Journal of Heredity 2012:103(6):853–862
doi:10.1093/jhered/ess061
Advance Access publication September 24, 2012

Genetic Diversity and Population Structure of Chinese White Poplar (*Populus tomentosa*) Revealed by SSR Markers

Qingzhang Du, Bowen Wang, Zunzheng Wei, Deqiang Zhang, and Bailian Li

From the National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, Peoples' Republic of China (Du, Wang, Wei, Zhang, and Li); and Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, Peoples' Republic of China (Du, Wang, Wei, Zhang, and Li)

Address correspondence to: Deqiang Zhang at the address above, or e-mail: DeqiangZhang@bjfu.edu.cn

Abstract

An understanding of allelic diversity and population structure is important in developing association studies and constructing core collections for tree breeding. We examined population genetic differentiation in the native *Populus tomentosa* by genotyping 460 unrelated individuals using 20 species-specific microsatellite markers. We identified 99 alleles with a mean of 4.95 observed alleles per locus, indicating a moderate level of polymorphism across all individuals. A model-based population structure analysis divided *P. tomentosa* into 11 subpopulations ($K = 11$). The pattern of individual assignments into the subsets ($K = 3$) provided reasonable evidence for treating climatic zones as genetic regions for population genetics. The highest level of genetic variation was found in the southern region (i.e., $N = 93$, $N_P = 11$, $H_E = 0.445$, $F = –0.102$), followed by the northeastern and northwestern regions. Thus, the southern region is probably the center of the current species distribution. No correlation was found between population genetic distance and geographic distance ($r = 0.0855$, $P = 0.3140$), indicating that geographical distance was not the principal factor influencing genetic differentiation in *P. tomentosa*. These data provide a starting point for conserving valuable natural resources and optimizing breeding programs.

Key words: *climatic regions, genetic differentiation, heterozygosity, microsatellite loci, Populus tomentosa*

Chinese white poplar (*Populus tomentosa* Carr.), a diploid ($2n = 2x = 38$), outcrossing perennial species, belongs to the section *Populus* (*Leuce*) in the genus *Populus*. In China, *P. tomentosa* is one of the main commercial tree species used for timber production. *Populus tomentosa* plays a major role in ecological and environmental protection along the Yellow River (Zhu and Zhang 1997). Many wild *P. tomentosa* ecotypes have arisen during the evolution of the species (Zhang et al. 2007). This accumulated genetic variation provides a potential source of beneficial alleles for poplar breeding and improvement. Conventional tree-breeding programs to improve this slow-growing species might not be effective (Zhang et al. 1992a), but molecular tools could remedy this deficiency and play an important role in the efficient management and utilization of genetic resources. Hence, a need exists for a more comprehensive analysis of genetic diversity and population structure in *P. tomentosa*, as the foundation of molecular marker-assisted selection breeding. For such analyses, using suitable genetic marker techniques is important.

To date, little research has investigated the genetic diversity and population structure in natural populations of *P. tomentosa*. However, some studies have examined climatic regionalization in its distribution, clonal geographical variation, and flowering and seeding patterns (Huang 1992a, 1992b; Zhang et al. 1992a). Previous studies that examined genetic diversity were based on allozymes and amplified fragment length polymorphism (AFLP) markers in limited samples of *P. tomentosa* (Zhang et al. 1992b; He 2005). However, an overview of the genetic and population structure of the germplasm is not available for this important poplar resource in China. To assess genetic diversity and population structure in crop and tree species, simple sequence repeat (SSR) markers are ideal because they are hypervariable,

Table 1 Geographical and climate parameters of native *P. tomentosa* used in this study

Climatic region	Number of genotypes	Populations used in this study	Geographical origin	Longitude (°E)	Latitude (°N)	Altitude (m asl)	Annual rainfall (mm)	Annual average temperature (°C)	Sunshine duration (h/year)	Accumulated temperature (°C)
Northeastern (NE)	175	Beijing (BJ) Hebei (HB) Shandong(SD)	Beijing, Hebei and almost all the region of Shandong	116.43 (110.7–120.4)	34.22 (31.6–37.1)	224	606	12.2	2697	4274
Southern (S)	190	Henan(HN) Shaanxi (SAX) Anhui (AH) Jiangsu (JS)	Northwest of Anhui, north and middle regions of Henan, and Kuan-chung Plain of Shaanxi.	113.22 (107.7–117.9)	38.18 (36.4–40.4)	66	733	15.9	2290	4546
Northwestern (NW)	95	Gansu (GS) Shanxi (SX)	Gansu, Ningxia, North and middle region of Shanxi province.	110.2 (105.2–113.6)	36.33 (34.2–37.8)	1098	544	8.84	2467	3063

co-dominant, and highly informative (López-Gartner et al. 2009; Blair et al. 2010). Many SSR markers have been developed for *Populus trichocarpa* and *Populus tremuloides* (Rahman et al. 2000; Yin et al. 2009), but only limited microsatellite markers for *P. tomentosa* were detected by Du et al. (2012). Therefore, to improve the utility and scope of marker-assisted selection (MAS) breeding programs, developing species-specific SSR markers that can be used to examine genetic diversity and population structure in *P. tomentosa* is essential.

In the present study, SSR markers for *P. tomentosa* were developed from genomic clones using a direct sequencing approach. These markers were then used to evaluate genetic diversity and population structure among and within provenances and climatic regions. The specific aims of our study were as follows: to assess the polymorphism and transferability of SSR markers in related species within the genus *Populus*; evaluate genetic diversity and the pattern of population structure; based on the results of the structure analysis, test whether the climatic regions described by Huang (1992a) can be treated as genetic regions in the analyses of molecular variance (AMOVAs) and to infer the degree of population genetic differentiation, and test the correlation between genetic distance and geographic distance in *P. tomentosa*.

Materials and Methods

Plant Materials

In 1982, 1047 native individuals of *P. tomentosa* were collected throughout the species' entire natural distribution range, covering an area of 1 million km^2. Root segments were used to establish a clonal arboretum in Guan Xian County, Shandong Province. The climatic zone from which these individuals were collected can be divided into three large climatic regions, southern (S), northwestern (NW), and northeastern (NE), on the basis of a principal components analysis and isodata fuzzy clustering using 16 meteorological factors (Huang 1992a). In the present study, 460 unrelated individuals representing almost the entire geographic distribution of *P. tomentosa* (190 specimens from the S region, 95 from the NW region, and 175 from the NE region) were used to assess genetic diversity and population structure (Table 1; Figure 1). Additionally, 45 *P. tomentosa* individuals were randomly selected and used to develop SSR loci using a re-sequencing method.

One thousand F_1 progeny that were derived from a controlled crossing between a female clone (P. alba × P. glandulossa) and a male clone "LM 50" (*P. tomentosa*) were used to conduct a Mendelian segregation testing for SSR loci. The progeny were grown in the Xiao Tangshan horticulture fields of Beijing Forestry University, Beijing, China (40°2'N, 115°50'E), using a randomized complete block design with three replications in 2008.

SSR Loci Development and Primer Design

In this study, to identify species-specific SSR markers in *P. tomentosa*, a series of specific primers were designed for

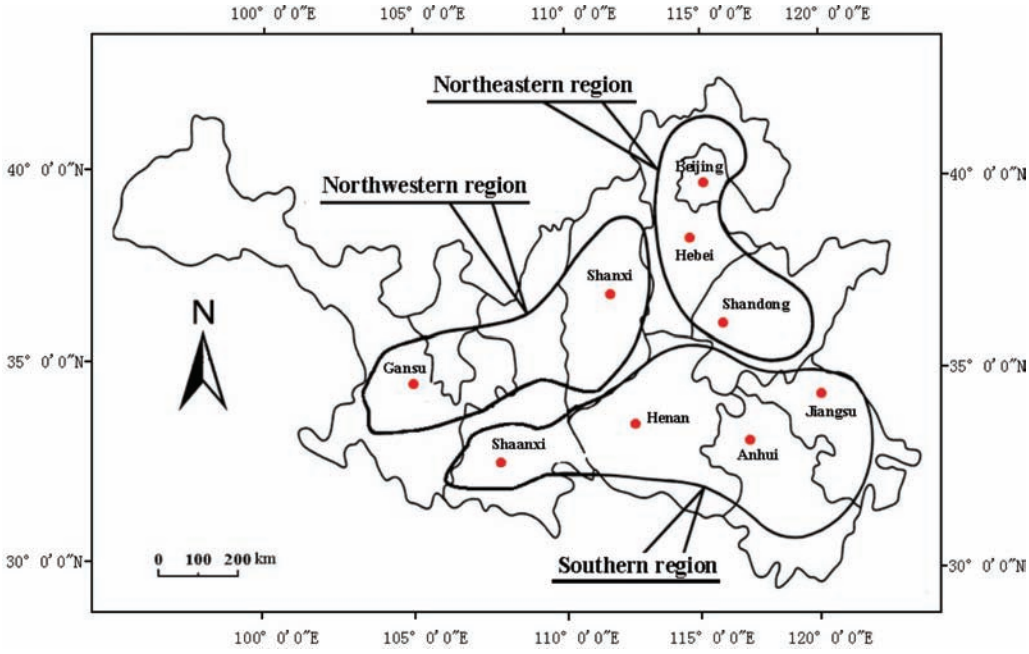

Figure 1. Map of climatic regionalization and sampled provenances of the wild *P. tomentosa* population used in this study. The three climatic regions are indicated by thick black lines (based on Huang 1992a), the locations of the nine sampled populations are shown by solid red circles (see Table 1 for the details of these nine sampled locations), and the thin black lines represent administrative divisions among different provinces in China.

PCR amplification of genomic random segments and total genomic DNA (20 ng per reaction), which were separately isolated from 45 unrelated *P. tomentosa* individuals and used as amplification templates. Subsequently, all of the PCR amplification products from the 45 genomic clones were sequenced (both strands) using conservative primers and a BigDye Terminator version 3.1 Cycle Sequencing kit (Applied Biosystems, Beijing, China) on a genetic analyzer (4300; Li-Cor, Lincoln, NE, USA). The re-sequenced genomic DNA segments from the 45 clones were aligned using MEGA version 5.1 (http://www.megasoftware.net/), and microsatellite loci were detected using the software SSRIT (Temnykh et al. 2001) with SSR motifs ranging from 2 to 15. Finally, polymorphisms in the SSR markers were identified by comparing them with allele ranges for each microsatellite locus among the 45 aligned sequences (with at least a 5% minor allele frequency). Primer pairs flanking the SSR regions were designed using the Primer 3 software (http://frodo.wi.mit.edu/primer3/primer3.FAQ.html), screened, and then optimized for reproducible amplification using standard PCR conditions with annealing temperatures altered according to the primer sequences. The cross-species transferability of SSR markers was tested in 20 species (a single genotype per species) within the genus *Populus* (*Salix matsudana* was used as the outgroup) (Supplementary Table 1 online). Linkage disequilibrium (LD) between pairs of microsatellite loci was estimated using GenePop version 4.0 (Raymond and Rousset 1995); this tested the independence of makers to improve the evaluation of among-population differences.

DNA Extraction and SSR Genotyping

For each sample, total genomic DNA was isolated from young leaves using a DNeasy Plant Mini kit (Qiagen China, Shanghai), following the manufacturer's protocol. The SSR amplification was performed in a 25-μL reaction volume containing 0.8 U Taq DNA polymerase (Promega, Sunnyvale, CA, USA), 0.3 μM each of forward and reverse primers (forward primers for the SSR loci were labeled with fluorescent dye (6-FAM)), 1× colorless GoTaq buffer (Promega), 0.2 mM dNTPs (Promega), and 20 ng total genomic DNA. The PCR conditions were as follows: denaturation at 94 °C for 3 min, followed by 30 cycles of PCR amplification, each consisting of denaturation at 94 °C for 3 min, annealing at 50–58 °C (depending on the primer pair) for 30 s, and extension at 72 °C for 40 s. The PCR products were finally extended at 72 °C for 5 min. The PCR products were separated by capillary electrophoresis using an ABI3730xl DNA Analyzer (Applied Biosystems) after confirmation of PCR amplification on a 1.5% agarose gel. The polymorphic loci analysis was performed using the software GeneMapper v4.0 (Applied Biosystems) with the LIZ 600 size standard (Applied Biosystems). Subsequently, MICRO-CHECKER 2.2.3 (http://www.microchecker.hull.ac.uk/) was used to check for null alleles, scoring errors, and large allele dropout. The results provided by this program are used as evidence of null alleles, and these null alleles should be removed or adjusted before they are used in subsequent genetic analyses (van Oosterhout et al. 2004).

Table 2 Diversity information parameters at 20 SSR loci

	N_A	N_E	I	PIC	H_O	H_E	F_{IS}	F_{IT}	F_{ST}	P-Value
SSR1	9	3.02	1.47	0.640	0.769	0.670	−0.2062	−0.1790	0.0325	0.0000
SSR2	6	1.27	0.48	0.205	0.208	0.215	0.0093	0.0441	0.0629	0.0000
SSR3	3	1.96	0.72	0.379	0.776	0.491*	−0.6408	−0.6076	0.0403	0.0000
SSR4	3	2.24	0.88	0.454	0.962	0.555*	−0.7700	−0.7656	0.0025	0.0156
SSR5	5	1.57	0.70	0.333	0.197	0.362	0.3999	0.4528	0.0981	0.0010
SSR6	6	2.12	0.94	0.460	0.716	0.530*	−0.4486	−0.3973	0.0454	0.0001
SSR7	4	2.09	0.82	0.416	0.804	0.523*	−0.5697	−0.5548	0.0395	0.0012
SSR8	4	1.28	0.47	0.206	0.183	0.216	0.0300	0.1230	0.0959	0.0001
SSR9	2	1.44	0.49	0.260	0.362	0.308	−0.3356	−0.2029	0.0994	0.0000
SSR10	5	2.16	0.87	0.429	0.773	0.537*	−0.5059	−0.4792	0.0377	0.0005
SSR11	5	2.16	0.86	0.430	0.858	0.538*	−0.6244	−0.6227	0.0211	0.2199
SSR12	5	2.13	0.84	0.423	0.756	0.531*	−0.5286	−0.4878	0.0267	0.0001
SSR13	6	1.66	0.85	0.374	0.365	0.396	0.0353	0.0684	0.0544	0.0020
SSR14	6	1.60	0.68	0.327	0.400	0.374	−0.0939	−0.0908	0.0079	0.1633
SSR15	7	2.79	1.18	0.574	0.947	0.643*	−0.5364	−0.5134	0.0450	0.0130
SSR16	4	2.58	1.09	0.541	0.961	0.613*	−0.5990	−0.5688	0.0389	0.0001
SSR17	5	1.31	0.52	0.224	0.175	0.236	0.2475	0.2621	0.0894	0.0002
SSR18	7	1.51	0.74	0.317	0.220	0.340	0.3049	0.3534	0.0697	0.0010
SSR19	3	1.59	0.61	0.311	0.438	0.371	−0.2278	−0.1136	0.0930	0.0001
SSR20	4	1.87	0.77	0.388	0.576	0.465	−0.2788	−0.2610	0.0339	0.0001
Average	4.95	1.92	0.80	0.385	0.572	0.446	−0.2671	−0.2257	0.0528	0.0033

N_A = Number of alleles per locus; N_E = Effective number of alleles; I = Shannon's Information index; H_O = Observed heterozygosity; H_E = Expected heterozygosity; F = Wright's fixation index; PIC = Polymorphism information content; F_{IS} = Inbreeding among individuals within subpopulations; F_{IT} = Inbreeding within entire population; F_{ST} = Variation due to differentiation among subpopulations; P-Value = Significant level for F-statistics (significance is $P \leq 0.0500$); * = Significant level for deviations from Hardy–Weinberg equilibrium (Significance is $P \leq 0.0100$).

Analysis of Genetic Diversity

The summary statistics reflected intra- and interpopulation genetic diversity levels, including the observed number of alleles per locus (N_A), the effective number of alleles (N_E), Shannon's Information index (I), polymorphism information content (PIC), observed heterozygosity (H_O), expected heterozygosity (H_E), and F-Statistics (F_{IS}, F_{IT}, F_{ST}) (Wright 1969). Values for these parameters were calculated using POPGEN VERSION 1.32 (Yeh et al. 1999). In addition, we carried out an AMOVA to partition the genetic variance among climatic regions, among populations within regions, and among individuals within populations. Hardy–Weinberg equilibrium (HWE) tests for these 20 SSR markers were conducted using Arlequin version 3.11 (http://cmpg.unibe.ch/software/arlequin3/); Arlequin 3.11 was also used to evaluate genetic diversity and genetic differentiation among and within climatic regions.

Mendelian inheritance of SSR loci, based on segregation analyses of progeny arrays, was examined to determine if excess heterozygotes of alleles existed in this species. The experiment was designed based on 14 SSR loci (randomly selected from Supplementary Table 2 online) and 1000 F_1 progeny. SAS FOR WINDOWS VER. 8.2 (SAS Institute, Cary, NC, USA) was used to conduct a chi square (χ^2) statistic for Mendelian inheritance segregation.

Analysis of Population Structure

We used the software package STRUCTURE VERSION 2.3.3 (http://pritch.bsd.uchicago.edu/structure.html) to assess population structure in *P. tomentosa*. The optimum number of subpopulations (K) was identified after 10 independent runs for each value of K ranging from 1 to 15, with a burn-in of 10 000 iterations followed by 100 000 iterations. This program applies a model-based clustering algorithm that implements the Markov chain Monte Carlo (MCMC) algorithm and a Bayesian framework. The algorithm identifies subgroups with distinctive allele frequencies and places individuals into K clusters using its estimated membership probability (Q). In this study, the model developed by Evanno et al. (2005) was used; it is an *ad hoc* procedure based on the second-order rate of change in the log probability (ΔK) of data between successive K values. ΔK is used to identify the break in the slope of the distribution of $L(K)$ at the true K. We used the height of this modal value as an indicator of the strength of the signal that was detected by the structure analysis (Evanno et al. 2005).

Furthermore, based on the identified population structure, we tested whether the climatic zones described by Huang (1992a) could be treated as genetic regions in the AMOVAs for this species.

Correlation between Genetic Distance and Geographic Distance

To assess whether more distant population pairs were more genetically different, the genetic and geographic distances between all pairs of subpopulations were measured. The comparable matrix of direct geographic distances among the nine geographic populations (See Table 1) was calculated according to latitude and longitude using Vincenty's

Table 3 Analysis of molecular variance (AMOVA) using 460 genotypes from nine native *P. tomentosa* populations from three climatic regions, based on 20 SSR markers

Source of variation	Degrees of freedom	Sum of squares	Mean square	Percentage variation (%)	*P*-Value
Among regions	2	368.252	184.126	10.7	<0.01
Among populations within regions	6	262.633	43.772	14.6	<0.01
Within populations	451	2261.736	5.015	74.7	<0.01
Total	459	2892.622	\	100	\

Table 4 Genetic diversity estimates for three climatic subregions of *P. tomentosa*

Subregion	Number of genotypes	*N*	N_A	N_E	Number of private alleles	*PIC*	H_O	H_E	F_{IS}	*P*-Value
Northeastern	175	81	4.05	1.90	5	0.374	0.622	0.442	−0.308	0.0010
Southern	190	93	4.65	2.00	11	0.401	0.519	0.445	−0.102	0.0012
Northwestern	95	69	3.45	1.80	2	0.325	0.586	0.388	−0.323	0.0001
Average	\	\	4.05	1.90	6	0.384	0.576	0.425	−0.244	0.0010

N = Number of observed alleles over all loci; N_A = Number of alleles per locus; N_E = Effective number of alleles; H_O = Observed heterozygosity; H_E = Expected heterozygosity; F_{IS} = Wright's fixation index; *PIC* = Polymorphism information content; *P*-Value = Significance level (F_{IS}) was $P \leq 0.0100$.

formula (http://www.movable-type.co.uk/scripts/latlong-vincenty.html). Genetic distance is represented by $F_{ST}/1-F_{ST}$ estimates (Rousset 1997) and pairwise F_{ST}; their respective *P*-values (to detect significant differences from zero) were calculated using Arlequin version 3.11. Finally, the program Distance Web Service (IBDWS) version 3.23 (http://ibdws.sdsu.edu/~ibdws/) was used to examine the relationship between $F_{ST}/1-F_{ST}$ and log-transformed geographic distance (Ln distance) for the analysis of isolation by distance (Mantel 1967; Rousset 1997; Jensen et al. 2005); significance was assessed by conducting 10 000 permutations.

Results

SSR Development and Transferability across *Populus* Species

According to the standard that the minor allele frequency was ≥5%, 20 novel SSR loci were identified from a total of 500 different random genomic sequences by comparing the data for these 45 clones with an average density of one SSR every 6.5 kb. Among these loci, eight SSR markers have been reported in the form of primer notes by Du et al. (2012) (Supplementary Table 2 online). The most abundant SSR loci were dinucleotide repeats (45%), followed by trinucleotide repeats (30%) (Supplementary Table 2 online). The remaining five SSR loci (25%) were repeated units that ranged in size from 4 to 12 nucleotides (Supplementary Table 2 online). To determine the potential transferability across species or genera, we tested the 20 SSR markers in 20 different species from five sections within the genus *Populus* (Supplementary Table 3 online). Cross-species transferability was scored positively only when clear band(s) were present. The average transferability rate for the 20 primer pairs tested was 100% in the section *Populus* (*Leuce*), but in other sections, it ranged from 80% (*Leucoides*) to 90% (*Tacamahaca*) (Supplementary Table 3 online). Thus, the developed SSR markers could be widely used in the genus *Populus* and provided considerable polymorphisms across 20 different species. Among the 20 species, 112 alleles were observed with the 20 SSR primers, for a mean number of 5.6 alleles per locus (Supplementary Table 3 online). No significant LD was detected between any of the SSR loci pairs in this study (data not shown), which suggests that these novel SSR markers could be used as independent genomic markers to assess genetic diversity and population structure in this species.

Assessment of Genetic Diversity

After the allele and genotype frequencies of the amplified alleles were adjusted using the program MICRO-CHECKER 2.2.3, the 20 SSR markers were used to evaluate genetic diversity across 460 genotypes of *P. tomentosa*. All of the markers showed a moderate level of polymorphism. Ninety-nine alleles were identified, with an average of 4.95 observed alleles per locus, ranging from two at SSR9 to nine at SSR1 (Table 2). As a measurement of the genetic diversity of the species, the *PIC* mean polymorphism level of the loci was 0.385, and it ranged from 0.205 for SSR2 to 0.640 for SSR1 (Table 2). The mean observed and expected levels of heterozygosity were 0.572 and 0.446, respectively (Table 2); H_O was higher than H_E at 14 SSR loci, with exceptions being SSR2, SSR5, SSR8, SSR13, SSR17, and SSR18, in accordance with the mean fixation index (F_{IS} = –0.2671; $P < 0.0500$). A 70% significant excess of heterozygotes was observed at 20 novel microsatellite loci (Table 2), and nine SSR loci showed significant departures from HWE ($P < 0.01$) in the natural population due to excess numbers of heterozygotes (Table 2).

Mendelian segregation tests for all 14 SSR loci showed that 10 of 12 heterozygote SSR loci for *P. tomentosa* followed Mendelian inheritance in the F_1 population (Supplementary Table 4 online), which supported the conclusion that excess heterozygotes did indeed exist in this species.

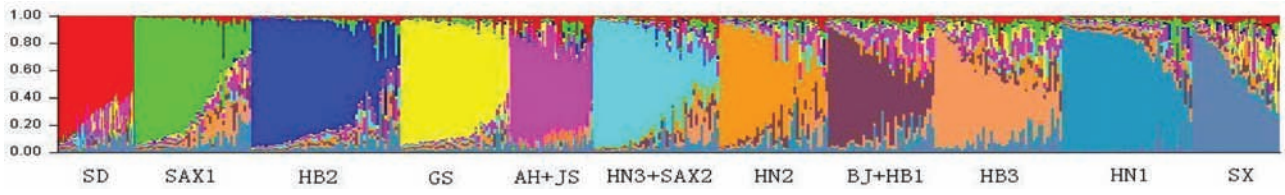

Figure 2. Estimation of subpopulations using 460 *P. tomentosa* individuals and 20 SSR markers without prior classification information. The entire population of *P. tomentosa* was divided into 11 subpopulations ($K = 11$) using STRUCTURE. Individuals are shown by thin vertical lines, which are divided into K colored segments standing for the estimated membership probabilities (Q) of each individual. Finally, all individuals are classified based on the Q-matrix.

Population Structure

Population structure analysis was conducted without prior classification information, yielding the optimal substructure result ($K = 11$). The statistical model described by Evanno et al. (2005) showed a clear peak ($\Delta K = 262.72$) at the value $K = 11$ (Supplementary Figure 1 online). Estimated subpopulations for the 460 *P. tomentosa* individuals are shown in Figure 2. Each individual is represented by a thin vertical line and classified based on its estimated membership probability (Q). Comparing the graphs of the Q-matrix across different values of K and combining information about the geographic distribution and evolutionary history of *P. tomentosa*, the subgroup value of $K = 11$ appeared to be optimal (Figure 2).

Similar to the hierarchical island model described by Evanno et al. (2005), the second highest peak ($\Delta K = 177.09$), which occurred at $K = 3$, was analyzed (Supplementary Figure 1 online). Division of the data set into three subclusters produced the best assignment of individuals to subclusters in STRUCTURE, and a subsequent analysis of each subset determined the appropriate number of populations within each subset. The pattern of individual assignment into the subsets ($K = 3$) (Supplementary Figure 2 online) generally agreed with the genotype clones that were assigned to the three different climatic regions (Table 1). The principle of the STRUCTURE model is the minimization of within-group LD and deviation from HWE, which provides evidence for treating climatic regions (also called geographical ecotypic zones) as genetic regions in analyses of population differentiation.

Analysis of Molecular Variance

Climatic zones, defined as genetic regions, provide a logical basis for organizing data for hierarchical AMOVA. The analysis indicated different levels of genetic variance among regions, among populations within regions, and among individuals within populations. Of the total genetic variance, 10.7% was attributed to regional divergence, 14.6% was ascribed to population differences within regions, and the remaining 74.7% was explained by individual differences within populations (Table 3). Significant genetic differentiation was noted among the three climatic regions ($P < 0.01$), among the nine populations within the regions ($P < 0.01$), and within individual populations ($P < 0.01$). Genetic differentiation among populations, based on average pairwise Φ_{PT} (similar to F_{ST}), was 0.253. This result also indicated higher variation within populations.

Genetic Differentiation among Three Climatic Regions

We evaluated the level of genetic variation among the three different climatic regions (NE, NW, and S) at this structure level. This variation was estimated in three ways: from the number of observed alleles per subregion (N_A), from the effective number of alleles (N_E), and using *PIC* values (Table 4). As is shown in Table 4, the S region had the largest diversity values ($N_A = 4.65$, $N_E = 2.00$, $PIC = 0.401$) compared to the NE ($N_A = 4.05$, $N_E = 1.90$, $PIC = 0.374$) and NW regions ($N_A = 3.45$, $N_E = 1.80$, $PIC = 0.325$). Among the three climatic regions, *PIC* values ranged from 0.325 to 0.401 with a mean of 0.384, indicating a moderately low allelic diversity in the *P. tomentosa* populations (Table 4). The negative Wright's fixation index (F_{IS}), which is related to an excess of heterozygotes within subregions, was consistent, with H_O being higher than H_E in each region (Table 4). In addition, the number of private alleles was relatively low in the NW and NE regions (two and five, respectively), and higher in the S region (11).

Genetic and Geographical Relatedness

With 20 SSR markers, pairwise F_{ST} comparisons among nine populations of *P. tomentosa* were conducted; the intervals ranged from –0.0061 to 0.1159, and the results were significantly different from zero after a Bonferroni-type adjustment (Rice 1989). Four of the five highest F_{ST} values were associated with the SAX population, which may have been affected by genetic drift because this population size is small, containing 23 individuals (Supplementary Table 5 online). The geographic distance matrices ranged from 200.18 to 1183.70 km, based on latitude and longitude values (details not shown). Rousset's genetic distance values (F_{ST} /1 – F_{ST}) indicated that the most closely related *P. tomentosa* populations were Jiangsu (JS) and Gansu (GS), even though the geographical distance between them (949.05 km) was not the closest. The greatest genetic distance was between Shanxi (SX) and Shaanxi (SAX), even though these two populations are not the farthest apart (479.62 km). The greatest geographic distance was between the Beijing (BJ) and Gansu (GS) populations, although this pairing did not have the largest Rousset's

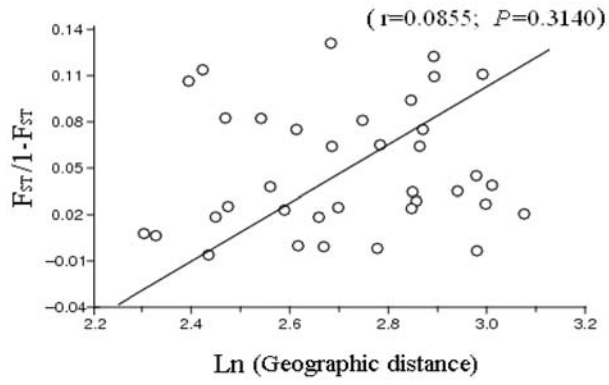

Figure 3. Relationship between genetic distance and geographic distance for *P. tomentosa* populations. Genetic distance is represented by pairwise $F_S/(1 - F_{ST})$ estimates among populations (Rousset 1997), which is regressed against the natural logarithm of the geographic distance (Ln distance) to test for isolation by distance. The RMA regression line (the solid line) overlays the scatterplot.

distance (0.0206). The shortest geographic distance was between Beijing (BJ) and Hebei (HB), but these two populations were not the most genetically related (0.0079). No correlation was found between genetic distance and geographic distance in *P. tomentosa* (r = 0.0855, P = 0.3140), based on a Mantel test of pairwise $F_{ST}/(1 - F_{ST})$ estimates among populations and the logarithm of geographic distance (Ln distance; Figure 3). This result provides further evidence that geographic distance is not the principal factor influencing genetic differentiation in *P. tomentosa*.

Discussion

Evaluation of SSR Markers and the Hybridization Origin of *P. tomentosa*

To lay the foundation for association analyses and MAS breeding programs, we developed 20 SSR markers and used them to evaluate genetic diversity and population structure in *P. tomentosa*. Of these 20 newly developed markers, 16 (80%) could be used to analyze species in different sections within the genus *Populus*. Our SSR markers had a higher cross-amplification rate within the genus *Populus* than SSR markers (72%) that were derived from *P. trichocarpa* by Yin et al. (2009). Microsatellite primer sequences are often conserved across related species and are broadly used to develop novel markers among species within the same genus or related species (Du et al. 2010). In the present study, we found a mean number of 4.95 observed alleles per locus within the 460 genotypes of *P. tomentosa*; this appreciable number of alleles may be related to the relatively large sample size and/or the co-dominant SSR markers. This value was higher than the level of N_A (4.3) in this *P. tomentosa* population that was determined using 15 microsatellite markers from cellulose synthase genes (Du et al. 2012) and higher than values in *Populus angustifolia* and *P. tremuloides*. However, 4.95 alleles per locus is lower than N_A values that have been reported for some related *Populus* species that were identified using SSR markers, such as *P. nigra*, *P. trichocarpa*, and *P. tremuloides* (see the review by Slavov and Zhelev 2010), as well as *P. euphratica* in NW China (N_A = 12.125; Wang et al. 2011). Outcrossing due to self-incompatibility is prevalent in this species (Kang et al. 1999), which could explain the considerable level of polymorphism in the microsatellite loci used in this study.

Expected heterozygosity is an important measurement of gene diversity (Slatkin and Barton 1989). In our study, the mean observed (H_O) and expected (H_E) levels of heterozygosity were 0.572 and 0.446, respectively. Similar values were also reported for two other poplar species: *P. tremuloides* (H_O = 0.47, H_E = 0.67; Namroud et al. 2005) and *P. tremula* (H_O = 0.47, H_E = 0.50; Lexer et al. 2005) when using SSR markers. However, higher levels of polymorphism have been reported for *P. nigra* (H_O = 0.70, H_E = 0.73; Rathmacher et al. 2010) and *P. euphratica* (H_O = 0.932, H_E = 0.787; Wang et al. 2011). H_E was lower than H_O at 14 SSR loci, indicating a significant excess of heterozygotes at these loci. However, H_E was higher than H_O at the other six SSR loci, which makes sense if the 20 markers are distributed at silence or hot spot regions throughout the whole genome. Excess heterozygotes were also observed at the population level. The S region showed the richest gene diversity, followed by the NE region, while the NW region showed the lowest level of gene diversity. Among the three climatic regions, the S region had a mean F_{IS} value of –0.102 (P = 0.0012), suggesting a high frequency of heterozygous loci in this region. Similarly, excess heterozygotes were found in the NE (F_{IS} = –0.308; P = 0.0010) and NW regions (F_{IS} = –0.323; P = 0.0001). These results suggest that excess heterozygotes across the entire natural distribution range could to be attributable to hybridization origins among multiple related species over a long period at different geographic locations, rather than from the persistence of one species at a particular zone for a short period of time (Li et al. 1997; Kang et al. 1999; He 2005).

Mendelian inheritance testing of all SSR loci based on segregation analyses of progeny arrays was used when a valid biological inference could not be drawn from the F_{IS} values. In this study, Mendelian inheritance testing suggested that excess heterozygotes were generated in this species. In 1986, to carry out the program "Chinese white poplar short-rotation industrial timber breeding improvement," this *P. tomentosa* resource pool was examined by the Institute of Chinese White Poplar (Beijing Forestry University, Beijing, China); they did not found polyploidy in any of the 1047 individuals examined (no paper has been published). Therefore, the excess number of heterozygotes cannot be explained by polyploidy or the amplification of some loci by specific primers, further confirming the hypothesis related to hybridization origins (Barnes 1967; Li et al. 1997; Kang et al. 1999; He 2005). The bottleneck effect is another reasonable cause of excess heterozygosity in this species; it results in reduced

allelic diversity and excess heterozygosity (Cornuet and Luikart 1996; Piry et al. 1999). In addition, positive selection or a heterozygote advantage in this species during its long evolutionary history should be considered (Nielsen 2005).

Climatic Region and Population Diversity

Huang (1992a) was the first to provide information about climatic regionalization in the distribution zones of *P. tomentosa*. Subsequently, Huang (1992b) reported geographical ecotypic zones, based on clonal growth variation and geographical factors. The three geographical ecotypic zones (central, suitable, and outlying regions) correspond well with the three climatic regions (S, NW, and NE regions, see Table 1 and Figure 1). Each region that included different populations was determined; several ecotypes are distributed in different regions, indicating that geographic and climatic factors play a key role in the different ecotypic patterns. Meanwhile, a considerable level of gene exchange may exist within regions, but limited gene flow could occur among regions (Huang 1992a, 1992b). The STRUCTURE model accounts for the presence of HWE or LD by introducing population structure and attempts to find population groupings that are not in disequilibrium (Pritchard et al. 2000). Thus, the subsets of populations in the structure analysis are in accordance with HWE. The pattern of individual assignment into the subsets ($K = 3$) (Supplementary Figure 2 online) generally agreed with the selection of clones from three different climatic regions (Table 1), indicating that each region can preserve a dynamic balance between gene frequency and genotype frequency. Therefore, the view that climatic zones can be treated as genetic regions in analyses of population differentiation was demonstrated. This was a crucial conclusion because it can guide genetic germplasm improvement and regional breeding strategies.

Generally, outcrossing of species produces a significant level of intrapopulation genetic variation (Narasimhamoorthy et al. 2008); our results indicate that *P. tomentosa* has this characteristic. Using SSR markers, the mean value of F_{ST} was 0.0528, which represents a low level of population differentiation; this was also confirmed by the AMOVA, which indicated that only 14.6% of the total molecular variance was attributable to between-population diversity (Table 3). The Mantel test also indicated that in the sampled populations, genetic variation was not correlated with geographic isolation (Figure 3 and Supplementary Table 5 online). Higher levels of total genetic variance were also reported at the intrapopulation level in this species using dominant AFLP markers; 75.23% of the genetic variation was explained by individual differences within populations and 24.77% was attributable to interpopulation differences (He 2005). Greater levels of total genetic variance were also detected at the intrapopulation level based on phenotypic characters (Huang 1992b; Li et al. 1996). Similarly, low levels of differentiation have been reported for congeneric species when using SSR markers, for example, as seen in the mean F_{ST} values for well-known species such as *P. tremula* (F_{ST} = 0.017/0.015; Lexer et al. 2005; Hall et al. 2007) and *P. trichocarpa* (F_{ST} = 0.078; Slavov and Zhelev 2010). These results provide valuable data for regional breeding; they may help minimize duplication and will assist in the establishment of core collections (López-Gartner et al. 2009).

The levels of genetic diversity within and among climatic regions that were found during the SSR marker analyses were consistent with levels in previous studies of geographical ecotype regionalization (Huang 1992b; Zhang et al. 1992a, 1992b; Li et al. 1996). Our results indicated that genetic variation was greater in the S region than in the NE and NW regions (Table 4). The NW region was located in arid and semiarid areas, with high altitude populations that experienced low temperatures and annual rainfall; during the past 100 years, precipitation in this zone has declined dramatically and most rivers have dried up (Huang 1992b; Zhang et al. 2003). Although *P. tomentosa* is able to adapt to different environmental conditions, the suitable ecological characteristics of this species are consistent with the S region. Drought, expansion of the desert, and low temperatures can impact growth and the fertility of pollen and seeds, leading to poor reproduction or the death of poplars at many sites; this could partially explain the divergent genetic diversities among the different climatic regions. The presence of private alleles in wild resources also gives us an opportunity to select useful recombinants in the field that showed greater adaptation and possibly accumulated genes for resistance to prevalent diseases. At a regional level, the genetic diversity in the S region was the highest, indicating that this region might represent the center of the current species distribution (Huang 1992b). This conclusion is also supported by more abundant phenotypic variations, such as leaf size and morphology, in the S region compared with the other two regions (Li et al. 1996). However, conjecture that the S climatic region is the central range of the origin of *P. tomentosa* was not supported. Complex analysis of the biogeography history, evolutionary forces, and geographic variables is required, which is beyond the scope of the present study.

Population Structure Analysis in *P. tomentosa*

Estimating population structure is a necessary first step in association analysis. This is important to avoid false positives or spurious associations and to constrain association studies in natural populations (Pritchard and Rosenberg 1999; Pritchard et al. 2000; Pritchard and Wen 2004; King et al. 2010).

To identify the true optimum number of subsets (K) in STRUCTURE, two models were chosen (Krutovsky et al. 2009; López-Gartner et al. 2009; D'hoop et al. 2010; Lanterbecq et al. 2010). In the first model described by Pritchard et al. (2000), the K value was obtained by comparing a probability $Pr(X/K)$ given the data (X) and an a value for each value of K. The K value that provides the maximum likelihood, called $LnP(D)$ in STRUCTURE, is generally considered the optimum number of subdivisions. However, Evanno et al. (2005) found that in many cases, the estimated $LnP(D)$ does not help determine the correct number of clusters (K). They recommended using an *ad hoc* statistic, ΔK, based on the rate of change in the log probability of data

between successive *K* values evaluated by STRUCTURE, which more accurately detects the real number of clusters.

In our study, although the population structure and migration scenarios of this species were complex, we applied the *ad hoc* statistic Δ*K* and found that it indeed improved the analysis of data generated by STRUCTURE. Examining the classified vertical lines for all individuals at $K = 11$ (Figure 2), we found that only the populations SD, Shanxi (SX), and Gansu (GS) could be unambiguously assigned to a particular cluster. Another six populations could be assigned to different subsets or combinations of a partial population, indicating that most *P. tomentosa* individuals had a mixed ancestry. Furthermore, the populations Anhui (AH) and Jiangsu (JS) were assigned to a particular subset, while the N portion of population Hebei (HB1) and population BJ were an admixture of independent subsets. The western area of population Henan (HN3) combined with the eastern range of population Shaanxi (SAX2). These results may be explained by the geographical origins of the samples, population size, and various topographies among or within different populations. For example, the populations Anhui (AH) and Jiangsu (JS) were assigned to a particular subset, possibly because of the small geographic sample size in this study. The Yellow River flows through the populations GS, SAX, and SX, and through the Taihang Mountains between populations HB and SX, which hinders genetic exchange within this species at different geographical provenances. In addition, elevation can play a major role in population structure. Different altitudes within a geographical zone would provide different degrees of selection pressure for adaptation and could accelerate population differentiation rather than maintain current gene flow (López-Gartner et al. 2009). Thus, different sampling locations at similar elevations without complex terrain pattern tended to cluster together. For example, Henan (HN3) was combined with the east range of population Shaanxi (SAX2), but the subset SAX1 (the marginal areas of the Kuan-chung Plain) was at a lower elevation than SAX2. The admixtures were from the same climatic region. Serving as a starting point for an association analysis, knowledge of population substructure is important as the basis for selecting samples that will be used in LD mapping. In the near future, based on genetic diversity, more diverse genotypes will be selected to design a LD mapping strategy to improve the exploitation of available *P. tomentosa* genetic resources.

Funding

State Key Basic Research Program of China (No. 2012CB114506), the Specific Program in Graduate Science and Technology Innovation of Beijing Forestry University (No. BLYJ201201), the Project of the National Natural Science Foundation of China (No. 31170622, 30872042), and the Introduction of Foreign Advanced Agricultural Science and Technology into China (No. 2009-4-22).

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

References

Barnes BV. 1967. Indications of possible mid-Cenozoic hybridization in aspens of the Columbia plateau. Rhodora. 69:70–81

Blair M, González L, Kimani P, Butare L. 2010. Genetic diversity, inter-gene pool introgression and nutritional quality of common beans (*Phaseolus vulgaris* L.) from Central Africa. Theor Appl Genet. 121:237–248.

Cornuet JM, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014.

D'hoop BB, Paulo MJ, Kowitwanich K, Sengers M, Visser RGF, van Eck HJ, van Eeuwijk FA. 2010. Population structure and linkage disequilibrium unravelled in tetraploid potato. Theor Appl Genet. 121:1151–1170.

Du QZ, Zhang DQ, Li BL. 2012. Development of 15 novel microsatellite markers from cellulose synthase genes in *Populus tomentosa* (Salicaceae). Am J Bot. 99:e46–e48.

Du XY, Chen ZW, Li W, Tan YQ, Lu J, Zhu XD, Zhao TY, Dong G, Zeng L. 2010. Development of novel microsatellite DNA markers by cross-amplification and analysis of genetic variation in gerbils. J Hered. 101:710–716.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 14:2611–2620.

Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. 2007. Adaptive population differentiation in phenology across a latitudinal gradient in European Aspen (*Populus tremula*, L.): A comparison of neutral markers, candidate genes and phenotypic traits. Evolution. 61:2849–2860.

He CZ. 2005. Study on genetic diversity and origin of *Populus tomentosa* Carr [Ph.D Thesis]. [China]: Beijing Forestry University.

Huang ZH. 1992a. The study on the climatic regionalization of the distributional region of *Populus tomentosa*. J Beijing Forestry Univ. 14:26–32.

Huang, ZH. 1992b. The study on the clonal geographical variation of *Populus tomentosa*. J Beijing Forestry Univ. 14:33–42.

Jensen JL, Bohonak AJ, Kelley ST. 2005. Isolation by distance, web service. BMC Genet. 6:13. doi:10.1186/1471-2156-6-13.

Kang X, Zhu Z, Zhang Z. 1999. Cytogenetic studies on the origin of Chinese white poplar. J Beijing Forestry Univ. 21:6–11.

King R, Harris S, Karp A, Barker J. 2010. Characterisation and inheritance of nuclear microsatellite loci for use in population studies of the allotetraploid *Salix alba–Salix fragilis* complex. Tree Genet Genomes. 6:247–258.

Krutovsky KV, Clair JBS, Saich R, Hipkins VD, Neale DB. 2009. Estimation of population structure in coastal Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco var. *menziesii*] using allozyme and microsatellite markers. Tree Genet Genomes. 5:641–658.

Lanterbecq D, Glaberman S, Vitousek MN, Steinfartz S, Benavides E, Wikelski M, Caccone A. 2010. Genetic differentiation between marine iguanas from different breeding sites on the island of Santa Fe (Galápagos Archipelago). J Hered. 101:663–675.

Lexer C, Fay MF, Joseph A, Nica MS, Heinze B. 2005. Barrier to gene flow between two ecologically divergent Populus species, *P .alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. Mol Ecol. 14:1045–1057.

López-Gartner G, Cortina H, McCouch S, Moncada M. 2009. Analysis of genetic structure in a sample of coffee (*Coffea arabica* L.) using fluorescent SSR markers. Tree Genet Genomes. 5:435–446.

Li K, Huang M, Wang M. 1997. Study on origin of *Populus tomentosa* Carr. using RAPD. Acta Phytotaxonomica Sinica. 35:24–31.

Li XG, Zhu ZT. 1996. Genetic diversity study on morphological characteristics of plus tree gene resources of *Populus tomentosa*. J Jilin Forestry Univ. 12:14–18.

Mantel N. 1967. Ranking procedures for arbitrarily restricted observation. Biometrics. 23:65–78.

Namroud MC, Park A, Tremblay F, Bergeron Y. 2005. Clonal and spatial genetic structures of aspen (*Populus tremuloides* Michx). Mol Ecol. 14:2969–2980.

Narasimhamoorthy B, Saha M, Swaller T, Bouton J. 2008. Genetic diversity in switchgrass collections assessed by EST-SSR markers. BioEnergy Res. 1:136–146.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.

Piry S, Luikart G, Cornuet JM. 1999. BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. J Hered. 90:502–503.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 65:220–228.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Pritchard JK, Wen W. 2004. Documentation for structure software: version 2. Available from: http://pritch.bsd.uchicago.edu/software/readme_structure2_1.pdf.

Rahman MH, Dayanandan S, Rajora OP. 2000. Microsatellite DNA markers in *Populus tremuloides*. Genome. 43:293–297.

Rathmacher G, Niggemann M, Kohnen M, Ziegenhagen B, Bialozyt R. 2010. Short-distance gene flow in *Populus nigra* L. accounts for small-scale spatial genetic structures: implications for in situ conservation measures. Conserv Genet. 11:1327–1338.

Raymond M, Rousset F. 1995. GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. J Hered. 86:248–249.

Rice WR. 1989. Analyzing tables of statistical tests. Evolution. 43:223–225.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics. 145:1219–1228.

Slatkin M, Barton NH. 1989. A comparison of three indirect methods for estimating average levels of gene flow. Evolution. 43:1349–1368.

Slavov GT, Zhelev P. 2010. Salient biological features, systematics, and genetic variation of *Populus*. In: Jansson S, Bhalerao R, Groover AT, editors. Genetics and genomics of *Populus*. NY: Springer. p. 15–38.

Temnykh S, Declerck G, Lukashova A, Lipovich L, Cartinhpur S, Mccouch S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11:1441–1452.

van Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P. 2004. Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 4:535–538.

Wang J, Li Z, Guo Q, Ren G, Wu Y. 2011. Genetic variation within and between populations of a desert poplar (*Populus euphratica*) revealed by SSR markers. Ann Forest Sci. 68:1143–1149.

Wright S. 1969. Evolution and the genetics of populations. The theory of gene frequencies. Chicago: University of Chicago Press.

Yeh FC, Yang RC, Boyle T. 1999. POPGENE version 1.32: Microsoft Windows-based freeware for population genetic analysis, quick user guide. Canada: Center for International Forestry Research, University of Alberta.

Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR, Tuskan GA. 2009. Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. New Phytol. 181:498–503.

Zhang DQ, Zhang ZY, Yang K. 2007. Identification of AFLP markers associated with embryonic root development in *Populus tomentosa*. Silvae Genet. 56:27–32.

Zhang H, Wu JW, Zheng QH, Yu YJ. 2003. A preliminary study of oasis evolution in the Tarim Basin, Xinjiang, China. J Arid Envir. 55:545–553.

Zhang Z, Huang Z, Zhang D. 1992a. The study on the clonal flowering and seeding of *Populus tomentosa*. J Beijing Forestry Univ. 14:43–50.

Zhang Z, Zhu Z, Müller-Starck G, Hattemer H. 1992b. Studying isozymes gene markers in clones of *Populus tomentosa* Carr. J Beijing Forestry Univ. 3:9–17.

Zhu Z, Zhang Z. 1997. The status and advances of genetic improvement of *Populus tomentosa* Carr. J Beijing Forestry Univ. 6:1–7.

Received Febuary 25, 2012; Revised June 21, 2012; Accepted July 3, 2012

Corresponding Editor: Dr. Brian Murray