SCIENTIFIC REPORTS

OPEN

Postglacial range expansion and the role of ecological factors in driving adaptive evolution of *Musa basjoo* var. *formosana*

Received: 18 January 2017
Accepted: 25 May 2017
Published online: 13 July 2017

Jui-Hung Chen[1], Chun-Lin Huang[2], Yu-Long Lai[1], Chung-Te Chang[3], Pei-Chun Liao[1], Shih-Ying Hwang[1] & Chih-Wen Sun[1]

Genetic variation evolves during postglacial range expansion of a species and is important for adapting to varied environmental conditions. It is crucial for the future survival of a species. We investigate the nuclear DNA sequence variation to provide evidence of postglacial range expansion of *Musa basjoo* var. *formosana*, a wild banana species, and test for adaptive evolution of amplified fragment length polymorphic (AFLP) loci underlying local adaptation in association with environmental variables. Postglacial range expansion was suggested by phylogeographical analyses based on sequence variation of the second intron of copper zinc superoxide dismutase 2 gene. Two glacial refugia were inferred by the average F_{ST} parameter (mean F_{ST} of a population against the remaining populations). Using variation partitioning by redundancy analysis, we found a significant amount of explained AFLP variation attributed to environmental and spatially-structured environmental effects. By combining genome scan methods and multiple univariate logistic regression, four AFLP loci were found to be strongly associated with environmental variables, including temperature, precipitation, soil moisture, wet days, and surface coverage activity representing vegetation greenness. These environmental variables may have played various roles as ecological drivers for adaptive evolution of *M. basjoo* var. *formosana* during range expansion after the last glacial maximum.

In the Quaternary, temperature oscillations are an important historical factor influencing the current distributions of a plant species[1]. During the last glacial maximum (LGM), most plant species in the Northern hemisphere would have retreated southward toward the tropics or warmer lowland areas, and survived in refugia[2]. Taiwan is a continental island situated off the coast of the Asian mainland and lies to the south of the Ryukyu Arc and north of the Philippines Archipelago. Although remnants of glaciations at the top of some peaks along the central mountain range (CMR) were found[3], the lowlands of Taiwan were not covered with ice but were drier and colder, and the climate changes would have confined species to refugia in low elevations[3,4]. In Taiwan, many conifers escaped to the middle elevations during the LGM[4]. These species were originally distributed at middle and low elevations and would have migrated to the lowlands. A reverse course of events occurred since the LGM[1,2,5], with species that were previously confined to refugia in the south expanded polewards and lowland forests colonizing at higher elevations[1,2,6]. Environmental gradients can potentially act as selective drivers during postglacial range expansion and result in locally adapted variants[2]. The current distributions of species are the results of a combinatorial effect by historical events, ecological factors, and stochastic or neutral mechanisms.

Population adaptive divergence is a central issue in evolutionary biology that focuses on understanding the correlations of population genetic variation with environmental heterogeneity[7]. Studies have shown that natural population divergence is driven by variable environmental conditions and leads to the evolution of locally adapted lineages[8–11]. However, gene flow between closely related genetic lineages can be reduced by a combinatorial effect

[1]Department of Life Science, National Taiwan Normal University, 88 Tingchow Road, Section 4, Taipei, 11677, Taiwan. [2]Laboratory of Molecular Phylogenetics, Department of Biology, National Museum of Natural Science, 1 Guanchien Road, Taichung, 40453, Taiwan. [3]Department of Geography, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei, 10617, Taiwan. Jui-Hung Chen and Chun-Lin Huang contributed equally to this work. Correspondence and requests for materials should be addressed to S.-Y.H. (email: hsy9347@ntnu.edu.tw) or C.-W.S. (email: cwsun@ntnu.edu.tw)

Population	Locality (latitude/longitude)	Elevation (m)	No. of sequences	Haplotypes[a]	*Hd* (SD)	θ_π (SD)	θ_S (SD)
Beishi	24°30′/121°45′	441	12	I (6), Id (2), IIc (2), IId (2)	0.727 (0.109)	0.00173 (0.00038)	0.00157 (0.00064)
Guanhu	23°40′/121°23′	763	12	I (8), Ig (2), IIb (2)	0.545 (0.144)	0.00096 (0.00030)	0.00105 (0.00052)
Sandimen	22°48′/120°41′	406	12	I (12)	0	0	0
Shanmai	23°03′/121°17′	303	10	I (4), Ic (2), Ij (2), II (2)	0.800 (0.089)	0.00141 (0.00035)	0.00140 (0.00063)
Shitou	23°42′/120°46′	937	12	I (4), Ia (2), Ib (2), Ih (2), II (2)	0.848 (0.067)	0.00144 (0.00032)	0.00157 (0.00064)
Shouka	22°14′/120°50′	324	12	I (9), II (3)	0.409 (0.133)	0.00097 (0.00032)	0.00079 (0.00045)
Wufeng	24°38′/121°06′	489	10	I (4), Ie (1), If (1), Ii (2), IIa (2)	0.822 (0.097)	0.00157 (0.00036)	0.00168 (0.00069)
Wulai	24°51′/121°34′	645	12	I (8), II (4)	0.485 (0.106)	0.00115 (0.00025)	0.00079 (0.00045)
Total			92	16	0.628 (0.056)	0.00119 (0.00014)	0.00280 (0.00066)

Table 1. Population, localities, and descriptive statistics of *Musa basjoo* var. *formosana* based on the second intron sequences of Cu/Zn SOD2. *Hd*, haplotype diversity; θ_π, nucleotide diversity estimated based on the average pairwise number of differences between sequences, θ_S, nucleotide diversity estimated based on the number of segregating sites per sequence. [a]The numbers in parentheses indicate the frequency of haplotype.

of geography and environment; and geographical isolation may play a larger role than environmental variation in shaping population structure[11–13]. Therefore, investigating the relative roles of geography and environment that influence genetic variation is critical to understand how environmental factors may act as selective drivers and lead to adaptive genetic variation underlying local adaptation in natural populations of a species[11,12]. Moreover, identifying environmental factors that play roles in driving adaptive divergence is particularly of interest in biological conservation and ecological restoration[14].

In Taiwan, genetic signatures of postglacial expansion into ranges of habitats from refugia were observed in many tree species such as *Cyclobalanopsis glauca*[15], *Cunninghamia konishii*[16], *Trochodendron aralioides*[17], *Castanopsis carlesii*[18], and *Cinnamomum kanehirae*[6,19]. Although Taiwan spans a small range of latitude geographically, varied geographical topologies support vegetation from tropical to cool climates[20]. The dramatic topological differences combined with the influence of tropical and subtropical climates have fostered high habitat diversity and may serve as a driving force for adaptive evolution[9–12]. During the postglacial recolonization process, a species would have evolved with adaptive variation invoked by differential selection along environmental gradients occurring in species' distribution ranges.

Musa basjoo Siebold is a cold hardy banana species and is thought to be originated from southern China[21,22] and is genetically differentiated from other *Musa* species[22,23]. *M. basjoo* Siebold & Zucc. ex linuma var. *formosana* (Warb. ex Schum.) Ying is a variety of *M. basjoo* endemic to Taiwan[24]. *M. basjoo* var. *formosana* is distributed in low elevations in Taiwan[25] and has a diploid chromosome number of $2n = 22$[26]. The generation time of bananas under cultivation can be up to 18 months[25] and could be longer for natural populations of *M. basjoo* var. *formosana*. Fruit from *M. basjoo* var. *formosana* is edible but has numerous large and hard seeds. This species is an important germplasm for banana breeding due to its characteristics of cold tolerance and disease resistance. For conservation of this species, it is important to identify the genetic relationships of individuals among populations and to understand the potential for evolutionary adaptation because biodiversity is increasingly threatened by human-induced anthropogenic climate change[27].

In this study, the second intron of copper zinc superoxide dismutase 2 gene (Cu/Zn *SOD2*) was sequenced for 46 individuals from eight populations (Table 1 and Fig. 1) to characterize the postglacial recolonization event. Moreover, amplified fragment length polymorphism (AFLP)[28] was also used to survey genetic variation of 112 individuals to examine the genetic relationships of individuals from different populations and test for association of AFLP loci with environmental variables underlying local adaptation driven by environmental gradients. We used variation partitioning based on redundancy analysis (RDA)[29] to assess the relative influences of geographical and ecological isolation contributes to AFLP variation. F_{ST} outliers were identified using genome scan methods. In light of understanding local adaptations associated with environmental gradients, multiple univariate logistic regression was used to correlate environmental variables with AFLP loci that have potentially evolved under selection. We hypothesized that frequencies of AFLP outliers may display correlations with environmental gradients underlying local adaptation because the distribution of *M. basjoo* var. *formosana* in different geographical regions represents habitat environmental heterogeneities across the CMR. The main goals of this study were to: (i) demonstrate postglacial expansion of *M. basjoo* var. *formosana*; and (ii) find evidence for potential adaptive evolution associated with environmental gradients.

Results

Nucleotide diversity, population differentiation, and test for postglacial expansion based on the second intron sequences of Cu/Zn *SOD2*. We obtained 92 sequences of the second intron of Cu/Zn *SOD2* with an aligned length of 1264 bp. No sign of recombination was found examined within populations. However, one non-significant recombination event ($P = 0.876$) was found between nucleotide sites 15 and 510 when total samples were analyzed. In addition, 2 of 46 samples collected from populations Shouka and Wufeng

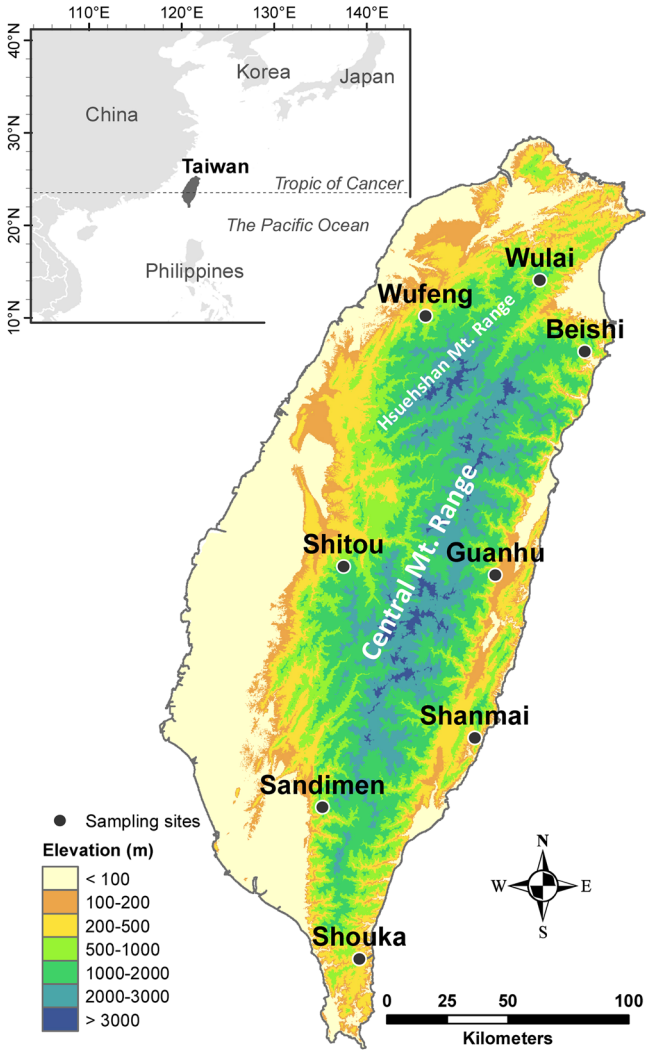

Figure 1. Sample localities of eight populations of *Musa basjoo* var. *formosana*. The countries' boundary (polygon) map was derived from the default map database in ArcGIS v. 10.3[92]. The elevation gradients of Taiwan (background) were presented in ArcGIS based on the 20 m digital elevation model (DEM) that was acquired from Data.GOV.TW[93]. The locations of the sampling sites were plotted using Tools in ArcGIS by their coordinates.

were heterozygotes and one single-base pair indel was found in the aligned sequences. Sixteen haplotypes were identified in the 92 sequences analyzed (Table 1 and Fig. 2). The number of haplotypes ranged from 1 to 5 for each population. The most common haplotype (haplotype I) was found in all populations examined and the second most common haplotype (haplotype II) was found in four populations (populations Shanmai, Shitou, Shouka, and Wulai) (Table 1 and Fig. 2). Populations Shitou and Wufeng had the greatest number of haplotypes and the population Sandimen had only one haplotype (the most common haplotype I). Although mismatch distribution of the frequency of pairwise differences among haplotypes did not fit tightly with a population expansion model (Kolmogorov-Smirnov test, $P < 0.001$, Supplementary Fig. S1), the haplotype network displayed "star-like" phylogeny in two separate groups indicating population expansion and subdivision (Fig. 2).

Population pairwise F_{ST} values calculated from Cu/Zn *SOD2* second intron sequence variation was mostly low (Supplementary Table S1) and averaged 0.053. The southern population Sandimen showed the highest mean pairwise F_{ST} (=0.154) against the remaining populations (Fig. 3), which may have been the southern glacial refugium for *M. basjoo* var. *formosana*. Table 1 presents the descriptive statistics of nucleotide diversity of the second intron of Cu/Zn *SOD2* include haplotype diversity (*Hd*) and nucleotide diversity θ_S and θ_π. *Hd* ranged between 0 (population Sandimen) and 0.848 (population Shitou), θ_π ranged between 0 (population Sandimen) and 0.00173 (population Beishi), and θ_S ranged between 0 (population Sandimen) and 0.00168 (population Wufeng). The levels of *Hd*, θ_π, and θ_S were not correlated with the number of sequences (sample size) via Pearson's correlation test (*Hd*, $r = -0.497$, $P = 0.209$; θ_π, $r = -0.383$, $P = 0.349$; and θ_S, $r = -0.469$, $P = 0.240$).

Neutrality tests based on Tajima's D[30] and Fu's Fs[31] revealed non-significant negative values in many populations and even positive values were observed (Table 2). In contrast, for pooled samples, significant negative values

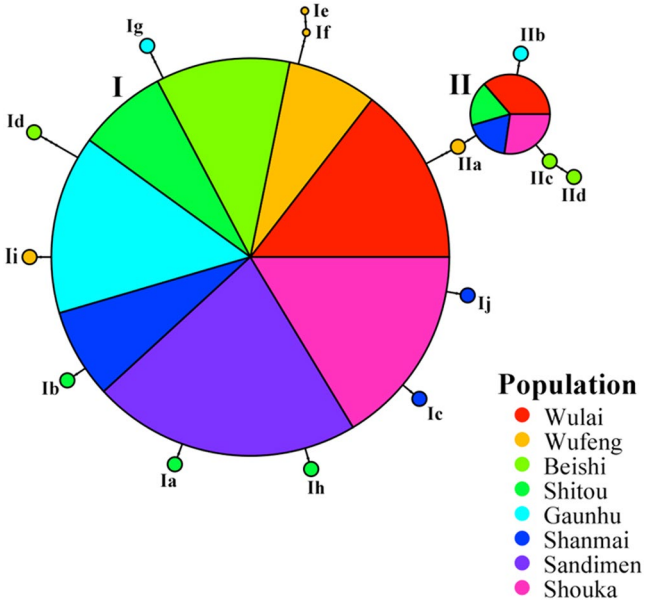

Figure 2. Haplotype network for *Musa basjoo* var. *formosana* based on nuclear superoxide dismutase gene intron 2 sequence data. The size of a circle corresponds to the haplotype frequency. Each line between haplotypes represents a mutational step; the dot represents another mutational step between haplotypes.

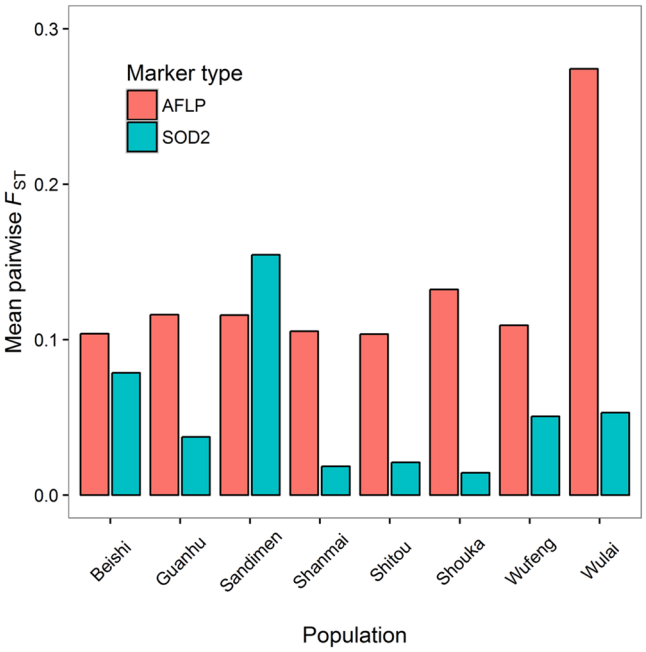

Figure 3. Degree of Mean pairwise F_{ST} values of each population in comparison with those of the remaining populations based on the second intron sequences of *Cu/Zn SOD2* and amplified fragment length polymorphic variation.

in Tajima's D (=−1.6544, $P=0.015$) and Fu's Fs (=−6.8767, $P=0.008$) were found. Moreover, non-significant raggedness index (rg)[32] was found for all individual populations and the pooled samples ($rg=0.0697$, $P=0.84$) (Table 2). Significant values of the sum of square deviations (SSD) statistic were found for several populations (populations Guanhu, Shitou, Shouka, and Wulai) and R_2 index[33] was non-significant for all individual populations (Table 2). Nevertheless, non-significant SSD (=01976, $P=0.19$) and significant R_2 index (=0.0416, $P=0.04$) were found for pooled samples, indicating spatial range expansions[34]. The mean demographic expansion factor (τ)[35], representing the time since the beginning of an expansion, was 1.6167 (0.30225–8.57544, 95% confidence intervals (CIs)). The time at which the expansion event took place was dated with a mean of 21317 (3985–113072,

	Beishi	Guanhu	Shandimen	Shanmai	Shitou	Shouka	Wufeng	Wulai	Total
Cu/Zn *SOD2* second intron									
Tajima's *D* (*P* value)	0.3743 (0.695)	−0.2986 (0.422)	NA	0.0235 (0.547)	−0.3237 (0.406)	0.7722 (0.779)	−0.2792 (0.397)	1.5227 (0.943)	−1.6544* (0.015)
Fu's *Fs* (*P* value)	1.1366 (0.742)	1.5988 (0.827)	NA	0.7370 (0.675)	−0.0792 (0.459)	3.6988 (0.957)	−0.2519 (0.411)	4.2310 (0.967)	−6.8767** (0.008)
rg (*P* value)	0.2323 (0.12)	0.3315 (0.90)	NA	0.1664 (0.37)	0.1625 (0.44)	0.6839 (0.96)	0.1037 (0.58)	0.7355 (0.94)	0.0697 (0.84)
SSD (*P* value)	0.0785 (0.11)	0.4187** (0.00)	NA	0.0652 (0.27)	0.1819* (0.02)	0.3347** (0.00)	0.0313 (0.54)	0.4702** (0.00)	0.1976 (0.11)
R_2 (*P* value)	0.1818 (0.63)	0.1515 (0.19)	NA	0.1778 (0.44)	0.1515 (0.30)	0.2045 (0.67)	0.1564 (0.22)	0.2424 (0.87)	0.0416* (0.04)
AFLP									
N	16	15	17	8	15	19	13	18	15.1
%P	65.6	52.2	56.8	75.4	64.1	56.0	73.3	50.7	63.1
N_p (N_{fp})	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0)	14 (0)	2 (0)
HE	0.280	0.241	0.259	0.324	0.290	0.231	0.336	0.248	0.276
SE (HE)	0.0085	0.0082	0.0083	0.0082	0.0079	0.0078	0.0071	0.0072	0.0079

Table 2. Neutrality test statistics and mismatch analysis based on the second intron sequences of *Cu/Zn SOD2* and summary of genetic diversity based on 521 amplified fragment length polymorphic loci of eight populations of *Musa basjoo* var. *formosana*. $^{*}P < 0.05$, $^{**}P < 0.01$ *N*, number of samples; *%P*, percent of polymorphic loci at the 5% level; H_E, Nei's gene diversity.

95% CIs) years before present for *M. basjoo* var. *formosana*, which corresponded roughly to the time frame at the end of the LGM, from 25,000–18,000 years before the present[2,5,36,37].

AFLP diversity and differentiation. Twelve primer pairs generated a total of 521 AFLP loci in the entire sample with an overall repeatability of 94.68% (Supplementary Table S2). The proportion of AFLP polymorphic loci ranged from 50.7% (population Wulai) to 75.4% (population Shanmai) with an average value of 63.1% (Table 2). The level of Nei's genetic diversity (H_E)[38] averaged 0.276 and ranged from 0.231 (population Shouka) to 0.336 (population Wufeng) (Table 2). The mean H_E was 0.276.

In AFLP data, the northern population Wulai had the largest average pairwise F_{ST} against the remaining populations (mean $F_{ST} = 0.274$, Fig. 2 and Supplementary Table S1). In the HICKORY analysis[39], the inbreeding coefficient (f) = 0 model, which estimated θ^B (an F_{ST} analogue) with $f = 0$ best fitted the data by having the lowest deviance information criterion (DIC) and $\overline{D}$ (a measure of how well the model fits the data) values, but there was little difference in the DIC values between the $f = 0$ model and the next best model (full model) (13771.7 vs. 13772.4) (Supplementary Table S3). The full model estimated an f of 0.0789 (0.0091–0.1728, 95% CIs). These results suggested that the degree of inbreeding was low in *M. basjoo* var. *formosana*. The *M. basjoo* var. *formosana* populations were moderately structured according to HICKORY estimates based on the $f = 0$ and full model ($\theta^B = 0.1354$ and 0.1429, respectively). The moderate level of genetic differentiation between populations estimated with HICKORY was consistent with the values estimated with AFLP-SURV[40] (average pairwise $F_{ST} = 0.1284$, Supplementary Table S1) and hierarchical analysis of molecular variance (AMOVA) ($\Phi_{ST} = 0.2117$, $P < 0.001$, Supplementary Table S4). When the population Wulai was excluded (because it was highly differentiated from other populations based on DAPC analysis; see result in the following section, Fig. 4), AMOVA revealed a Φ_{ST} value of 0.1308 ($P < 0.001$). AMOVA Φ_{ST} value based on the three DAPC clusters was 0.2091 ($P < 0.001$).

In the STRUCTURE analysis[41], the maximal ΔK value (change in the log probability) occur at $K = 2$ (Supplementary Fig. S2). However, the highest mean log likelihood (LnP(D)) was obtained when $K = 8$. In the LEA analysis[42], the minimal cross-entropy was lowest when $K = 7$ (Supplementary Fig. S3). However, analysis based on discriminant analysis of principal components (DAPC)[43] revealed three genetic clusters with the first two linear discriminants described 74.48% of the total AFLP genetic variation (Fig. 3). DAPC results provided prominent phylogenetic breaks among the clusters. In addition, the values of symmetric similarity coefficient[44] (SSC) were higher when $K = 2$, 3, and 4 (0.995, 0.986, and 0.983, respectively) and lower when $K \geq 5$ (Supplementary Fig. S4). Therefore, $K = 3$ could be the most probable genetic clustering scenario according to the results of DAPC, STRUCURE, and LEA analyses (Fig. 3 and Supplementary Fig. S5).

Effect of environmental variables on AFLP variation among populations. Seven environmental variables, including annual mean temperature (BIO1), annual precipitation (BIO12), mean wind speed (WS_{mean}), normalized difference vegetation index (NDVI), soil pH, soil moisture index (TMI)[45], and wet days (Supplementary Table S5) were retained as explanatory variables for AFLP variation. Variation partitioning[29] revealed a vast amount of unexplained variation (80.94%, fraction [d]), and the proportion of explained variation was 19.06% (fraction [a+b+c]) (Supplementary Table S6). Within the explained variation, 10.81% ($P < 0.001$) explained by pure environmental variables (fraction [a]) and 8.25% ($P < 0.001$) by geographically structured environmental variables (fraction [b]). No AFLP variation attributed to pure geographical difference (fraction [c]) was found.

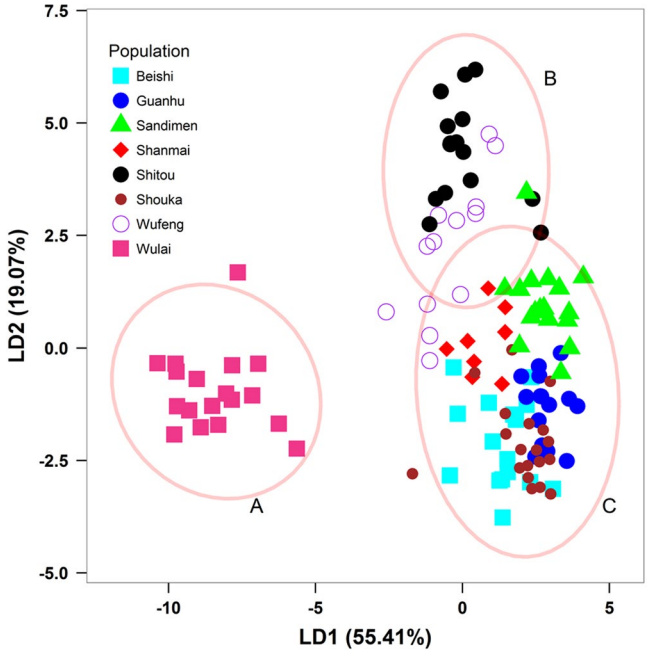

Figure 4. Scatter plot of the first two linear discriminants (LDs) from the discriminant analysis of principal components based on 521 amplified fragment length polymorphic loci for individuals of eight populations of *Musa basjoo var. formosana*. The ellipses represent 95% confidence intervals distinguishing the three genetic clusters A, B, and C.

AFLP loci potentially evolved under selection and test for association with environmental variables. We performed outlier detection for AFLP loci potentially evolved under selection with two neutrality test methods: DFDST and BAYESCAN[46,47]. Multiple univariate logistic regression was used to test for correlations of frequencies of AFLP loci with values of environmental variables using Samβada[48]. DFDSIT and BAYESCAN identified six (1.15%) and five (0.96%) AFLP loci that potentially evolved under selection (Table 3). Samβada analysis found seven AFLP loci strongly correlated with environmental variables (Table 3). We considered four AFLP loci (P1_17, P3_23, P3_24, and P6_12) as adaptive loci potentially evolved under selection because they were identified by either neutrality test method and correlated strongly with environmental variables. Four AFLP loci that identified as outliers by DFDIST or BAYESCAN and correlated strongly with environmental variables, logistic regression plots were drawn (Fig. 5). In the Samβada analysis, significant positive relationships were found for locus P1_17 with soil moisture and wet days (pseudo-$R^2 = 0.251$, $P < 0.0001$; pseudo-$R^2 = 0.106$, $P = 0.0025$, respectively) and locus P3_24 with annual mean temperature and NDVI (pseudo-$R^2 = 0.094$, $P = 0.00897$; pseudo-$R^2 = 0.105$, $P = 0.00897$, respectively). Significant negative relationships were found between locus P3_23 and NDVI (pseudo-$R^2 = 0.115$, $P = 0.0056$) and between locus P6_12 and annual mean temperature (pseudo-$R^2 = 0.101$, $P = 0.00368$). Samβada analysis reported several types of pseudo-R^2 values, we adopted the Nagelkerke pseudo-R^2 because its calculation is based on log likelihoods rather than on residual deviance and scaled approximately from 0 to 1 equivalent to the unadjusted R^2 in linear regression[49].

Discussion

In this study, we surveyed sequence variation in the second intron of Cu/Zn *SOD2* and AFLP variation to investigate whether the postglacial range expansion occurred in *M. basjoo* var. *formosana* and adaptive evolution in association with the environmental gradients of contemporary populations in this species. Sequence variation in the second intron of Cu/Zn *SOD2* suggested range expansion since the LGM. A vast amount of unaccounted variation was observed, which is typical for RDA multivariate analysis[50]. In addition, a considerable amount of unexplained variation could also be attributed to random processes triggered by ecological drift and dispersal, non-spatially structured biological, and/or unmeasured environmental differences[11,51]. In addition, the natural landscapes of Taiwan have been dramatically altered by humans, particularly in the lowlands, which may have played a role in influencing genetic variation surveyed in *M. basjoo* var. *formosana*. Nevertheless, a significant amount of AFLP variation can be explained by the environmental variables. Local adaptation in *M. basjoo* var. *formosana* populations was suggested by the AFLP loci that were potentially evolved under selection in association with environmental variables, including annual mean temperature, annual precipitation, surface coverage activity (NDVI), soil moisture, and number of days with >0.1 mm of rain per month (wet days).

Glaciers have had a serious impact on the current distribution of plant species and most subtropical species have retreated toward the tropics or warmer lowland areas[2]. The discrepancy in detecting population expansion for individual populations could be because of either population reduction, population subdivision, a recent bottleneck, or migration[52]. Our data revealed that the mismatch distribution for the frequency of pairwise differences

Locus	DFDIST[a]	BAYESCAN[b] (log_{10} (PO))	Samβada[c] (pseudo-R^2, P-value)
P1_17	X		TMI (0.251, <0.0001); Wet day (0.106, 0.0025)
P1_19			BIO1 (0.00363)
P2_15			TMI (0.266, 0.0021), Wet days (0.326, 0.0024)
P3_23	X	X (4.699)	NDVI (0.115, 0.0056)
P3_24	X	X (4.398)	BIO12 (0.094, 0.0086), NDVI (0.105, 0.0090)
P4_35			Soil pH (0.241, 0.0053)
P6_12	X	X (1.558)	BIO1 (0.101, 0.0037)
P6_17	X	X (1000)	
P6_25		X (1.889)	
P6_30			WSmean (0.362, 0.0064)
P12_29			BIO1 (0.135, 0.0049)
P12_49	X		

Table 3. Amplified fragment length polymorphic (AFLP) loci identified potentially evolved under selection by DFDIST and BAYESCAN neutrality test methods and association of allelic frequencies with values of environmental variables. [a]For DFDIST, 95% significance level and 5% false discover rates were used. [b]For BAYESCAN, a log_{10} (probability odds) (log_{10}PO) above 1.5 indicated a very strong outlier with a 5% false discovery rate. [c]For Samβada, values of Nagelkerke's pseudo-R^b are indicated (Nagelkerke 1991). Statistically significant at $P < 0.01$, corresponding to a confidence threshold after Bonferroni corrections of 2.741 x −6. BIO1, annual mean temperature; BIO12, annual precipitation; NDVI, normalized difference vegetation index; Wet days, number of days with >0.1 mm of rain per month; WS_{mean}, mean wind speed; TMI, soil moisture index.

among haplotypes did not fit tightly with a population in the expansion model (Kolmogorov-Smirnov test, $P < 0.001$, Supplementary Fig. S1), which indicates diminishing populations or structured sizes of *M. basjoo* var. *formosana*[53,54]. In un-subdivided populations, molecular signature characteristics of sudden expansions might not be observed, even when populations had expanded by several orders of magnitude after the LGM[55]. Although no consistent evidence for population growth was found for individual populations, our results showed evidence for historical spatial range expansion in pooled samples because of significant negative values of neutrality test statistics (Tajima's $D = -1.6544$, $P = 0.015$; Fu's $Fs = -6.8767$, $P = 0.008$) (Table 2). Moreover, the hypothesis of historical spatial range expansion could not be rejected in pooled samples based on non-significant small *rg* value ($rg = 0.0697$, $P = 0.84$)[32], non-significant SSD (SSD = 00096, $P = 0.81$)[56], and significant R_2 ($R_2 = 0.0416$, $P = 0.04$)[33] (Table 2).

Based on the average F_{ST} value for each population in comparison with the remaining populations, glacial refugia of plant species have been inferred in many plant species in Taiwan. The most divergent populations can be found located in the north-central and south (particularly in the southeastern part) parts of Taiwan for tree species[6]. Populations served as refugia in the north-central part of Taiwan were found for tree species, including *T. aralioides*[17], *Cu. koinishii*[57], *Ca. carlesii*[18], *Machilus thunbergii*[58], *Ma. Kusanoi*[58], and *Ci. kanehirae*[6] based on chloroplast DNA variation; and *Ci. kanehirae*[19] and *Quercus glauca*[59] based on nuclear DNA variation. Species with southern glacial refugia include *Cy. glauca*[15], *T. aralioides*[17], *Ca. Carlesii*[18], *Ma. thunbergii*[58], *Ma. Kusanoi*[58], and *Ci. kanehirae*[6] based on chloroplast DNA variation; and *Ci. kanehirae*[19] and *Q. glauca*[59] based on nuclear DNA variation. Our results in *M. basjoo* var. *formosana* reflect recolonization events after the latest glacial period that probably originated from the southern population Sandimen due to its highest mean pairwise F_{ST} (=0.154) against the remaining populations based on sequence variation of the second intron of Cu/Zn *SOD2* (Fig. 3). However, the population Wulai had the highest average F_{ST} based on AFLP variation (Fig. 3) and clearly distinguished from other populations according to the DAPC genetic clustering analysis (Fig. 4), suggesting that the population Wulai may have been the northern glacial refugium for *M. basjoo* var. *formosana*.

In comparison with other angiosperm species that occurred in Taiwan, *M. basjoo* var. *formosana* had lower nucleotide diversity in the Cu/Zn *SOD2* second intron ($Hd = 0.628$, $\theta_\pi = 0.00119$, and $\theta_S = 0.00280$) compared to introns of chalcone synthase (*Chs*) and leafy (*Lfy*) genes in *Ci. kanehirae* (*Chs*: $Hd = 0.841$, $\theta_\pi = 0.00716$, and $\theta_S = 0.00371$; *Lfy*: $Hd = 0.895$, $\theta_\pi = 0.00479$, and $\theta_S = 0.00805$)[19] and intron of glyceraldehyde-3-phosphate dehydrogenase gene in *Q. glacuca* ($Hd = 0.840$ and $\theta_\pi = 0.00500$)[59]. However, these comparisons may not be appropriate based on only one gene and the gene sequences compared between species were different. Nevertheless, the level of sequence nucleotide diversity reflects the past demographic events in natural populations. Although populations Sandimen and Wulai may have been the glacial refugia for *M. basjoo* var. *formosana*, populations Beishi, Shitou, and Wufeng located in north and central Taiwan across the CMR had higher levels of nucleotide diversity (Table 1). These results suggested that populations Beishi, Shitou, and Wufeng may be the melting pots of diversity[5]. However, the migration routes may have occurred mostly northward from the southern refugium (population Sandimen) and between adjacent populations except for the population Wulai; and scarcely southward from the northern refugium (population Wulai). This is because the population Wulai had a high average F_{ST} (=0.274) in comparison with all other populations based on AFLP variation (Supplementary Table S1) and individuals of the population Wulai are clearly differentiated from individuals of all other populations (Fig. 4). Compared to other populations, the population Wulai is identified as potential refugium located in higher elevation because the rugged geographic topology might serve as an insular area for the survival of this population during the LGM.

AFLP displayed higher level of genetic diversity (average $H_E = 0.276$) in *M. basjoo* var. *formosana* compared with other broadleaf tree species occurred in Taiwan, such as *Rhododendron oldhamii* (average $H_E = 0.216$)[11] and

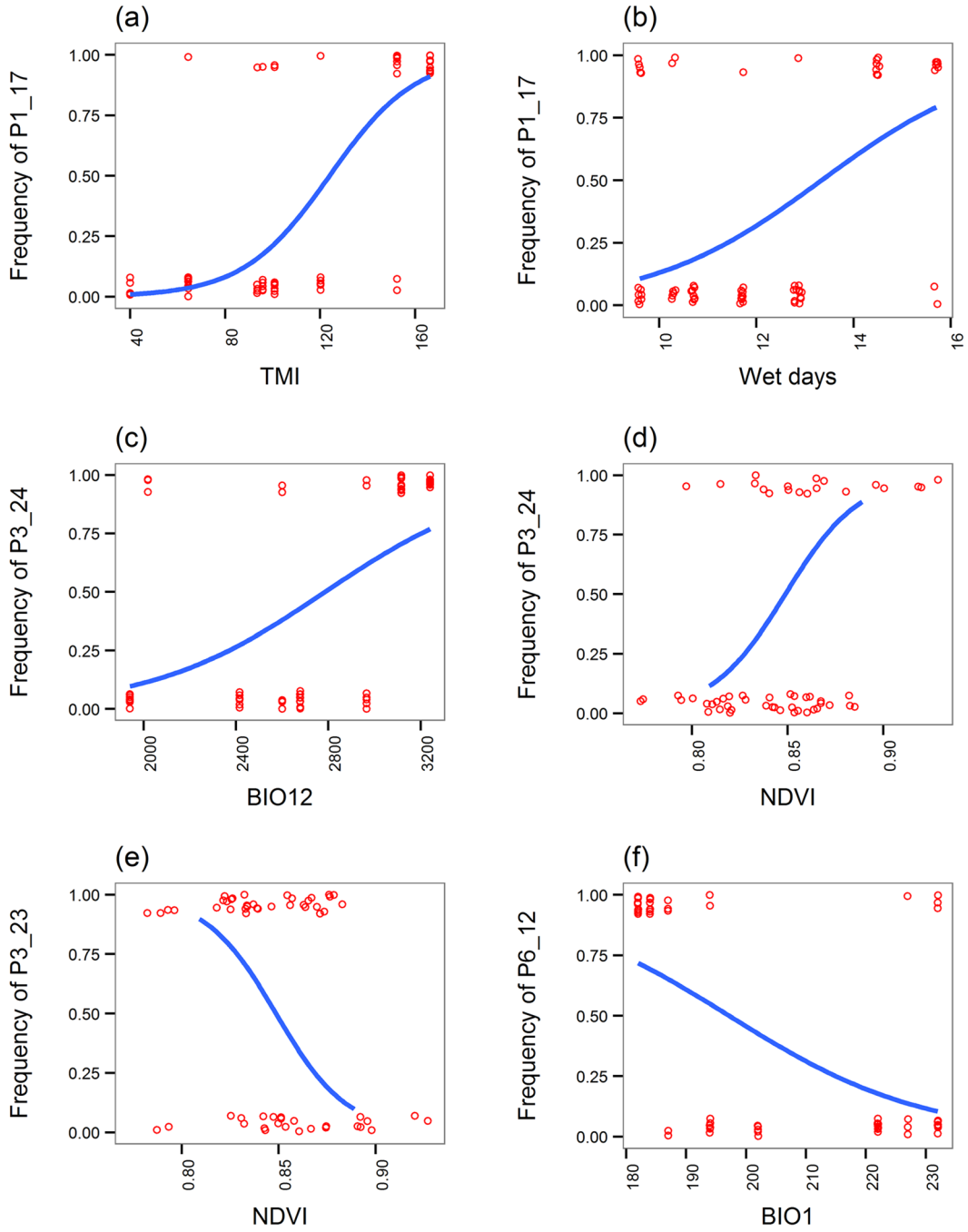

Figure 5. Logistic regression plots of four AFLP loci (P1_17, P3_23 P3_24, and P6_12) potentially evolved under selection against values of specific environmental variables. Logistic regression depicts significant positive relationships of locus P1_17 with environmental variables TMI (**a**) and wet days (**b**) (pseudo-$R^2 = 0.251$, $P < 0.0001$; pseudo-$R^2 = 0.106$, $P = 0.0025$, respectively) and of locus P3_24 with environmental variables BIO12 (**c**) and NDVI (**d**) (pseudo-$R^2 = 0.094$, $P = 0.00897$; pseudo-$R^2 = 0.105$, $P = 0.00897$, respectively), and significant negative relationship of locus P3_23 with environmental variable NDVI (**e**) (pseudo-$R^2 = 0.115$, $P = 0.0056$) and of locus P6_12 with environmental variable BIO1 (**f**) (pseudo-$R^2 = 0.101$, $P = 0.00368$). BIO1, annual mean temperature; BIO12, annual precipitation, NDVI, normalized difference vegetation index; TMI, Thornthwaite moisture index; wet days, number of days with > 0.1 mm of rain per month.

species in the genus *Salix* (average $H_E = 0.166$)[12]. *M. basjoo* var. *formosana* also had a relatively high level of AFLP variation compared to the average H_E (=0.230) for 13 plant species summarized in Nybom[60]. Moreover, the level of AFLP variation of *M. basjoo* var. *formosana* was comparable to that of *M. balbisiana*, another wild banana species, occurred in China (average $H_E = 0.241$)[61]. Patterns of genetic variation in contemporary populations of a species are influenced by the historical processes that shaped the distribution of a species[1,2], by the landscape ecological properties[11,62], and by life history traits[63]. High levels of H_E in *M. basjoo* var. *formosana* may have been related

to the low degree of inbreeding revealed by the HICKORY analysis (Supplementary Table S3). Long life span and predominant outcrossing by animal pollinators may account for high AFLP diversity in wild bananas[61,64]. The potential build-up of genetic variation during the course of expansion since the LGM, particularly under climate change conditions, is important for species' survival facing global climate change[27]. *M. basjoo* var. *formosana* harbors a substantial amount of AFLP variation and is an important resource for populations to adapt to changing environmental conditions under natural selection.

In *M. basjoo* var. *formosana*, pure environmental and spatially structured environmental factors explained a significant amount of AFLP variation (Supplementary Table S6), which suggests that environments play important roles in influencing genetic variation of this species. Environmental variables such as temperature, precipitation, surface coverage activity, soil moisture, and wet days are the most important ecological drivers influencing genetic variation of *M. basjoo* var. *formosana* and correlated strongly with four AFLP loci based on multiple univariate logistic regression analysis (Table 3 and Fig. 5), indicating fitness-related change in AFLP variation. Temperature and precipitation are commonly found to be the ecological drivers strongly correlated with adaptive AFLP variation for many plant species that occur in Taiwan, such as *R. oldhamii*[11], *Keteleeria davidiana* var. *formosana*[9], and *Salix* species[12]. Temperature and precipitation appeared to be common ecological drivers for adaptive AFLP variation in natural populations of various plant species[8,65–68].

Soil properties can explain a non-negligible proportion of the spatial distribution of tree species[69] and influence genetic variation among populations of tree species[70–72]. Soil moisture is associated with allozyme genotypes at the glycerate dehydrogenase locus and may play an important role in the adaptation of *Pinus edulis*[70]; and with AFLP loci such as in *Fagus sylvatica*[71] and *Eperua falcate*[72]. NDVI is a measure of surface coverage activity indicating the level of vegetation greenness and acts as a proxy for a biotic competitive environment. NDVI, estimated with the moderate resolution imaging spectroradiometer, has been shown to be linear with the fraction of absorbed photosynthetically active radiation (fPAR)[73], which was not retained as explanatory variable in this study (Supplementary Methods). NDVI and/or fPAR can be influential factors acting on the genetic variation of a species in response to interactions with other species in a local ecological community[74]; and has been shown to be correlated with population adaptive divergence[12] and adaptive divergence between species[12,13].

Understanding environmental variables acting as ecological drivers and playing roles in shaping the contemporary gene pool structure of *M. basjoo* var. *formosana* from experienced postglacial range expansion is important. Our results suggest that there are two glacial refugia located in the northern and southern part of the *M. basjoo* var. *formosana* distribution range. Postglacial expansion confronting ecological discontinuity may have triggered the evolution of environmentally associated AFLP variation underlying local adaptation. Although a vast amount of AFLP variation was attributed to residual effects, a significant amount of explained variation attributed to environmental effects was found. In conclusion, environmental variables include temperature, precipitation, soil moisture, surface coverage activity, and wet days may have been the most important ecological drivers for adaptive evolution of postglacial expanded *M. basjoo* var. *formosana* populations, and have played roles in shaping the current distributions of this species.

Methods

Sampling. A total of 112 individuals from eight populations of *M. basjoo* var. *formosana* were collected (Fig. 1; Tables 1 and 2). All samples were subjected to AFLP genotyping and 46 were used to obtain the Cu/Zn *SOD2* second intron sequences.

Cloning and sequencing of Cu/Zn *SOD2* second intron. The extraction of genomic DNA and total RNA from leaves followed the methods of Dellaporta *et al.*[75] and Clendennen and May[76], respectively. First-stranded cDNA was synthesized using total RNA and reverse transcribed (MMLV reverse transcriptase, Promega), and amplified using RACE kit (Invitrogen) with degenerate primers (SOD-F1, 5′-CTCRMKCCDGGNCTCCATGGCTTCC-3′; and SOD-R1, 5′-TTTCCKTCRTCRCCMRCATG-3′). RACE products were then ligated into pGEM-T vector (Promega) and transformed to *E. coli*. Two full-length Cu/Zn *SOD* coding sequences were cloned and sequenced (mbCSD1: ABI34606 and mbCSD2: ACY24898).

The *Bam*HI-, *Bgl*II-, *Eco*RI-, or *Sac*I-digested genomic DNA was self-religated and used for the first inverse polymerase chain reaction (IPCR) with CSD-F2 and CSD-R2 primers (5′-GGTGATACCACCAACGGCTGC-3′ and 5′-GCAGCCGTTGGTGGTATCACC-3′). The first IPCR was performed at 94 °C for 6 min, 25 cycles of 20 sec at 94 °C, 15 sec at 60 °C, 2 min at 72 °C, followed by 5 min at 72 °C. The first IPCR amplified fragments were used as template in the second IPCR with CSD-F3 (5′-CACGTGATGAGGAACGACATGC-3′) and CSD-R3 (5′-GCAGCCGTTGGTGGTATCACC-3′) primers in a PCR reaction at 94 °C for 4 min, 25 cycles of 30 sec at 94 °C, 30 sec at 60 °C, 2 min at 72 °C, followed by 5 min at 72 °C. The second IPCR products were subcloned to pUC119 and sequenced (Genbank accession numbers for genomic sequences of Cu/Zn *SOD1* and Cu/Zn *SOD2* are DQ866814 and GU045759, respectively).

An initial screening found no variation in introns of Cu/Zn *SOD1* and Cu/Zn *SOD2* except the second intron of Cu/Zn *SOD2* in several samples from different populations. Specific CSD2-F1 (5′-CTCCACTGGTAAACCCTCG-3′) and CSD2-R1 (5′-GAGGTCCTGCATGACAACAAG-3′) primers were used to amplify Cu/Zn *SOD2* second intron sequences of 46 individuals in PCR reactions with 6 min at 94 °C for 6 min, followed by 30 cycles of 30 sec at 94 °C, 30 sec at 57 °C, 1.5 min at 72 °C, and 5 min holding at 72 °C. The sequential PCR fragments were ligated into pGEM-T vector, transformed into *E. coli*, and double sequenced using M13 primers and haplotype sequences deposited (GenBank accessions: KX688826~KX688841).

Sequence alignment, nucleotide diversity, haplotype, and demography. Sequences were aligned with Clustal X[77] and nucleotide diversity θ_π and θ_S, haplotype diversity (*Hd*), R_2 index, and pairwise mismatch

distribution estimated using DnaSP v5.0[78]. θ_π and θ_S were calculated based on the average pairwise number of differences between sequences and the number of segregating sites per sequence, respectively. Population pairwise F_{ST}, raggedness (*rg*) index of observed mismatch distribution, and neutrality statistics (Tajima's *D* and Fu's *Fs*) were estimated using Arlequin v3.5[79]. Moreover, goodness-of-fit of the observed mismatch distribution to that expected under the spatial expansion model was tested using the sum of square deviations (SSD) statistic with Arlequin. We generated a haplotype network with the R "pegas" package[80,81].

Neutrality statistics are nearly zero in constant-size populations, whereas a significant negative value and significant positive value reflect processes such as population subdivision or recent bottlenecks. Positive R_2 value and significance of coalescent simulation against neutral model indicate population expansion. The raggedness index is a measure that quantifies the smoothness of the observed mismatch distribution and small *rg* values represent a population that has experienced sudden expansion[32]. A significant SSD value indicates departure from spatial expansion[34].

To estimate the time since the beginning of an expansion, we used $t = \tau/2\mu k$, where t is the time elapsed between initial and current population sizes, τ is the estimated number of generations since the expansion, μ is the mutation rate, and k is the sequence length. A mutation rate of 1.5% per 10^6 year per site was used[19]. Demographic expansion factor (τ) was estimated using Arlequin. We assumed a mean generation time of 2 years for *M. basjoo* var. *formosana*.

AFLP. Genomic DNA (200 ng) was digested with 5 U *Eco*RI and 5 U *Mse*I (Yeastern Biotech, Taipei, Taiwan) for AFLP genotyping[28]. Digested products were ligated with *Eco*RI (0.5 μM) and *Mse*I (5 μM) adaptors in a 10-μl ligation reaction mixture using 30 U T4 DNA ligase (Yeastern Biotech) at 22 °C for 1 hr. Preselective amplification was performed in a total volume of 10-μl reaction buffer, including 1 μl of the digested sample (1:9 dilution), 100 nM of *Eco*RI (E00: 5′-GACTGCGTACCAATTC-3′) and *Mse*I (M00: 5′-GATGAGTCCTGAGTAA-3′) primers (Supplementary Table S2), 0.25 mM dNTPs, and 1 U *Taq* (Bernardo Scientific, Taipei, Taiwan). The thermal cycling parameters for preselective amplification were as follows: 2 min at 72 °C and 3 min at 94 °C, followed by 25 cycles of 30 sec at 94 °C, 30 sec at 56 °C, and 5 min holding at 72 °C. Twelve *Eco*RI (labeled with FAM) and *Mse*I selective primer combinations (E00 and M00 primers with three additional bases; Supplementary Table S2) were used in selective amplification. Selective amplification was performed with 1 μl diluted preselective amplification product (1:9 dilution) in a 10-μl 1 x PCR buffer containing *Eco*RI and *Mse*I (both 100 nM) selective primers, 0.25 mM dNTPs, 0.75 U *Taq* with an initial holding at 94 °C for 3 min, followed by 12 cycles of 30 sec at 94 °C, 30 sec at 65 °C with a 0.7 °C touchdown per cycle and 1 min at 72 °C, followed by 24 cycles of 30 sec at 94 °C, 30 sec at 56 °C, 1 min at 72 °C, with a final 5 min holding at 72 °C. Selective amplification were visualized on an ABI PRISM 3100 sequencer (Applied Biosystems, Foster City, CA, USA). AFLP fragments were scored with Peak Scanner v2.0 (Applied Biosystems) for each individual in the range of 150–500 bp with relative fluorescent unit threshold set at 100, and genotyping error rate estimated (Supplementary Table S2). AFLP genotyping data are available in the Supplementary Data.

Environmental variables. Seven environmental variables, including annual mean temperature (BIO1), annual precipitation (BIO12), number of days with > 0.1 mm of rain per month (wet days), mean wind speed (WS_{mean}), normalized difference vegetation index (NDVI), soil pH, and soil moisture index (TMI) were retained as explanatory variables (Supplementary Methods and Supplementary Table S4).

AFLP diversity, structure, and relationships. Nei's genetic diversity (H_E)[38] and pairwise F_{ST} based on a Bayesian method with non-uniform prior distribution of allele frequencies[82] was calculated using AFLP-SURV[40]. The numbers of private and fixed private bands were calculated using FAMD[83]. Hierarchical analysis of molecular variance (AMOVA) was performed with the "poppr" package of R[84] and significance tested based on 9999 permutations with the "ade4" package of R[85]. The Bayesian program, HICKORY v1.1[39], was used to estimate an F_{ST} analogue (designated θ^B) from dominant markers accounting for inbreeding coefficient (f) using default settings for sampling and chain length parameters (burnin = 5,000; samples = 100,000; thinning = 20). Four models, including a full model, $f = 0$ model, $\theta^B = 0$ model, and f-free model were fitted with the AFLP data (Supplementary Methods).

Initial detection of genetic structure of *M. basjoo* var. *formosana* was carried out with two assignment test methods: Bayesian clustering[41] and sparse non-negative factorization (sNMF)[42] methods. Bayesian clustering method adapted for dominant markers implemented in STRUCTURE[41] was used to estimate an individual's probability of belonging to a homogeneous cluster (K populations). An admixture model was adopted and tested with ten runs, for K ranging from 1 to 9, with 10^6 iterations and 10^5 burn-in steps. R package "pophelper"[86] was used to summarize mean log likelihood (LnP(D)), change in the log probability (ΔK)[87], and symmetric similarity coefficient (SSC)[88] to evaluate the fit of different clustering scenarios. Genetic assignment of individuals was also inferred for K values ranging from 1 to 9 based on sNMF algorithm and least-squares optimization with the "LEA" package of R[42]. The settings for the LEA analysis were: regularization parameter = 100, iterations = 200, and repetitions = 10 with other arguments set to defaults, and the best K evaluated with the means of minimal cross-entropy.

The relationships between individuals and populations of *M. basjoo* var. *formosana* were also visualized based on the discriminant analysis of principal components (DAPC) using the R "adegenet" package[43,89]. DAPC first performed a principal component analysis (PCA), and an optimal number of PCs was estimated and retained for further discriminant analysis.

Effect of environmental variables on AFLP variation. Redundancy analysis (RDA) was used to partition AFLP variation and explained by environmental and geographical variables using the R "vegan" package[90] and significance tested with 9,999 permutations. Proportions of AFLP genetic variation (adjusted R^2) explained by pure environmental variables, geographically structured environmental variables, pure geographical variables, and residual components were estimated[29]. The longitude and latitude of sample localities were used as geographical variables.

AFLP outlier detection and association with environmental variables. Two F_{ST} outlier detection methods (DFDIST and BAYESCAN) were used to identify AFLP loci that potentially evolved under selection. DFDIST implemented in Mcheza software[46] estimated allele frequencies based on the Bayesian approach[82] and the highest and lowest 30% of the initial F_{ST} were removed for calculating the mean F_{ST}. Outliers were identified by observed F_{ST} and H_E compared to simulated neutral distributions generated using 10^5 iterations of coalescent simulations. Loci with unusually high F_{ST} values at the 95% confidence level by applying a false discovery rate of 5% were considered potentially evolved under selection. BAYESCAN uses a Bayesian likelihood approach via a reversible-jump Markov Chain Monte Carlo algorithm in comparing the selection versus neutrality model to identify AFLP loci that are potentially evolved under selection[47]. Posterior odds (PO), the ratio of posterior probabilities of selection over neutrality, was estimated with the settings of a sample size of 50,000 and thinning interval of 20 among 10^6 iterations, following 100 pilot runs of 50,000 iterations. When an AFLP locus with $\log_{10}(PO) > 1.5$[91] was considered to have strong evidence for selection.

Samβada[48] was further used to evaluate the associations between frequencies of AFLP loci and values of environmental variables using logistic regression model; and significant fit was identified based on the Wald and likelihood ratio tests with Bonferroni correction at $P < 0.01$. Given the relationships between values of environmental variables and frequencies of AFLP outlier loci, logistic regression plots were depicted.

References

1. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913 (2000).
2. Hewitt, G. M. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc. Lond.* **58**, 247–276 (1996).
3. Siame, L. *et al*. Glacial retreat history of Nanhuta Shan (north-east Taiwan) from preserved glacial features: the cosmic ray exposure perspective. *Quat. Sci. Rev.* **26**, 2185–2200 (2007).
4. Tsukada, M. Late Pleistocene vegetation and climate in Taiwan (Formosa). *Proc. Natl. Acad. Sci. USA* **55**, 543–548 (1966).
5. Petit, R. *et al*. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**, 1563–1565 (2003).
6. Kuo, D. C. *et al*. Two genetic divergence centers revealed by chloroplastic DNA variation in populations of *Cinnamomum kanehirae* Hay. *Conserv. Genet.* **11**, 803–812 (2010).
7. Stölting, K. N. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Mol. Ecol.* **22**, 842–855 (2013).
8. Jump, A. S., Hunt, J. M., Martínez-Izquierdo, J. A. & Peñuelas, J. Natural selection and climate change: temperature-linked spatial and temporal trends in gene frequency in *Fagus sylvatica*. *Mol. Ecol.* **15**, 3469–3480 (2006).
9. 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 (2013).
10. 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 (2013).
11. Huang, C. L. *et al*. Influences of environmental and spatial factors on genetic and epigenetic variations in *Rhododendron oldhamii* (Ericaceae). *Tree Genet. Genom.* **11**, 823 (2015).
12. Huang, C. L. *et al*. Genetic relationships and ecological divergence in *Salix* species and populations in Taiwan. *Tree Genet. Genomes* **11**, 39 (2015).
13. Nakazato, T., Warren, D. L. & Moyle, L. C. Ecological and geographic modes of species divergence in wild tomatoes. *Am. J. Bot.* **97**, 680–693 (2010).
14. Escudero, A., Iriondo, J. M., Torres, M. E. (2003) Spatial analysis of genetic diversity as a tool for plant conservation. *Biol. Conserv.* **113**, 351–365 (2003).
15. Huang, S. F., Hwang, S. Y. & Lin, T. P. Spatial pattern of chloroplast DNA variation of *Cyclobalanopsis glauca* in Taiwan and East Asia. *Mol. Ecol.* **11**, 2349–2358 (2002).
16. 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 (2003).
17. Huang, S. F., Hwang, S. Y., Wang, J. C. & Lin, T. P. Phylogeography of *Trochodendron aralioides* (Trochodendraceae) in Taiwan and its adjacent areas. *J. Biogeogr.* **31**, 1251–1259 (2004).
18. Cheng, Y. P., Hwang, S. Y. & Lin, T. P. Potential refugia in Taiwan revealed by the phylogeographical study of *Castanopsis carlesii* Hayata (Fagaceae). *Mol. Ecol.* **14**, 2075–2085 (2005).
19. 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 (2010).
20. Su, H. J. Studies on the climate and vegetation types of the natural forests in Taiwan (III): a scheme of geographic regions. *Quar. J. Chin. For.* **18**, 33–44 (1985).
21. Makino, T. Makino's New Illustrated Flora of Japan (The Hokuryukan Co., Ltd. Tokyo, 1979).
22. Liu, A. Z., Li, D. Z. & Li, X. W. Taxonomic notes on wild bananas (*Musa*) from China. *Bot. Bull. Acad. Sin.* **43**, 77–81 (2002).
23. Janssens, S. B. *et al*. Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. *New Phytol.* **210**, 1453–1465 (2016).
24. Ying, S. S. Musaceae in *Flora of Taiwan*. (2nd ed.) **5**, 704–706 (National Taiwan University Press, Taipei).
25. Pillay, M., Tenkouano, A. & Hartman, J. Bananas and plantains: future Challenges In *Musa breeding in Crop Improvement, Challenges in the Twenty-First Century*. Chapter 8, 223–252 (Food Products Press, New York, 2002).
26. Chiu, H. L., Chen, L. F., Shii, C. T. & Chang, Y. C. Study on ploidy of *Musa formosana* (Warb.) Hayata in Taiwan. *J. Taiwan Agric. Res.* **59**, 78–85 (2010).
27. Alberto, F. J. *et al*. Potential for evolutionary responses to climate change - evidence from tree populations. *Glob. Change Biol.* **19**, 1645–1661 (2013).
28. Vos, P. *et al*. AFLP: a new technique for DNA fingerprinting. *Nucleic. Acids Res.* **23**, 4407–4414 (1995).
29. Borcard, D., Legendre, P. & Drapeau, P. Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045–1055 (1992).
30. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

31. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
32. Harpending, H. C. Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. *Hum. Biol.* **66**, 591–600 (1994).
33. Ramos-Onsins, S. E. & Rozas, J. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19**, 2092–2100 (2002).
34. Schneider, S. & Excoffier, L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates very among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089 (1999).
35. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569 (1992).
36. Yu, G. *et al*. Palaeovegetation of China: a pollen data-based synthesis for the mid-Holocene and last glacial maximum. *J. Biogeogr.* **27**, 635–664 (2000).
37. Bartlein, P. J. *et al*. Pollen-based continental climate reconstructions at 6 and 21 Ka: a global synthesis. *Clim. Dyn.* **37**, 75–802 (2011).
38. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* **70**, 3321–3323 (1973).
39. Holsinger, K. E. & Lewis, P. O. Hickory: a package for analysis of population genetic data v1.0. http://www.academia.edu/1839794/HICKORY (2003).
40. Vekemans, X., Beauwens, T., Lemaire, M. & Roldán-Ruiz, I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Mol. Ecol.* **11**, 139–151 (2002).
41. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).
42. Frichot, E. & Francois, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925–929 (2015).
43. 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 (2010).
44. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
45. Thornthwaite, C. W. An approach toward a rational classification of climate. *Geogr. Rev.* **38**, 55–94 (1948).
46. Antao, T. & Beaumont, M. A. Mcheza: A workbench to detect selection using dominant markers. *Bioinformatics* **15**, 1717–1718 (2011).
47. Foll, M. & Gaggiotti, O. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993 (2008).
48. Stucki, S. *et al*. High performance computation of landscape genomic models integrating local indices of spatial association. *Mol. Ecol. Resour. Accepted Author Manuscript*.. doi:10.1111/1755-0998.12629 (2016).
49. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
50. Cottenie, K. Integrating environmental and spatial processes in ecological community dynamics. *Ecol. Lett.* **8**, 175–1182 (2005).
51. Legendre, P. *et al*. Partitioning beta diversity in a subtropical broad-leaved forest of China. *Ecology* **90**, 663–674 (2009).
52. Maruyama, T. & Fuerst, P. A. Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111**, 675–689 (1985).
53. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
54. Rogers, A. R. Genetic evidence for a Pleistocene population expansion. *Evolution* **49**, 608–615 (1995).
55. Ray, N., Currat, M. & Excoffier, L. Intra-deme molecular diversity in spatially expanding populations. *Mol. Biol. Evol.* **20**, 76–86 (2003).
56. 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 (2004).
57. Chung, J. D., Lin, T. P., Tan, Y. C., Lin, M. Y. & Hwang, S. Y. Genetic diversity and biogeography of *Cunninghamia konishii* (Cupressaceae), an island species in Taiwan: a comparison with *Cunninghamia lanceolata*, a mainland species in China. *Mol. Phylogenet. Evol.* **33**, 791–801 (2004).
58. Wu, S. H. *et al*. Contrasting phylogeographic patterns of two closely related species, *Machilus thunbergii* and *Machilus kusanoi* (Lauraceae), in Taiwan. *J. Biogeogr.* **33**, 936–947 (2006).
59. 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 (2006).
60. Nybom, H. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.* **13**, 1143–1155 (2004).
61. Wang, X. L., Chiang, T. Y., Roux, N., Hao, G. & Ge, X. J. Genetic diversity of wild banana (*Musa balbisiana* Colla) in China as revealed by AFLP markers. *Genet. Resour. Crop Evol.* **54**, 1125–1132 (2007).
62. 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 (2016).
63. Hamrick, J. L. & Godt, M. J. W. Effects of life history traits on genetic diversity in plant species. *Philos. Trans. R. Soc. Lond. B* **351**, 1291–1298 (1996).
64. Wong, C. *et al*. Genetic diversity of the wild banana *Musa acuminate* Colla in Malaysia as evidenced by AFLP. *Ann. Bot.* **88**, 1017–1025 (2001).
65. Nakazato, T., Bogonovich, M. & Moyle, L. C. Environmental factors predict adaptive phenotypic differentiation within and between two wild Andean tomatoes. *Evolution* **62**, 774–792 (2008).
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 (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 (2012).
68. Bothwell, H. *et al*. Identifying genetic signatures of selection in a non-model species, alpine gentian (*Gentiana nivalis* L.), using a landscape genetic approach. *Conserv. Genet.* **14**, 467–481 (2013).
69. John, R. *et al*. Soil nutrients influence spatial distributions of tropical tree species. *Proc Natl Acad Sci USA* **104**, 864–869 (2007).
70. Mitton, J. B., Grant, M. C. & Yoshino, A. M. Variation in allozymes and stomatal size in pinyon (*Pinus edulis*, Pinaceae), associated with soil moisture. *Am. J. Bot.* **85**, 1262–1265 (1998).
71. Pluess, A. R. & Weber, P. Drought-adaptation potential in *Fagus sylvatica*: linking moisture availability with genetic diversity and dendrochronology. *PLoS One* **7**, e33636 (2012).
72. Brousseau, L., Foll, M., Scotti-Saintagne, C. & Scotti, I. Neutral and adaptive drivers of microgeographic genetic divergence within continuous populations: the case of the Neotropical tree *Eperua falcate* (Aubl.). *PLoS One* **10**, e0121394 (2015).
73. Huemmrich, K. F., Privette, J. L., Mukelabai, M., Myneni, R. B. & Knyazikhin, Y. Time-series validation of MODIS land biophysical products in a Kalahari woodland, Africa. *Int. J. Remote Sens.* **26**, 4381–4398 (2005).
74. Violle, C. *et al*. The return of the variance: intraspecific variability in community ecology. *Trends Ecol. Evol.* **27**, 244–252 (2011).

75. Dellaporta, S. L., Wood, J. & Hicks, J. B. A plant DNA minipreparation: Version II. *Plant Mol. Biol. Rep.* **1**, 19–21 (1983).
76. Clendennen, S. K. & May, G. D. Differential gene expression in ripening banana fruit. *Plant Physiol.* **115**, 462–469 (1997).
77. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).
78. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).
79. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 64–567 (2010).
80. Paradis, E. Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420 (2010).
81. R Development Core Team. R: A Language and Environment for Statistical Computing, Version 3.0.0. http://www.R-project.org/ (2013).
82. Zhivotovsky, L. A. Estimating population structure in diploids with multilocus dominant DNA markers. *Mol. Ecol.* **8**, 907–913 (1999).
83. Schlüter, P. M. & Harris, S. A. Analysis of multilocus fingerprint data sets containing missing data. *Mol. Ecol. Notes* **6**, 569–572 (2006).
84. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *Peer J.* **2**, e281 (2014).
85. Dray, S. & Dufour, A. B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
86. Francis, R. M. Pophelper: an r package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).
87. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
88. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
89. Jombart, T. Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
90. Oksanen, J. *et al*. Vegan: community ecology package. R package version 2.0-1. https://cran.r-project.org/web/packages/vegan/ (2011).
91. Jeffreys, H. *Theory of probability* (3rd ed, Oxford University Press, Oxford, 1961).
92. Menon, S. ArcGIS 10.3: The next generation of GIS is here. Environmental Systems Research Institute, Inc., CA, USA. http://www.esri.com/software/arcgis (2014).
93. Open Government Data Providing Organization in Taiwan. http://data.gov.tw/node/35430. Open Government Data License, Version 1.0: http://data.gov.tw/license#eng.

Acknowledgements

This work was supported partly by the Taiwan Council of Agriculture under grant numbers 93AS-4.1.1-FB-e2-3 and 94AS-9.1.7-FB-e1-11 to C.W.S. and by the Taiwan Ministry of Science and Technology under grant number of MOST 103-2313-B-003-001-MY3 to S.Y.H. and MOST 104-2311-B-003-002-MY3 to C.W.S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

C.W.S. and S.Y.H. conceived and designed the experiments. Y.L.L. and P.C.L. collected the samples. Y.L.L. performed the IPCR experiment and characterized the *Cu/Zn SOD2* sequences. J.H.C. performed AFLP experiment and haplotype analysis. C.L.H., C.T.C. and S.Y.H. provided analysis tools and performed the analyses for AFLP diversity and its association with environmental variables. J.H.C., C.L.H., S.Y.H. and C.W.S. wrote the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-05256-6

Competing Interests: The authors declare that they have no competing interests.

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