SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

https://doi.org/10.1007/s11427-018-9455-2

De novo assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China

Yan-Jing Liu[1,2], Xiao-Ru Wang[3] & Qing-Yin Zeng[1,2*]

[1]*State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing 100091, China;*
[2]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;*
[3]*Department of Ecology and Environmental Science, UPSC, Umeå University, SE-90187, Umeå, Sweden*

Received November 25, 2018; accepted December 2, 2018; published online January 17, 2019

The white poplar (*Populus alba*) is widely distributed in Central Asia and Europe. There are natural populations of white poplar in Irtysh River basin in China. It also can be cultivated and grown well in northern China. In this study, we sequenced the genome of *P. alba* by single-molecule real-time technology. *De novo* assembly of *P. alba* had a genome size of 415.99 Mb with a contig N50 of 1.18 Mb. A total of 32,963 protein-coding genes were identified. 45.16% of the genome was annotated as repetitive elements. Genome evolution analysis revealed that divergence between *P. alba* and *Populus trichocarpa* (black cottonwood) occurred ~5.0 Mya (3.0, 7.1). Fourfold synonymous third-codon transversion (4DTV) and synonymous substitution rate (k_s) distributions supported the occurrence of the salicoid WGD event (~ 65 Mya). Twelve natural populations of *P. alba* in the Irtysh River basin in China were sequenced to explore the genetic diversity. Average pooled heterozygosity value of *P. alba* populations was 0.170±0.014, which was lower than that in Italy (0.271±0.051) and Hungary (0.264±0.054). Tajima's *D* values showed a negative distribution, which might signify an excess of low frequency polymorphisms and a bottleneck with later expansion of *P. alba* populations examined.

***Populus alba*, *de novo* assembly, genetic diversity, population expansion**

Citation: Liu, Y.J., Wang, X.R., and Zeng, Q.Y. (2019). *De novo* assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China. Sci China Life Sci 62, https://doi.org/10.1007/s11427-018-9455-2

INTRODUCTION

The white poplar (*Populus alba*), commonly called abele or silver poplar, is a deciduous tree belonging to the family Salicaceae. This taxon is widely distributed in Europe, and also presents in Central Asia (Brundu et al., 2008). As a fast-growing tree and a good biomass product, white poplar is included in the European Program of Forest Genetic Resources (EUFORGEN, 1999). It is resistant to insect pests, fungal and bacterial pathogens, and can tolerate diverse environmental stresses such as drought, wind, salinity and low temperatures (Brundu et al., 2008). Natural populations of white poplar are widely distributed in the northwest of China, mainly in the Irtysh River basin (Fang et al., 1999). The Irtysh River basin is under a temperate continental climate characterized by a seasonal climate with large changes in temperature during one year (Table S1 in Supporting Information). The white poplars can withstand freeze damage and survive at temperatures as low as –20°C in this region (Table S1 in Supporting Information).

The genus *Populus* has six sections: *Tacamahaca*, *Turanga*, *Populus*, *Leucoides*, *Aigeiros*, and *Abaso* (Hamzeh and Dayanandan, 2004). The sequenced genomes of *Populus trichocarpa* (black cottonwood) and *Populus euphratica* belong to sect. *Tacamahaca* and *Turanga*, respectively (Argus et al., 2010; Fang et al., 1999; Ma et al., 2013; Tuskan et al., 2006). The *P. trichocarpa* is a forest tree species that is

*Corresponding author (email: qingyin.zeng@caf.ac.cn)

native to western North America and grows at elevations up to 2,600 m, such as the west of the Rocky Mountains (Argus et al., 2010). The *P. euphratica* is a desert poplar species that adapts to salt stress, extreme temperatures, and drought conditions in the natural desert forest ecosystems of China and Middle Eastern countries (Fang et al., 1999; Ferreira et al., 2006). Both of these two species are not suitable for planting in the Chinese plain, while the white poplar can be cultivated and grown well in North China (Fang et al., 1999). Recently, genomes of two aspen species, *Populus tremuloides* and *Populus tremula*, were assembled (Lin et al., 2018). Along with *P. alba*, these three species all belong to sect. *Populus* (Argus et al., 2010). *P. tremuloides* is a major tree species of North America that is widely distributed throughout cold and cool-temperate regions from coast to coast and from within the Arctic Circle to the north rim of the Valley of Mexico. It hybridizes with *P. alba* (*P.* × *heimburgeri* B. Boivin) in southeastern Canada and the northeastern United States (Argus et al., 2010). *P. tremula* is native to cool temperate regions of Europe and Asia. Hybridizations between *P. alba* and *P. tremula* in natural populations are very common and frequent (Lexer et al., 2005). These two Eurasian species were temporarily separated demographic histories after the last glacial maximum (Christe et al., 2016; Christe et al., 2017; Fussi et al., 2010). They are genetically and ecologically divergent (flood plain species vs. upland pioneer) and have incomplete reproductive barriers (Lexer et al., 2005; Stölting et al., 2013). Assembly of the *P. alba* genome will not only provide more information on gene and genome evolution, but also facilitate the study of how evolutionary processes affect speciation of *Populus* species. Here, we *de novo* assembled the genome of *P. alba*. Its natural populations are widely distributed, and retain a large amount of natural variation. We then surveyed the genetic diversity of *P. alba* natural populations in Irtysh River basin in China. The results indicated that the *P. alba* populations examined might have experienced an expansion. This genome will establish a new model poplar species and facilitate the poplar functional genomics research.

RESULTS AND DISCUSSION

Genome sequencing and assembly

This study applied the Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing technology to sequence and *de novo* assemble the genome of *P. alba*. The genome size was estimated to be approximately 495.96±0.03 Mb by flow cytometry. Contigs were assembled from ~130.17× (54.15 Gb) PacBio reads. Three rounds of error correction were performed using whole-genome shotgun sequences (Table S2 in Supporting Information). 1,285 contigs with a total length of 415.99 Mb were assembled, covering ~83.86% of the genome size (Table 1). This assembly had a contig N50 length of 1.18 Mb. The longest contig was 7.09 Mb.

To assess the completeness of the genome, the full-length transcriptome of *P. alba* plantlets and core eukaryotic genes dataset were compared with the draft *P. alba* genome. Using PacBio full-length cDNA sequencing, 97,950 isoforms were obtained accounting for 214.05 Mb sequences from three different size fractions (1–2 kb, 2–3 kb, and 3–6 kb; Table S3 in Supporting Information). 99.99% (97,943 isoforms) of all these isoforms could be mapped to the draft *P. alba* genome. Among all the mapped isoforms, 99.30% (97,255 isoforms) had sequence identities higher than 95% with the draft *P. alba* genome. The core eukaryotic gene coverage of draft *P. alba* genome was estimated to be ~95.56% by CEGMA software (Table S4 in Supporting Information), and ~96.52% by BUSCO software, which was comparable with other *Populus* or *Salix* genomes (Figure S1 in Supporting Information). These assessments were slightly higher than that of *Populus tremula* and *Populus tremuloides* (Lin et al., 2018). In these two aspen genomes, 93.55% and 91.94% of the CEGMA genes were completely identified in the *P. tremula* and *P. tremuloides* genomes, respectively. And 95% of BUSCO genes were completely recovered in *P. tremula* and *P. tremuloides* genome assembly. All of these results support that the assembly of *P. alba* has a high coverage.

Genome annotation

To identify the repetitive elements of draft *P. alba* genome, this study used a pipeline, which consisted of both homology-based searches and *de novo* annotation. A total of 188.30 Mb sequence in draft *P. alba* genome were annotated as repetitive elements (Table 2). The proportion of the repetitive elements in draft *P. alba* genome was 45.16%, which was slightly lower than that of *P. trichocarpa* genome (48.07%). This lower proportion was mainly due to the lower ratio of DNA transposons in draft *P. alba* genome (9.91% of *P. alba* vs. 12.31% of *P. trichocarpa*). This tiny difference in total abundance was also found in comparison between other aspens and *P. tricocarpa*. The amounts of repetitive elements in *P. tremula* and *P. tremuloides* genomes were both lower than that in *P. tricocarpa* genome (Lin et al., 2018). Overall, the components of repetitive elements in *P. alba* (Figure 1) were similar to that in *P. trichocarpa*, *Populus euphratica*, and *Sallix suchowensis* genome (Table 2).

Protein-coding gene models were predicted by combining homology-based, *de novo* and RNA-Seq based methods. A total of 32,963 protein-coding genes were identified (Table S5 in Supporting Information). Among these identified genes, 32,288 (97.95%) had homologues in other plants (Table S5 in Supporting Information), 31,384 (95.21%) could be functionally annotated, and 14,981 (45.45%) were

Table 1 Statistics for the assembly of *P. alba*[a)]

	Size (bp)	Counts	Content (%)
Total counts of contig sequences		1,285	
Total length of contig sequences	415,990,855		
Largest contig length	7,086,944		
Smallest contig length	4,992		
Average sequence length	323,728.3		
Median sequence length	75,011		
Contig N50*	1,180,906	101	
Contig N90*	144,820	473	
GC content			33.76
N content			0

a) *, N50 and N90 refer to the size above which 50% and 90% of the total length of the sequence assembly can be found.

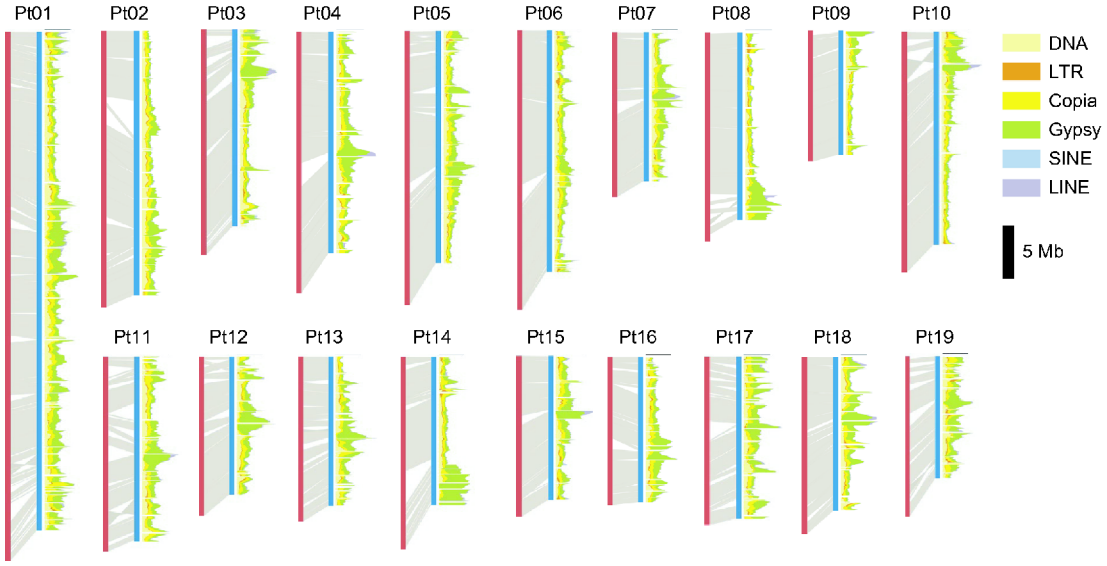

Figure 1 Collinearity between *P. alba* (blue) and *P. trichocarpa* (red) genomes. Orthologous relationships between 23,708 protein-coding *P. alba* genes and 23,091 *P. trichocarpa* genes arranging in 460 collinear blocks are linked by grey lines. The proportion of the main *P. alba* repeat elements (colored) were calculated in 400 kb windows sliding in 50 kb steps.

found to have splice variants (21,328 predictions in total; Table S6 in Supporting Information). The average CDS length, exon length, intron length, and exon number of *P. alba* genes were 1,273.15 bp, 235.11 bp, 462.22 bp, and 5.42, respectively, while those of *P. trichocarpa* genes were 1,108.50 bp, 332.39 bp, 393.16 bp, and 4.85, respectively. Paired-samples *t* test (2-tailed) showed that *P. alba* genes had significant difference with *P. trichocarpa* ($P \leq 0.036$) in coding sequence (CDS) length, exon length, and exon number (Figure S2 in Supporting Information). These differences might be caused by the smaller predicted transcriptome of *P. alba* than that of *P. trichocarpa* (42,950 protein-coding genes in annotation version 3.1). However, *P. alba* genes did not show significant difference in CDS length, exon number, exon length, and intron length with *P. euphratica* (paired-samples *t* test, $P \geq 0.073$; Figure S2 in Supporting Information). GC content in genic regions of *P. alba* (44.03%) was significantly higher than the average of whole genome (34.96%), which was consistent with that of other plant genomes (Singh et al., 2016).

14,482 gene families (30,139 genes) in *P. alba* genome were identified by homologue clustering of genes from 12 plant species (Figure 2A). Besides these grouped genes, 2,395 *P. alba* genes were found to have no homologues within the 12 compared species, which were predicted to be orphan genes. Compared with *P. trichocarpa*, *P. euphratica*, *S. suchowensis*, and *A. thaliana*, 10,986 (75.86%) of 14,482 *P. alba* gene families were shared by these five species, whereas 110 families (contained 232 genes) were unique to *P. alba*. Gene Ontology (GO) enrichment analysis of these 110 unique families showed enrichment in copper ion binding ($P<10^{-10}$) (Table S7 in Supporting Information). Compared with *P. trichocarpa*, 1,150/2,466 families were expanded/contracted in *P. alba* (in comparison to ancestor).

Table 2 Classification of repetitive elements in the *P. alba*, *P. trichocarpa*, and *S. suchowensis* genome[a)]

Type*	*P. alba*		*P. trichocarpa*		*P. euphratica*		*S. suchowensis*	
	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
SINEs	881,048	0.21	1,168,914	0.28	209,288	0.04	1,006,145	0.33
LINEs	5,072,756	1.22	4,951,259	1.17	3,752,096	0.79	5,617,848	1.85
CR1	344,817	0.08	78,842	0.02	189,386	0.04	535,238	0.18
I	263,128	0.06	104,649	0.02	141,311	0.03	561,615	0.19
L1	4,352,351	1.04	4,348,060	1.03	2,951,464	0.62	3,147,889	1.04
L2	36,517	0.01	347,797	0.08	225,781	0.05	508,719	0.17
LTR elements	120,180,749	28.82	122,845,585	29.05	165,397,346	35.00	77,322,015	25.48
Caulimovirus	787,964	0.19	843,975	0.20	1,249,429	0.26	477,461	0.16
Copia	31,494,548	7.55	33,630,983	7.95	24,359,131	5.16	29,462,819	9.71
ERV	75,316	0.02	58,869	0.01	164,678	0.03	551,547	0.18
Gypsy	79,320,298	19.02	76,624,917	18.12	132,516,935	28.04	29,827,753	9.83
Pao	37,525	0.01	69,014	0.02	226,816	0.05	813,843	0.27
DNA	41,312,880	9.91	52,048,303	12.31	33,501,653	7.09	21,433,480	7.06
CMC-EnSpm	9,811,105	2.35	14,192,684	3.36	7,335,513	1.55	2,254,030	0.74
hAT-Ac	1,548,459	0.37	2,528,969	0.60	1,215,694	0.26	839,124	0.28
hAT-Tag1	2,520,929	0.60	2,735,641	0.65	1,025,376	0.22	1,035,574	0.34
hAT-Tip100	1,143,089	0.27	1,676,016	0.40	1,676,084	0.35	161,866	0.05
MuLE-MuDR	814,205	0.20	1,317,446	0.31	785,694	0.17	578,621	0.19
PIF-Harbinger	5,100,121	1.22	6,302,561	1.49	4,478,469	0.95	1,898,801	0.63
Tc1/Mariner (TcMar)	79,074	0.02	180,737	0.04	154,265	0.03	329,211	0.11
Helitron	17,380,007	4.17	19,013,358	4.50	14,751,218	3.12	12,120,634	3.99
Simple_repeat	17,030,335	4.08	16,961,853	4.01	21,794,530	4.61	10,878,248	3.59
Unknown	3,823,815	0.92	5,352,205	1.27	5,948,659	1.26	21,494,505	7.08
Total	188,301,583	45.16	203,328,119	48.07	230,603,572	48.80	137,752,241	45.40
Genome size	416,960,588[#]		422,943,459		472,531,041		303,415,476	

a) *, All repeat types were assigned according to homology to the Repbase database (http://www.girinst.org/repbase). The names of the main classes of repetitive elements are shown in bold. #, The mitochondrion and chloroplast sequences were counted.

5,088 genes in the 1,150 expanded families were remarkably enriched in terpene synthase activity, magnesium and zinc ion binding (Table S8 in Supporting Information). 6,236 genes in the 2,466 contracted families were enriched in recognition of pollen and multiple molecular functions, including ADP binding, extracellularly glutamate-gated ion channel activity, protein kinase activity, etc. (Table S9 in Supporting Information).

Genome evolution and phylogeny

To address the phylogenetic relationship of different *Populus* species, we reconstructed a phylogenetic tree based on a concatenated sequence alignment of 444 single-copy genes shared by *P. alba* and other 11 green plant species (Figure 2B). All the relationships were well supported with >99% bootstrap values. *P. alba* and *P. trichocarpa* were grouped together. The ancestor of these two species was sister of *P. euphratica*. The splice time of *P. alba* and *P. trichocarpa* was 5.0 (3.0, 7.1) million years ago (Mya), while the splice time of *P. euphratica* and the ancestor of *P. alba* and *P. trichocarpa* was 5.5 (3.4, 7.7) Mya. There were neither expanded nor contracted gene families in the ancestor of *P. alba* and *P. trichocarpa* (in comparison to its ancestor). One possibility was that the ancestor of *P. alba* and *P. trichocarpa* had evolved for a short time (0.5 Mya). Among internal nodes, the ancestor of the Salicaceae experienced the most significant gene family expansion. This dramatic change might be due to whole genome duplication (WGD) event occurring in the ancestor of Salicaceae, ~60 to 65 Mya (Tuskan et al., 2006). WGD event can lead to rapid expansion of gene families (Van de Peer et al., 2009a, 2009b). For instance, *Glycine* experienced a recent WGD event (~13 Mya) (Schmutz et al., 2010), which resulted in the rapid expansion of *Glycine max* gene families (Figure 2B).

Collinearity analysis between *P. alba* and *P. trichocarpa* genomes identified 3,098 syntenic blocks, covering 77.1% (25,423) of *P. alba* genes and 61.6% (26,437) of *P. trichocarpa* genes (Figure 1). This extensive collinearity indicated that most of genomic regions were conserved after speciation. Collinearity analysis of inter- and intra- *Populus* genomes provided evidence of the salicoid WGD event (Figure 3A and B). The sharp peaks in fourfold synonymous third-codon transversion (4DTV≈0.08) values and synonymous substitution rate (k_s≈0.25) values represented a burst of gene duplication, corresponding to the salicoid WGD event (~65

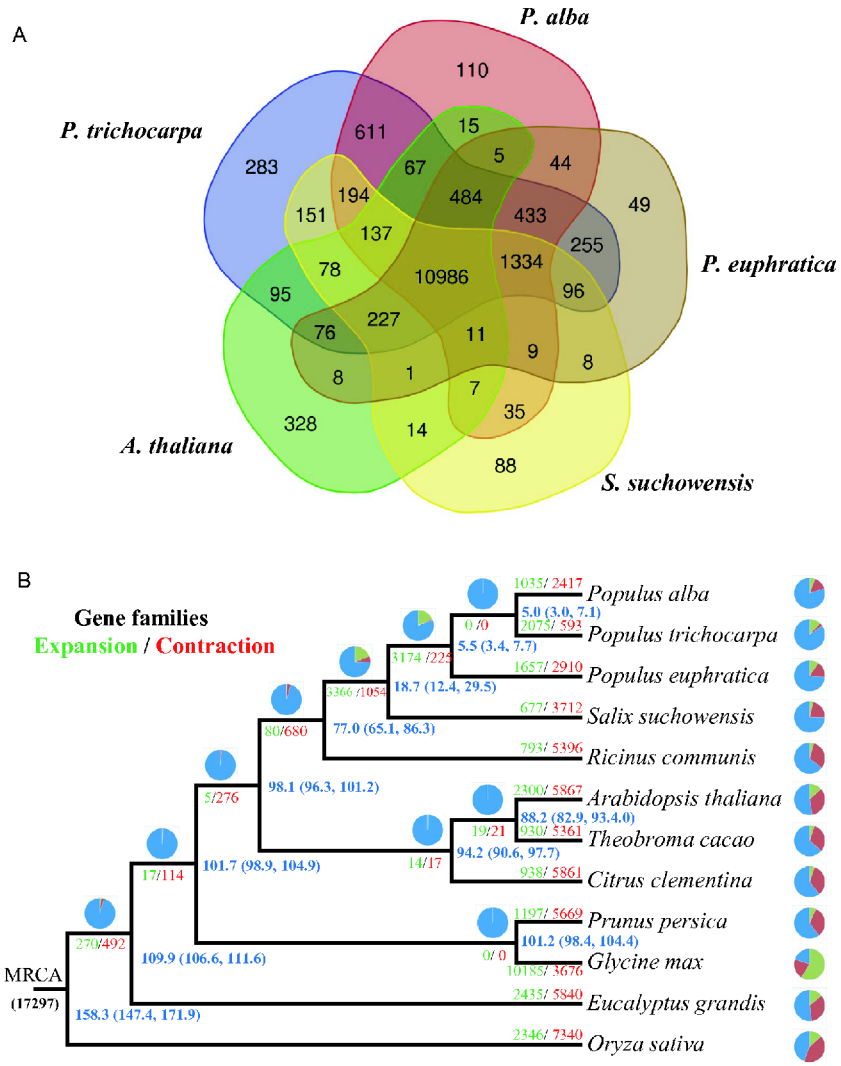

Figure 2 Evolution of *P. alba* and comparative genomic analysis. A, Comparison of the number of gene families in *P. alba*, *P. trichocarpa*, *P. euphratica*, *S. suchowensis*, and *A. thaliana*. B, Phylogenetic tree and number of gene families displaying expansion (green) and contraction (red) among 12 plant species. The pie charts show the expansion (green), contraction (red) and conserved (blue) gene family proportions among all gene families. Estimated divergence time (in millions of years) is denoted at each internal node in blue. MRCA, most recent common ancestor.

Mya) (Ma et al., 2013; Tuskan et al., 2006). Distribution of 4DTV values suggested that divergence between *P. alba* and *P. trichocarpa* occurred ~4.3 Mya (4DTV, ~0.005), which was similar to the result of phylogenetic analysis (~5.0 Mya; Figure 2B). This estimation was smaller than the divergence time between *P. trichocarpa* and *P. tremula*, which was dated at ~15 Mya (Lin et al., 2018). It indicated that the *P. alba* and *P. tremula* had experienced long-term divergence and speciation.

Genetic diversity of white poplar populations in Irtysh River basin in China

White poplar is considered native to the continental Mediterranean basin and widely distributed throughout the floodplains of northern Africa, southern Europe, and Central Asia (Brundu et al., 2008; Roiron et al., 2004). In the Irtysh River basin in China, *P. alba* has adapted to the local climates and formed natural populations. To survey the patterns of genetic variation, 12 *P. alba* populations from upstream of Irtysh River in China were selected for pooled whole-genome sequencing (WGS) (Figure 4, Table S1 in Supporting Information). Ten individuals were selected based on their geographic coordinates (minimum linear distance 500 m) to reduce the effect of clonal reproduction on allele frequency within each population.

Whole-genome resequencing of the pooled libraries resulted in a total of 109.82 Gb (790,835,846 reads) clean data (Table S2 in Supporting Information). 98.36% of the reads (777,870,039 reads) were accurately mapped against the *P. alba* genome. A total of 7,497,684 single nucleotide polymorphisms (SNPs) were identified from *P. alba* populations.

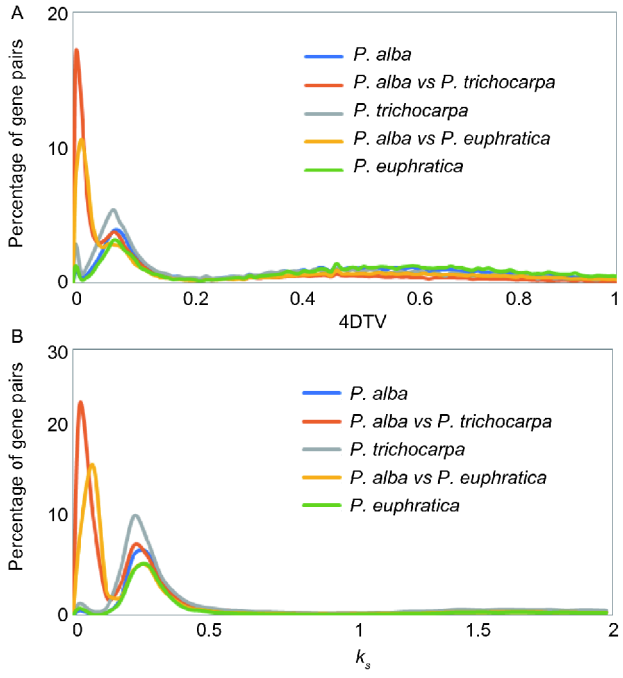

Figure 3 The 4DTV (A) and k_s distributions (B) for duplicated gene pairs in *Populus* species.

The average distance among SNP loci was 55 bp according to the assembly size of 415,990,855 base pairs. The higher SNP counts in studied populations than that in Italy and Hungary (1,775,768 SNPs) (Stölting et al., 2015) might be attributed to introgressions. On average, pooled heterozygosity value (±SD) of *P. alba* populations calculated from 263,417 1-kb non-overlapping sliding windows along the genome was 0.170±0.014. This was lower than the pooled heterozygosity values of *P. alba* populations in Italy (0.271 ±0.051) and Hungary (0.264±0.054) (Stölting et al., 2015).

To infer potential demographic events that could strongly influence genetic diversity within populations, we calculated Tajima's *D* using windowed measures. A genome-wide negative Tajima's *D* is indicative of an expansion after a bottleneck, whereas a positive *D* is compatible with a scenario of a decrease in population size (Biswas and Akey, 2006; Tajima, 1989). In studied populations, only windows with at least one SNP were used for the analyses. Watterson's Theta ranged from 0.0004 to 0.0622, and Tajima's *D* values showed a negative distribution, which might signify an excess of low frequency polymorphisms and a bottleneck with later expansion of studied *P. alba* populations (Figure 5).

MATERIALS AND METHODS

Plant materials and libraries preparation

For *de novo* genome sequencing, a single genotype of *Populus alba* cultivated in Institute of Botany, the Chinese Academy of Sciences (IBCAS), Beijing, China was tissue cultured. We established a regeneration system of this genotype. All genomic DNA were extracted from regenerated plantlets cultured on 1/2 Murashige and Skoog (MS) medium containing 0.5 mg L^{-1} IBA. Total RNA from roots, stems, and leaves of regenerated plantlets were extracted separately and then mixed equally for library preparation. PacBio libraries for *de novo* sequencing were constructed at Tianjin Biochip Corporation, Tianjin, China. For Illumina sequencing, genomic DNA was extracted using a DNeasy Plant Maxi Kit (Qiagen, Germany) and fragmented by M220 Focused-ultrasonicator (Covaris, Inc., USA) following the

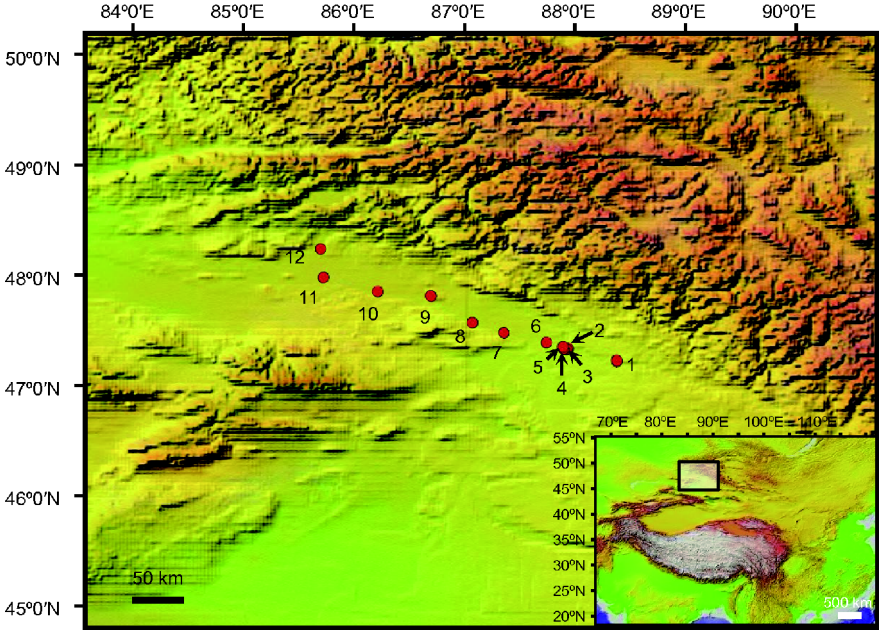

Figure 4 Geographic distribution of the 12 sampled populations of *P. alba*.

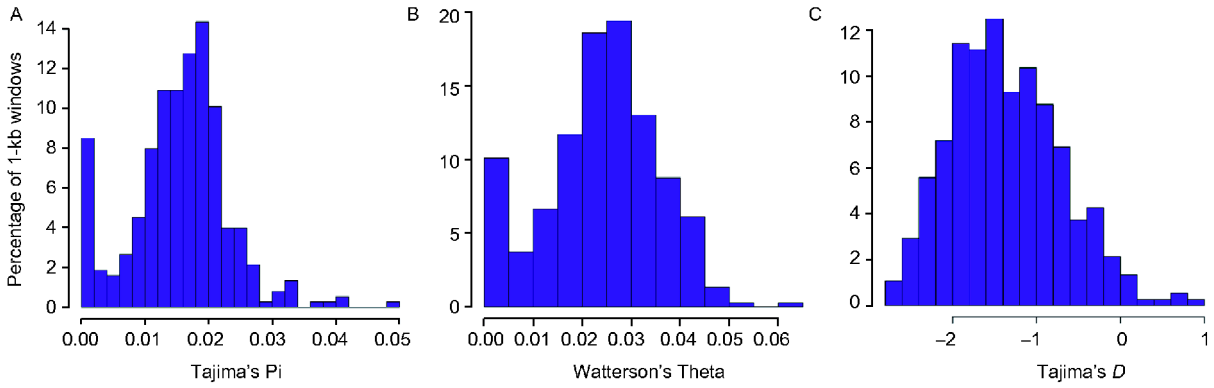

Figure 5 (Color online) Distributions of genomic parameters among *P. alba* populations. A, Tajima's Pi. B, Watterson's Theta. C, Tajima's *D* ($P<2.2\times10^{-16}$).

manufacturer's instructions. Short-insert paired-end (PE) (300–800 bp) DNA libraries were constructed using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (UK). Short-insert PE RNA library was prepared according to the manufacturer's instructions (Illumina, San Diego, California, USA).

Population-level whole-genome resequencing of *P. alba* was performed using a pool-Seq strategy. Populations were sampled from locations in the fields along Irtysh River basin in northwest of China (Figure 4, Table S1 in Supporting Information). Samples were randomly selected at linear distances over 500 m. Leaves of each sample were dried with silica gel and stored at –20°. Within one week after sampling, genomic DNA from each sample was extracted using a DNeasy Plant Maxi Kit (Qiagen, Germany) and fragmented by M220 Focused-ultrasonicator (Covaris, Inc., USA), separately. Quantified DNA of 10 individuals were mixed equally for each population. Short-insert paired-end (PE) (500–700 bp) DNA libraries were constructed using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (UK).

Genome sequencing and assembly

Long-read DNA and RNA (1–20 kb) sequencing data were generated by PacBio RSII Sequencer (Pacific Biosciences, USA). Three short-insert PE libraries (325, 600, 795 bp) were qualified by Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) and sequenced on Illumina Hi-Seq2500 sequencing platforms. The draft genome was first assembled by SMARTdenovo (https://github.com/ruanjue/smartdenovo) using long-read DNA sequencing data. Three rounds of point correction were applied using short-insert PE reads. PE reads were trimmed by ngsShoRT v2.2 (Chen et al., 2014) and mapped to the raw assembly with Burrows-Wheeler Aligner (BWA) v0.7.13 using the *bwa mem* algorithm (Li and Durbin, 2009). Errors were corrected using in-house script under the following criteria: base coverage of Illumina short reads was higher than 10 and the ratio of consensus sequence is higher than 0.8.

Assessment of the assembly completeness

To assess the completeness of the final genome assembly, both experimental and *in silico* evaluation were performed. First, three cDNA libraries of different sizes (1–2 kb, 2–3 kb, and 3–6 kb) were sequenced on PacBio RSII sequencer. Polished high-quality isoforms were mapped to the genome assembly using BLAT v.35 (Kent, 2002) to evaluate transcriptome coverage. Second, core eukaryotic genes were mapped to the genome assembly to calculate the gene region coverage by CEGMA v2.5 (Parra et al., 2007). Third, sets of Benchmarking Universal Single-Copy Orthologs from OrthoDB (www.orthodb.org) were employed by BUSCO v3 (Simão et al., 2015) to assess the completeness of *P. alba* genome assembly.

Repeat annotation

A combination of *de novo* and homology-based approaches was used to identify repetitive sequences, including tandem repeats and interspersed transposable elements (TEs). The tandem repeats were annotated using Tandem Repeats Finder (TRF) v4.09 (Benson, 1999). TEs were identified at both DNA and protein levels. At the DNA level, a *de novo* repeat library for *P. alba* was constructed using RepeatModeler v1.0.4 (http://www.repeatmasker.org/RepeatModeler/), LTR_finder v1.0.5 (Xu and Wang, 2007), and PILER (Edgar and Myers, 2005). This library contained the resulting consensus sequences and classification for each repeat family identified by the three programs. Both the *de novo* library and RepBase repeatmaskerlibraries-20160829 (Bao et al., 2015), which is the most commonly used repetitive DNA elements database, were employed to identify the transposable elements by RepeatMasker Open-4.0 (Smit, 2013–2015). At the protein level, genome sequences were searched against the TE protein database with RepeatProteinMask in

the RepeatMasker package. All of these predictions were combined to give a final annotation of the repetitive sequences in studied genomes.

Gene prediction and functional annotation

Protein-coding gene models were identified by homology-based and *de novo* prediction incorporating RNA sequencing data. RNA-seq data processed by Tophat v2.1.1 and cufflinks v2.2.1 (Trapnell et al., 2009; Trapnell et al., 2010) was used for transcriptome-based prediction. Homology-based detection used a collection of proteins from 11 plant species, including *P. trichocarpa*, *A. thaliana*, *O. sativa*, *G. max*, *Medicago truncatula*, *T. cacao*, *Citrullus lanatus*, *C. clementina*, *Malus domestica*, *P. persica*, and *Pyrus* × *bretschneideri*. *De novo* prediction was performed using GeneMark-ET and AUGUSTUS packaged in BRAKER1 (Hoff et al., 2016). Finally, a reference gene set was constructed by combining all above predictions using EVidenceModeler v1.1.1 (Haas et al., 2008).

Functions of the predicted gene models were assigned based on BLASTp searches against public databases, including NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/), Swiss-Prot (Bairoch and Apweiler, 2000), KEGG (Kyoto Encyclopedia of Genes and Genome) (Kanehisa et al., 2016), and KOGs (Eukaryote Clusters of Orthologous Groups of proteins, http://eggnog.embl.de/version_3.0/downloads.html). GO terms were determined using Blast2GO v3 (Götz et al., 2008). Gene Ontology (GO) enrichment analysis was performed using topGO (topology-based Gene Ontology scoring) (Alexa and Rahnenfuhrer, 2010).

Gene family evolution

For gene family analysis, protein-coding sequences of 11 green plant species were obtained from public databases. Data sets of nine of the 11 species were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html), including *Populus trichocarpa* v3.0 (Tuskan et al., 2006), *Ricinus communis* v0.1 (Chan et al., 2010), *Arabidopsis thaliana* TAIR10 (Lamesch et al., 2012), *Theobroma cacao* v1.1 (Motamayor et al., 2013), *Citrus clementina* v1.0 (Wu et al., 2014), *Prunus persica* v2.1 (Verde et al., 2013), *Glycine max* Wm82.a2.v1 (Schmutz et al., 2010), *Eucalyptus grandis* v2.0 (Myburg et al., 2014), and *Oryza sativa* v7_JGI (Ouyang et al., 2007). Data of *Populus euphratica* v1.0 (Ma et al., 2013) was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome/) genome database. Data of *Salix suchowensis* v1.0 (Dai et al., 2014) was obtained from the website of Nanjing Forestry University, Jiangsu, China (https://bio.njfu.edu.cn/home/index.php). Together with *P. alba*, protein-coding genes of 12 species were filtered to keep only primary transcript with open reading frame ≥150 bp and without any stop codon. After filtering, 429, 624, 500, 357, 412, 508, 152, 34, 106, 250, 6, and 56 truncated or small genes of *P. alba*, *P. trichocarpa*, *P. euphratica*, *S. suchowensis*, *R. communis*, *A. thaliana*, *T. cacao*, *C. clementina*, *P. persica*, *G. max*, *E. grandis*, and *O. sativa* were excluded from the analysis, respectively. Orthogroups among these 12 species were detected by an all-against-all BLASTp search of all retained proteins. Then, the OrthoFinder v1.1.4 (Emms and Kelly, 2015) was used to cluster the blast results into paralogous and orthologous groups. Evolution of gene families was analyzed with CAFÉ v3.1 (Computational Analysis of Gene Family Evolution) (Han et al., 2013). Single-copy orthologs were then extracted to reconstruct the phylogenetic tree of the 12 species using RAxML v8.0.20 (Stamatakis, 2014). Bayesian estimation of species divergence times was performed by MCMCTree program packaged in the Phylogenetic Analysis by Maximum Likelihood (PAML) v4.8 (Yang, 2007) based on the phylogenetic tree constructed.

Collinearity analysis

Pairwise collinearity analysis between species was applied using MCScanX (Wang et al., 2012). At least five genes are required to identify syntenic blocks. The values of synonymous (K_s), non-synonymous (K_a) substitution rates and the K_a/K_s ratio were estimated for each collinear gene pair using codeml in PAML v4.8. Transversion rate at 4DTV for each collinear gene pair was calculated based on 4-fold degenerate sites following the HKY substitution model.

Pool-Seq data for population

Libraries of *P. alba* populations were sequenced on illumine platform. Illumina 150 bp paired-end sequencing data for each population were trimmed by ngsShoRT v2.2 (ngsShoRT.pl -methods lqr_5adpt_tera_nsplit_rmHP_5end -lqs 20 -lq_p 5 -5a_f i-m -5a_mp 95 -5a_ins 0 -5a_del 0 -5a_axn ka -tera_avg 20 -nsplit_len 5 -rmHP_ml 8 -rmHP_bases a -min_rl 70 -n5 10) (Chen et al., 2014). The trimmed sequences were mapped to the *P. alba* genome using *bwa mem* and realigned with IndelRealigner in GATK v3.8.0 (DePristo et al., 2011; Van der Auwera et al., 2013) pipeline. The resulting bam-files were combined to final sync-files and processed to analyze allele frequency differences. Tajima's *D* was calculated using PoPoolation (Kofler et al., 2011) (PoPoolation: Variance-sliding.pl --measure D --pool-size 20 --min-count 2 --min-coverage 6 --max-coverage 100 --window-size 1000 --step-size 1000 --min-covered-fraction 0.5).

Accession numbers

Raw sequencing data for genome assembly and population

resequencing, the assembly and annotation of *P. alba* genome have been deposited in NCBI under BioProject accession number PRJNA491245.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We thank Dr. Jian Wang for assisting with the population sampling from Irtysh River basin. This work was supported by the National Science Fund for Distinguished Young Scholars (31425006) and Chinese Academy of Forestry (CAFYBB2018ZX001).*

References

Alexa, A., and Rahnenfuhrer, J. (2010). topGO: Enrichment Analysis for Gene Ontology. R package version 2.30.1.

Argus, G.W., Eckenwalder, J.E., Kiger, R.W. (2010). *Salicaceae*. In Flora of North America, Flora of North America Editorial Committee, ed. vol. 7. (New York: Oxford University Press).

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28, 45–48.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580.

Biswas, S., and Akey, J.M. (2006). Genomic insights into positive selection. Trends Genets 22, 437–446.

Brundu, G., Lupi, R., Zapelli, I., Fossati, T., Patrignani, G., Camarda, I., Sala, F., and Castiglione, S. (2008). The origin of clonal diversity and structure of *Populus alba* in Sardinia: evidence from nuclear and plastid microsatellite markers. Ann Bot 102, 997–1006.

Chan, A.P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K.M., Redman, J., Chen, G., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol 28, 951–956.

Chen, C., Khaleel, S.S., Huang, H., and Wu, C.H. (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med 9, 8.

Christe, C., Stölting, K.N., Bresadola, L., Fussi, B., Heinze, B., Wegmann, D., and Lexer, C. (2016). Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F_1 fertility and recurrent gene flow. Mol Ecol 25, 2482–2498.

Christe, C., Stölting, K.N., Paris, M., Fraïsse, C., Bierne, N., and Lexer, C. (2017). Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. Mol Ecol 26, 59–76.

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G.A., Milne, R., Chen, Y., Wan, Z., Wang, Z., et al. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res 24, 1274–1277.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498.

Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics 21, i152–i158.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16, 157.

EUFORGEN. (1999). *Populus nigra* network: Report of the fifth meeting..

Fang, C., Zhao, S., Skvortsov, A. (1999). Salicaceae. In Flora of China, Z. Y. Wu, P.H. Raven, D.Y. Hong, ed. vol. 4. (Beijing: Science Press; St. Louis, MO: Missouri Botanical Garden Press).

Ferreira, S., Hjernø, K., Larsen, M., Wingsle, G., Larsen, P., Fey, S., Roepstorff, P., and Salomé Pais, M. (2006). Proteome profiling of *Populus euphratica* Oliv. upon heat stress. Ann Bot 98, 361–377.

Fussi, B., Lexer, C., and Heinze, B. (2010). Phylogeography of *Populus alba* (L.) and *Populus tremula* (L.) in Central Europe: secondary contact and hybridisation during recolonisation from disconnected refugia. Tree Genets Genomes 6, 439–450.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36, 3420–3435.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9, R7.

Hamzeh, M., and Dayanandan, S. (2004). Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast TRNT-TRNF region and nuclear rDNA. Am J Bot 91, 1398–1408.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 30, 1987–1997.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32, 767–769.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M.T., Grimwood, J., Cattonaro, F., et al. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45, 487–494.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44, D457–D462.

Kent, W.J. (2002). BLAT—The BLAST-like alignment tool. Genome Res 12, 656–664.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., and Schlötterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS ONE 6, e15925.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40, D1202–D1210.

Lexer, C., Fay, M.F., Joseph, J.A., Nica, M.S., and 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.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Lin, Y.C., Wang, J., Delhomme, N., Schiffthaler, B., Sundström, G., Zuccolo, A., Nystedt, B., Hvidsten, T.R., de la Torre, A., Cossu, R.M., et al. (2018). Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. Proc Natl Acad Sci USA 115, E10970–E10978.

Ma, T., Wang, J., Zhou, G., Yue, Z., Hu, Q., Chen, Y., Liu, B., Qiu, Q., Wang, Z., Zhang, J., et al. (2013). Genomic insights into salt adaptation in a desert poplar. Nat Commun 4, 2797.

Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., Cornejo, O., Findley, S.D., Zheng, P., Utro, F., Royaert, S., et al. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol 14, r53.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., et al. (2014). The genome of *Eucalyptus grandis*. Nature 510, 356–362.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The

TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35, D883–D887.
Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.
Roiron, P., Ali, A.A., Guendon, J.L., Carcaillet, C., and Terral, J.F. (2004). Preuve de l'indigénat de Populus alba L. dans le Bassin méditerranéen occidental. Comptes Rendus Biologies 327, 125–132.
Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.
Singh, R., Ming, R., and Yu, Q. (2016). Comparative analysis of GC content variations in plant genomes. Tropical Plant Biol 9, 136–149.
Smit, A., Hubley, R., and Green, P. (2013–2015). RepeatMasker Open-4.0 (http://www.repeatmasker.org).
Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stölting, K.N., Nipper, R., Lindtke, D., Caseys, C., Waeber, S., Castiglione, S., and Lexer, C. (2013). Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol Ecol 22, 842–855.
Stölting, K.N., Paris, M., Meier, C., Heinze, B., Castiglione, S., Bartha, D., and Lexer, C. (2015). Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of *Populus alba* (Salicaceae), a widespread forest tree. New Phytol 207, 723–734.
Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515.
Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313, 1596–1604.
Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009a). The flowering world: a tale of duplications. Trends Plant Sci 14, 680–688.
Van de Peer, Y., Maere, S., and Meyer, A. (2009b). The evolutionary significance of ancient genome duplications. Nat Rev Genet 10, 725–732.
Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 11, 11.10.11-11.10.33.
Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40, e49.
Wu, G.A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., Perrier, X., Ruiz, M., Scalabrin, S., Terol, J., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol 32, 656–662.
Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35, W265–W268.
Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586–1591.

SUPPORTING INFORMATION

Figure S1 Gene region coverage assessed by BUSCOs (Benchmarking Universal Single-Copy Orthologs).

Figure S2 Comparison of CDS length (A), exon number per gene (B), exon length (C), and intron length (D). Asterisk indicates significant difference (P<0.05) in distribution between that species and *P. alba*.

Table S1 Locations of population samples for pool sequencing

Table S2 PacBio and whole-genome shotgun (WGS) reads used in genome sequencing

Table S3 PacBio and WGS reads used in the *P. alba* transcriptome sequencing

Table S4 Gene region coverage assessed by CEGMA

Table S5 Functional annotation of predicted genes for *P. alba*

Table S6 Alternative splicing statistics of protein-coding genes for *P. alba*

Table S7 GO enrichment analysis of 110 gene families unique to *P. alba*

Table S8 GO enrichment analysis of 1,150 expanded *P. alba* families

Table S9 GO enrichment analysis of 2,466 contracted *P. alba* families

The supporting information is available online at http://life.scichina.com and http://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.