Poudel, R. C., Li, A., Xu, J. & Guan, K. Genetic diversity of *Rhododendron delavayi* var. *delavayi* (CB Clarke) Ridley inferred from nuclear and chloroplast DNA: implications for the conservation of fragmented populations. *Plant Syst. Evol.* **300**, 1853–1866. https://doi.org/10.1007/s00606-014-1012-1 (2014).

53. Yoichi, W. *et al.* Population demographic history of a temperate shrub, *Rhododendron weyrichii* (Ericaceae), on continental islands of Japan and South Korea. *Ecol. Evol.* **6**, 8800–8810. https://doi.org/10.1002/ece3.2576 (2016).

54. Huang, C.-L. *et al.* Disentangling the effects of isolation-by-distance and isolation-by-environment on genetic differentiation among *Rhododendron* lineages in the subgenus *Tsutsusi*. *Tree Genet. Genomes* **12**, 53. https://doi.org/10.1007/s11295-016-1010-2 (2016).

55. Liew, P.-M. & Chung, N.-J. Vertical migration of forests during the last glacial period in subtropical Taiwan. *West. Pac. Earth Sci.* **1**, 405–414 (2001).

56. Jump, A. S., Huang, T.-J. & Chou, C.-H. Rapid altitudinal migration of mountain plants in Taiwan and its implications for high altitude biodiversity. *Ecography* **35**, 204–210. https://doi.org/10.1111/j.1600-0587.2011.06984.x (2012).

57. Jump, A. S., Mátyás, C. & Peñuelas, J. The altitude-for-latitude disparity in the range retractions of woody species. *Trends Ecol. Evol.* **24**, 694–701. https://doi.org/10.1016/j.tree.2009.06.007 (2009).

58. Hamrick, J. L., Murawski, D. A. & Nason, J. D. The influence of seed dispersal mechanisms on the genetic structure of tropical tree populations. *Vegetatio* **107**, 281–297. https://doi.org/10.1007/BF00052230 (1993).

59. Chen, C.-Y. *et al.* Demography of the upward-shifting temperate woody species of the *Rhododendron pseudochrysanthum* complex and ecologically relevant adaptive divergence in its trailing edge populations. *Tree Genet. Genomes* **10**, 111–126. https://doi.org/10.1007/s11295-013-0669-x (2014).

60. Hwang, S.-Y. *et al.* Postglacial population growth of *Cunninghamia konishii* (Cupressaceae) inferred from phylogeographical and mismatch analysis of chloroplast DNA variation. *Mol. Ecol.* **12**, 2689–2695. https://doi.org/10.1046/j.1365-294X.2003.01935.x (2003).

61. Reed, D. H. & Frankham, R. Correlation between fitness and genetic diversity. *Conserv. Biol.* **17**, 230–237. https://doi.org/10.1046/j.1523-1739.2003.01236.x (2003).

62. Mitchell-Olds, T. & Schmitt, J. Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* **441**, 947–952. https://doi.org/10.1038/nature04878 (2006).

63. Antonelli, A. Biogeography: drivers of bioregionalization. *Nat. Ecol. Evol.* **1**, 0114. https://doi.org/10.1038/s41559-017-0114 (2017).

64. Su, H.-J. Studies on the climate and vegetation types of the natural forests in Taiwan (II). Altitudinal vegetation zones in relation to temperature gradient. *Quart. J. Chin. Forest.* **17**, 57–73 (1984).

65. Li, C.-F. *et al.* Classification of Taiwan forest vegetation. *Appl. Veg. Sci.* **16**, 698–719. https://doi.org/10.1111/avsc.12025 (2013).

66. Manel, S., Poncet, B. N., Legendre, P., Gugerli, F. & Holderegger, R. Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Mol. Ecol.* **19**, 3824–3835. https://doi.org/10.1111/j.1365-294X.2010.04716.x (2010).

67. Manel, S. *et al.* Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Mol. Ecol.* **21**, 3729–2738. https://doi.org/10.1111/j.1365-294X.2012.05656.x (2012).

68. Fang, J.-Y. *et al.* Divergent selection and local adaptation in disjunct populations of an endangered conifer, *Keteleeria davidiana* var. *formosana* (Pinaceae). *PLoS ONE* **8**, e70162. https://doi.org/10.1371/journal.pone.0070162 (2013).

69. Chen, J.-H. *et al.* Postglacial range expansion and the role of ecological factors in driving adaptive evolution of *Musa basjoo* var. *formosana*. *Sci. Rep.* **7**, 5341. https://doi.org/10.1038/s41598-017-05256-6 (2017).

70. Shih, K.-M., Chang, C.-T., Chung, J.-D., Chiang, Y.-C. & Hwang, S.-Y. Adaptive genetic divergence despite significant isolation-by-distance in populations of Taiwan Cow-tail fir (*Keteleeria davidiana* var. *formosana*). *Front. Plant Sci.* **9**, 92. https://doi.org/10.3389/fpls.2018.00092 (2018).

71. Richner, H. & Phillips, P. D. A comparison of temperatures from mountaintops and the free atmosphere—their diurnal variation and mean difference. *Mon. Weather Rev.* **112**, 1328–1340. https://doi.org/10.1175/1520-0493(1984)112%3c1328:ACOTFM%3e2.0.CO;2 (1984).

72. Pepin, N. C. & Seidel, D. J. A global comparison of surface and free-air temperatures at high elevations. *J. Geophys. Res.* **110**, D3. https://doi.org/10.1029/2004JD005047 (2005).

73. Chiou, C.-R. *et al.* Altitudinal distribution patterns of plant species in Taiwan are mainly determined by the northeast monsoon rather than the heat retention mechanism of Massenerhebung. *Bot. Stud.* **51**, 89–97 (2010).

74. Jessen, D., Roth, C., Wiermer, M. & Fulda, M. Two activities of long-chain acyl-coenzyme A synthetase are involved in lipid trafficking between the endoplasmic reticulum and the plastid in *Arabidopsis*. *Plant Physiol.* **167**, 351–366. https://doi.org/10.1104/pp.114.250365 (2015).

75. Podolec, R. & Ulm, R. Photoreceptor-mediated regulation of the COP1/SPA E3 ubiquitin ligase. *Curr. Opin. Plant Biol.* **45**, 18–25. https://doi.org/10.1016/j.pbi.2018.04.018 (2018).

76. De Bigault Du Granrut, A. & Cacas, J.-L. How very-long-chain fatty acids could signal stressful conditions in plants?. *Front. Plant Sci.* **7**, 1490. https://doi.org/10.3389/fpls.2016.01490 (2016).

77. Glémin, S., Clément, Y., David, J. & Ressayre, A. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet.* **30**, 263–270. https://doi.org/10.1016/j.tig.2014.05.002 (2014).

78. Porceddu, A. & Camiolo, S. Spatial analyses of mono, di and trinucleotide trends in plant genes. *PLoS ONE* **6**, e22855. https://doi.org/10.1371/journal.pone.0022855 (2011).

79. Haddrill, P. R., Charlesworth, B., Halligan, D. L. & Andolfatto, P. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* **6**, R67. https://doi.org/10.1186/gb-2005-6-8-r67 (2005).

80. Miyashita, N. T., Kawabe, A., Innan, H. & Terauchi, R. Intra-and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. *Mol. Biol. Evol.* **15**, 1420–1429. https://doi.org/10.1093/oxfordjournals.molbev.a025870 (1998).

81. Morton, B. R. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J. Mol. Evol.* **46**, 449–459. https://doi.org/10.1007/PL00006325 (1998).

82. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).

83. Wei, H., Fu, Y. & Arora, R. Intron-flanking EST–PCR markers: from genetic marker development to gene structure analysis in *Rhododendron*. *Theor. Appl. Genet.* **111**, 1347–1356. https://doi.org/10.1007/s00122-005-0064-6 (2005).

84. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999. https://doi.org/10.1093/bioinformatics/btv494 (2015).

85. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680. https://doi.org/10.1093/nar/22.22.4673 (1994).

86. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2018).

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

88. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276. https://doi.org/10.1016/0040-5809(75)90020-9 (1975).

89. Rozas, J. *et al.* DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* **34**, 3299–3302. https://doi.org/10.1093/molbev/msx248 (2017).

90. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

91. Fu, Y.-X. & Li, W.-H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).

92. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).

93. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x (2010).

94. Morimoto, J., Shibata, S. & Hasegawa, S. Habitat requirement of *Rhododendron reticulum* and *R. macrosepalum* in germination and seedling stages- filed experiment for restoration of native Rhododendron by seeding. *J. Jpn. Soc. Reveg. Tech.* **29**, 135–140 (2003) ((**in Japanese**)).

95. Yasada, M. For conserving an endangered species, *Rhododendron dilatatum* var. *boreale. Kousynaikihou* **143**, 18–22 (2006) ((**in Japanese**)).

96. De Mendiburu, F. agricolae: statistical procedures for agricultural research. R package version 1.2-8. https://CRAN.R-project.org/package=agricolae (2017). Accessed September 9th 2018.

97. Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x (2005).

98. Archer, F. I., Adams, P. E. & Schneiders, B. B. strataG: an R package for manipulating, summarizing and analysing population genetic data. *Mol. Ecol. Res.* **17**, 5–11. https://doi.org/10.1111/1755-0998.12559 (2017).

99. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94. https://doi.org/10.1186/1471-2156-11-94 (2010).

100. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071. https://doi.org/10.1093/bioinformatics/btr521 (2011).

101. Oksanen, J. *et al.* vegan: community ecology package. R package version 2.4-2. https://CRAN.R-project.org/package=vegan (2017). Accessed January 15th 2018.

102. Hervé, M. RVAideMemoire: Testing and plotting procedures for biostatistics. R package version 0.9-69. https://CRAN.R-project.org/package=RVAideMemoire (2018). Accessed June 18th 2018.

103. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* **67**, 1–48. https://doi.org/10.18637/jss.v067.i01 (2015).

104. Breheny, P. & Burchett, W. Visualization of regression models using visreg. *R J.* **9**, 56–71. https://doi.org/10.32614/rj-2017-046 (2017).

105. Borcard, D., Legendre, P. & Drapeau, P. Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045–1055. https://doi.org/10.2307/1940179 (1992).

106. Borcard, D. & Legendre, P. All-scale spatial analysis of ecological data by means of principal coordinates of neighbor matrices. *Ecol. Model.* **153**, 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4 (2002).

107. Peres-Neto, P. R., Legendre, P., Dray, S. & Borcard, D. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* **87**, 2614–2625. https://doi.org/10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2 (2006).

Acknowledgements

Author contributions

S.Y.H. conceived and designed the experiments. J.D.C. and S.Y.H. collected the samples. Y.C.H. performed the experiment and haplotype analysis. C.T.C. and S.Y.H. provided analysis tools and performed the data analyses. Y.C.H. and S.Y.H. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73748-z.

Correspondence and requests for materials should be addressed to S.-Y.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.