New Phytologist

Research

Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection

Mingcheng Wang[1]*, Lei Zhang[1]*, Zhiyang Zhang[1], Mengmeng Li[1], Deyan Wang[1], Xu Zhang[2], Zhenxiang Xi[1], Ken Keefover-Ring[3], Lawrence B. Smart[4], Stephen P. DiFazio[5], Matthew S. Olson[6], Tongming Yin[7], Jianquan Liu[1,2] and Tao Ma[1]

[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China; [2]State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology & College of Life Sciences, Lanzhou University, Lanzhou 730000, China; [3]Departments of Botany and Geography, University of Wisconsin-Madison, 430 Lincoln Dr., Madison, WI 53706, USA; [4]Horticulture Section, School of Integrative Plant Science, New York State Agricultural Experiment Station, Cornell University, Geneva, NY 14456, USA; [5]Department of Biology, West Virginia University, Morgantown, WV 25606, USA; [6]Department of Biological Sciences, Texas Tech University, Box 43131, Lubbock, TX 79409-3131, USA; [7]Co-Innovation Center for Sustainable Forestry in Southern China, College of Forestry, Nanjing Forestry University, Nanjing 210037, China

Author for correspondence:
Tao Ma
Tel: +86 13519669951
Email: matao.yz@gmail.com

Received: *9 February 2019*
Accepted: *16 September 2019*

New Phytologist (2019)
doi: 10.1111/nph.16215

Key words: balancing selection, gene flow, phylogenomics, *Populus*, trans-specific polymorphisms.

Summary

- Phylogenetic analysis is complicated by interspecific gene flow and the presence of shared ancestral polymorphisms, particularly those maintained by balancing selection. In this study, we aimed to examine the prevalence of these factors during the diversification of *Populus*, a model tree genus in the Northern Hemisphere.
- We constructed phylogenetic trees of 29 *Populus* taxa using 80 individuals based on re-sequenced genomes. Our species tree analyses recovered four main clades in the genus based on consensus nuclear phylogenies, but in conflict with the plastome phylogeny. A few interspecific relationships remained unresolved within the multiple-species clade because of inconsistent gene trees. Our results indicated that gene flow has been widespread within each clade and also occurred among the four clades during their early divergence.
- We identified 45 candidate genes with ancient polymorphisms maintained by balancing selection. These genes were mainly associated with mating compatibility, growth and stress resistance.
- Both gene flow and selection-mediated ancient polymorphisms are prevalent in the genus *Populus*. These are potentially important contributors to adaptive variation. Our results provide a framework for the diversification of model tree genus that will facilitate future comparative studies.

Introduction

The phylogenetic histories of species are complicated, and it is now well understood that the persistence of ancestral polymorphisms across multiple speciation events contributes to the presence of gene genealogies that conflict with speciation history (Mayr, 1966; Schluter, 2001; Coyne & Orr, 2004). Incomplete lineage sorting (ILS), which results from the persistence of ancestral polymorphisms across multiple speciation events and the subsequent random fixation of these polymorphisms in different lineages, is one process that generates genealogical histories that are inconsistent with the species tree (Tajima, 1983; Pamilo & Nei, 1988; Degnan & Rosenberg, 2009). Because ILS requires the long-term maintenance of ancestral polymorphisms relative to speciation events, ILS is expected to be much more prevalent in clades with rapid radiations (Wu, 1991; Schluter, 2000; Arnold, 2006; Feng *et al.*, 2019). Another factor that influences the long-term maintenance of polymorphisms, and may also affect the extent of ILS, is balancing selection (Guerrero & Hahn, 2018), which may be more common in plants than has been historically recognised (Delph & Kelly, 2014). In the context of ILS, historical balancing selection actively maintains polymorphisms, and will result in a higher proportion of genes exhibiting ILS once the lineages fix, likely because of weakened selection pressures relative to drift. In such cases, orthologous sequences from the same loci will cluster by allele, rather than by species, thereby distorting phylogenetic trees (Charlesworth 2006; Fijarczyk & Babik, 2015; Gao *et al.*, 2015). However, although balancing selection should increase the frequency of ILS because ancient linked polymorphisms are maintained across multiple speciation events, it should not bias the genealogical topologies of linked sites towards specific genealogical histories (Fijarczyk & Babik, 2015; Gao *et al.*, 2015).

To date, the influence of balancing selection on the maintenance of polymorphisms has been reported for only a limited number of taxonomic groups at a few loci. For example, the

*These authors contributed equally to this work.

diverged alleles at the major histocompatibility locus (MHC) are often shared across distantly related vertebrate species (Klein *et al.*, 2007), and the divergent ABO blood alleles exist together in humans, gibbons and in Old World monkeys (Ségurel *et al.*, 2012). In plants, classic examples of divergent alleles maintained by balancing selection include self-incompatibility (S) and disease resistance (R) genes (Takebayashi *et al.*, 2003; Roux *et al.*, 2013; Karasov *et al.*, 2014). Recently trans-specific polymorphisms were reported for five genes in Arabidopsis and distantly related *Capsella* with divergence around 8 Myr, the function of which involved in adaptation to divergent habitats (Wu *et al.*, 2017). With the ability to now generate large whole-genome data sets to explore phylogenetic histories, we also can better identify the potential for long-term balancing selection to influence the maintenance of polymorphisms across speciation events.

Unlike ILS, historical hybridisation will bias the numbers of genealogies that exhibit histories in conflict with the history of speciation (Leaché *et al.*, 2014; Solís-Lemus *et al.*, 2016). In fact, the presence of this bias underlies the rationale for the ABBA–BABA test, which has been used to identify historical patterns of hybridisation throughout the tree of life (Green *et al.*, 2010; Durand *et al.*, 2011). One characteristic of this test is that the genealogical effects of ancient hybridisation will persist in the genomes of extant species (Lamichhaney *et al.*, 2015; Novikova *et al.*, 2016; Feng *et al.*, 2019). Plants are well known to maintain the ability to hybridise even after clear morphological differentiation between species has occurred (Eaton *et al.*, 2015; Baute *et al.*, 2016; Pease *et al.*, 2016; Y. Liu *et al.*, 2017; Crowl *et al.*, 2019), which will complicate the reconstruction of speciation histories.

In this study, we aimed to examine gene flow and ancient polymorphisms within diversification of the genus *Populus*. All species of the genus are collectively known as poplars and are widely distributed in the Northern Hemisphere from subtropical to boreal forests (Eckenwalder, 1996), where they can act as keystone species (Whitham *et al.*, 2006). In addition, most poplars exhibit ecological flexibility, with diverse adaptations and large population sizes. Numerous species have been artificially planted around the world, and poplars account for more than half of the planted trees in China, where they are used for the wood, pulp and paper industries and for environmental restoration projects (Isebrands & Richardson, 2014). Although this genus has long been used as a model for diverse studies in trees (Tuskan *et al.*, 2006; Jansson & Douglas, 2007), phylogenetic relationships within the genus remain unclear. Frequent interspecific hybridisation and clonal expansion have perplexed taxonomists of the genus, with acknowledged species, varieties, and hybrids ranging from 22 to 85 (Eckenwalder, 1996). Six sections are traditionally recognised (*Abaso*, *Turanga*, *Populus*, *Leucoides*, *Tacamahaca* and *Aigeiros*) based on morphological traits (Eckenwalder, 1996). However, these sections have not been consistently supported by molecular evidence, and relationships among and within the sections have been the subjects of controversy (Hamzeh & Dayanandan, 2004; Cervera *et al.*, 2005; Wang *et al.*, 2014; X. Liu *et al.*, 2017, Zhang *et al.*, 2018). For example, based on morphological traits and fossil evidence, the basal lineages of *Populus* remain unresolved between lineages from sect. *Turanga* ranging from central Asia to Africa and *P. mexicana* of the monotypic sect. *Abaso* (Eckenwalder, 1996). Moreover, phylogenies constructed with molecular data also conflict. A phylogenetic analyses of chloroplast genomes identified *Turanga* as the basal section (Zhang *et al.*, 2018), whereas a phylogeny based on multiple low-copy genes suggested that *P. mexicana* was basal (X. Liu *et al.*, 2017). These inconsistencies, together with a lack of resolution in some prior phylogenetic analyses, may result from influences of both interspecific gene flow and ILS of ancient polymorphisms.

In order to investigate the phylogeny of *Populus* and factors that have influenced conflicting genealogical histories among loci, we generated a whole-genome data set consisting of 80 individuals of 29 taxa for the genus. Our sampling covers all six sections of the genus and most distinct species as well as hybrids (Eckenwalder, 1996). We first reconstructed a backbone phylogeny for the genus based on variant sites within the single-copy genes. We then examined the presence of hybridisation between the main lineages by identity-by-descent (IBD) and ABBA–BABA analyses. Finally, we selected six species from the three main lineages of the genus to identify genes with ancient polymorphisms that were likely maintained by balancing selection. These results provide a detailed examination of the phylogenetic relationships and evolutionary diversification of the model genus *Populus*.

Materials and Methods

Sample collection, sequencing and mapping

Leaves representing 63 individuals of 24 species were collected from natural populations and dried on silica gel (Supporting Information Table S1). For each individual, whole genomic DNA was extracted using the CTAB protocol (Doyle & Doyle, 1987). Paired-end Illumina genomic libraries were prepared and sequenced on the HiSeq 2000 and HiSeq 2500 Illumina platforms following the manufacturer's instructions (Illumina, San Diego, CA, USA). Previously published genome sequences of 17 individuals from six species were also included in our analysis (Slavov *et al.*, 2012; Geraldes *et al.*, 2015; Wang J. *et al.*, 2016; Ma *et al.*, 2018) (Table S1). In total, genome resequencing data for 80 individuals from 29 taxa covering all six sections of the genus were obtained.

The raw reads were subjected to quality control. Low-quality reads were removed if they met any of the following criteria: (1) ≥5% unidentified nucleotides; (2) a phred quality ≤ 7 for >65% of read length; and (3) reads overlapping > 10 bp with the adapter sequence, allowing < 2 bp mismatch. We then mapped these high-quality reads to the *P. trichocarpa* reference genome v.3.0 (Tuskan *et al.*, 2006) using BWA-MEM v.0.7.12-r1039 with default parameters (Li & Durbin, 2009). Duplicated reads were removed using the 'rmdup' function of SAMTOOLS v.1.3.1 (Li *et al.*, 2009). Finally, the Genome Analysis Toolkit (GATK) v.3.6 (McKenna *et al.*, 2010) was used to perform local realignment of reads to enhance alignments in regions around putative InDels.

Single nucleotide polymorphisms and genotype calling

Single nucleotide polymorphisms (SNPs) and short InDels were called with GATK UnifiedGenotyper with default parameters for each species separately. Some filtering steps were performed to reduce false positives: (1) SNPs and InDels with a quality score < 30 were removed; (2) SNPs with more than two alleles were removed; (3) SNPs at or within 5 bp from any InDels were removed; (4) genotypes with extremely low (less than one-third average depth) or extremely high (greater than three-fold average depth) coverage were assigned as missing sites; and (5) SNPs with missing genotypes in all individuals were removed. The final single nucleotide variants (SNVs) were generated by merging the results of each species, and only biallelic variant sites were retained for downstream analysis. The final SNVs were annotated using SnpEff v.4.3 (Cingolani *et al.*, 2012). A principal component analysis was performed on these SNVs using the smartpca program in EIGENSOFT (Patterson *et al.*, 2006).

Ancestral state reconstruction

To obtain data from an appropriate outgroup for the phylogenetic analysis, the *Salix suchowensis* genome (v.5.2) (Dai *et al.*, 2014) and the *S. purpurea* genome (v.1.0) (Zhou *et al.*, 2018) were downloaded and aligned to the *P. trichocarpa* reference genome (v.3.0) using last v.869 (Kiełbasa *et al.*, 2011). Here, *c.* 151.04 and 151.68 Mb of the *P. trichocarpa* genome sequences could be accurately aligned by *S. suchowensis* and *S. purpurea*, respectively. The sites that were covered by both *Salix* genomes were then extracted from the alignment file and directly used as the ancestral state. In total, we identified the ancestral state at about 81% of the whole-genome SNVs, which covered about 98% of the coding SNVs from the single-copy genes identified below.

Phylogenetic analyses

A maximum-likelihood (ML) tree of the concatenated whole-genome SNVs was constructed using RAxML v.8.0.17 (Stamatakis, 2006). One hundred bootstrapped alignment files were generated with the option '-f j'. For each file, a ML tree was built with the option '-f d -N 3' based on the GTRCAT model, which has been designed to accelerate the computations on large datasets with > 50 taxa. Finally a majority-rule consensus tree of the bootstrapped trees was generated using the 'consensus' function of the R package ape (Paradis *et al.*, 2004) and support values of tree splits were calculated using the SumTrees program from the DendroPy package (Sukumaran & Holder, 2010). The same pipeline was applied to SNVs at four-fold degenerate sites (4D SNVs).

We also identified single-copy genes using the OrthoMCL (Li *et al.*, 2003) method for all protein-coding genes from seven Salicaceae species: *S. suchowensis* (Dai *et al.*, 2014), *S. purpurea* (Zhou *et al.*, 2018), *P. trichocarpa* (Tuskan *et al.*, 2006), *P. euphratica* (Ma *et al.*, 2013), *P. pruinosa* (Yang *et al.*, 2017), *P. deltoides* (https://genome.jgi.doe.gov/) and *P. alba* var. *pyramidalis* (Ma *et al.*, 2019). The SNVs within these genes were then extracted across all 29 species and divided into three datasets: (1) the first and second codon positions (C_{12}); (2) the third codon position (C_3); and (3) complete coding sequences (CDS). For each dataset, the individual gene trees were constructed using RAxML v.8.0.17 and a species tree was estimated using recently developed coalescence methods in MP-EST v.1.5 (Liu *et al.*, 2010) and Astral v.4.11 (Mirarab *et al.*, 2014). Gene trees were superimposed using DensiTree (Bouckaert, 2010). The ETE2 package (Huerta-Cepas *et al.*, 2010) was used to examine different topologies of gene trees. We also estimated a 'concatenation tree' using RAxML v.8.0.17 for concatenated sequences of the C_{12}, C_3 and CDS datasets, respectively. Finally, we reconstructed a ML chloroplast DNA phylogeny based on 77 concatenated genes present in all the Salicaceae species using RAxML with 500 bootstrap replicates.

Identification of gene flow

To detect shared haplotypes and thus possible gene flow between species, we performed an IBD blocks analysis based on whole-genome SNVs using Beagle v.4.1 (Browning & Browning, 2013) with the following parameters: window = 50 000; overlap = 5000; ibdtrim = 100; ibdlod = 10. To evaluate the correlations between IBD block length and recombination, the population-scaled recombination rates (ρ) of *P. trichocarpa* and *P. tremula* were obtained from a previous study (Wang J. *et al.*, 2016). We also performed ABBA–BABA analysis using *S. suchowensis* as an outgroup in all comparisons. In brief, for the ordered alignment (((S_1, S_2), S_3), O), two classes of shared derived alleles were identified: the ABBA site refers to a pattern in which S_1 has the outgroup allele and S_2 and S_3 share the derived allele, the BABA site corresponds to patterns in which S_1 and S_3 share the derived allele and S_2 has the outgroup allele. *D* statistics were then calculated as (ABBA − BABA)/(ABBA + BABA) (Green *et al.*, 2010). Under the null hypothesis of ILS, the number of ABBA and BABA sites is expected to be equal ($D = 0$). Alternatively, significant deviation of *D* from 0 suggests other events, in particular S_3 exchanging genes with S_1 or S_2 (Durand *et al.*, 2011). *D* statistics were estimated using angsd 0.9.21 (Korneliussen *et al.*, 2014) with a block size of 5 Mb, and *Z*-scores were calculated using the *m*-block jackknife method (Busing *et al.*, 1999). A *Z*-score with an absolute value > 3 was considered statistically significant.

Identification of trans-specific polymorphisms under balancing selection

To investigate trans-specific polymorphisms within *Populus*, we analysed 72 additional individuals of six species from previously published data sets representing three of the major lineages (Slavov *et al.*, 2012; Geraldes *et al.*, 2015; Wang J. *et al.*, 2016; Ma *et al.*, 2018). We applied the same criteria as used above for reads mapping, SNP and genotype calling and filtering, and only retained SNVs with missing genotype rates < 20% in all six species. Shared biallelic SNVs were counted and the genomic divergence (F_{ST}) between each pair of species was estimated using

VCFTOOLS v.0.1.14 (Danecek *et al.*, 2011). Joint allele frequency spectra were calculated for biallelic sites between species, unfolded using *S. suchowensis* as the outgroup, and plotted using DADI v.1.7.0 (Gutenkunst *et al.*, 2009).

To identify genes under balancing selection, we focused only on SNVs located in genic regions and shared by all six species. To filter out potential duplicated genes, we estimated the copy number for these genes in each individual based on the ratio of exon coverage depth divided by genome coverage depth (Hastings *et al.*, 2009), and retained genes with a ratio between 0.4 and 1.6 in all individuals. To avoid the potential for sampling sites influenced by convergence resulting from repeated mutations among shared SNVs, we considered genes with more than one shared SNV and at least one shared SNV in coding regions. Finally, the genes with at least two shared SNVs in linkage disequilibrium ($r^2 > 0.3$) in all three major lineages were selected as candidate genes under balancing selection.

Data accessibility

The sequencing data have been deposited in the Genome Sequence Archive in the BIG Data Center (BIG Data Center Members, 2019), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA001510 that is publicly accessible at http://bigd.big.ac.cn/gsa. The whole-genome SNV data and gene trees have been deposited in GitHub (https://github.com/wangmcyao/Whole-genome-SNPs-and-gene-trees-of-genus-Populus).

Results

Phylogenetic analyses

To clarify the phylogenetic relationships within *Populus*, we collected 1.12 Tb of whole-genome sequencing data of 80 individuals from 29 species, representing all six sections and almost all of the currently recognised species of this genus (Isebrands & Richardson, 2014). Sequences were first aligned to the reference genome of *P. trichocarpa* and about 88% were successfully mapped, covering 81% of the genome and yielding an average depth of 29 × per individual (Table S1). We applied stringent variant calling and quality filters to identify a final set of 12.93 million biallelic SNVs (Table S2). Among these, 1.94 million nonsynonymous SNVs and 1.8 million synonymous SNVs were identified.

We next inferred a concatenated genome tree using a ML method for genome-wide and four-fold degenerate (4D) SNVs, respectively. Both trees were highly resolved for most clades and showed nearly identical topologies, with the only differences in the positions of *P. lasiocarpa* and *P. ningshanica* (Fig. S1). In both phylogenies, *P. mexicana* of sect. *Abaso* diverged first with high support, followed by sect. *Populus*, which was monophyletic and clearly divided into two subclades, one with only Asian species and a second clade with species representing Asian, Europe and North America. The next split included the monophyletic sect. *Turanga* and a polyphyletic clade consisting of members of sects. *Aigeiros*, *Tacamahaca* and *Leucoides*, which we refer to from this point forward as 'ATL'. The divergence of these four clades were also supported by a principal component analysis (Fig. S2). We found that *c.* 6% of the identified genome-wide SNVs were shared among these clades (Fig. S3), which may contribute significantly to phylogenetic inconsistencies.

To further investigate the phylogenetic conflicts of this genus, we identified 5305 single-copy orthologous genes across seven Salicaceae genomes and extracted 620 531 SNVs within the coding region of these genes. Individual gene trees constructed from these data generally had low support values and the relationships among sections of *Populus* were highly variable among them (Figs S4, S5). We also constructed concatenation trees based on different partitions of these orthologues, and the results consistently supported the basal position of sect. *Abaso* and followed by the successive divergences of the other three clades: sect. *Turanga*, sect. *Populus* and ATL (Fig. S6). The C_{12} and CDS concatenation trees supported the placement of sect. *Turanga* as basal to sects. *Populus* and ATL, whereas in the C_3 concatenation tree, sect. *Populus* diverged earlier than sect. *Turanga*. These phylogenetic conflicts among different gene partitions of orthologues were also observed in our species tree analyses. Both ASTRAL and MP-EST methods generated species trees nearly identical with the concatenation tree when applied to different gene partitions (Figs S7, S8): sect. *Populus* was placed sister to the ATL clade in all the species trees except the C_3 partition. However, this phylogenetic relationship was supported by C_3 partition after eliminating gene trees with low values of average bootstrap support (Fig. S9). Among these trees, the C_{12} species tree had the highest bootstrap support in all the major nodes, and thus was considered as the most reliable topology to resolve the *Populus* phylogeny (Fig. 1a). However, based on the phylogenetic analyses of plastome sequences, only the monophyly of sect. *Turanga* was supported, which occupied the basal position, while *P. mexicana* of sect. *Abaso* clustered with species from sects. *Aigeiros* (*P. deltoides* and *P. fremontii*), *Tacamahaca* (*P. angustifolia*, *P. trichocarpa* and *P. balsamifera*) and *Leucoides* (*P. heterophylla*) (Fig. S10).

Extensive gene flow

The highly variable relationships indicated by the gene trees, and the striking discordance between the plastome phylogeny and our coalescence tree, may be due to ILS and interspecific gene flow. To gain further insight into the relationships among the species, we searched for IBD haplotypes using BEAGLE. No IBD blocks were identified across sects. *Abaso*, *Turanga*, *Populus* or ATL; however, abundant shared IBD blocks were detected in comparisons within the four major sections, both between and within species (Figs 2a, S11). Of the shared blocks between species, 75.9% (62 539) were detected within the ATL clade, while 15.2% (12 483) were detected within sect. *Populus*, and only 8.9% (7323) within sect. *Turanga*. The length of IBD haplotypes shared between species (Fig. S11a) ranged from 1.9 kb to 1.5 Mb (median = 23.2 kb). As expected, extensively shared haplotypes were found between recently diverged species, for example, between *P. trichocarpa* and *P. balsamifera* (median = 15.1 kb,

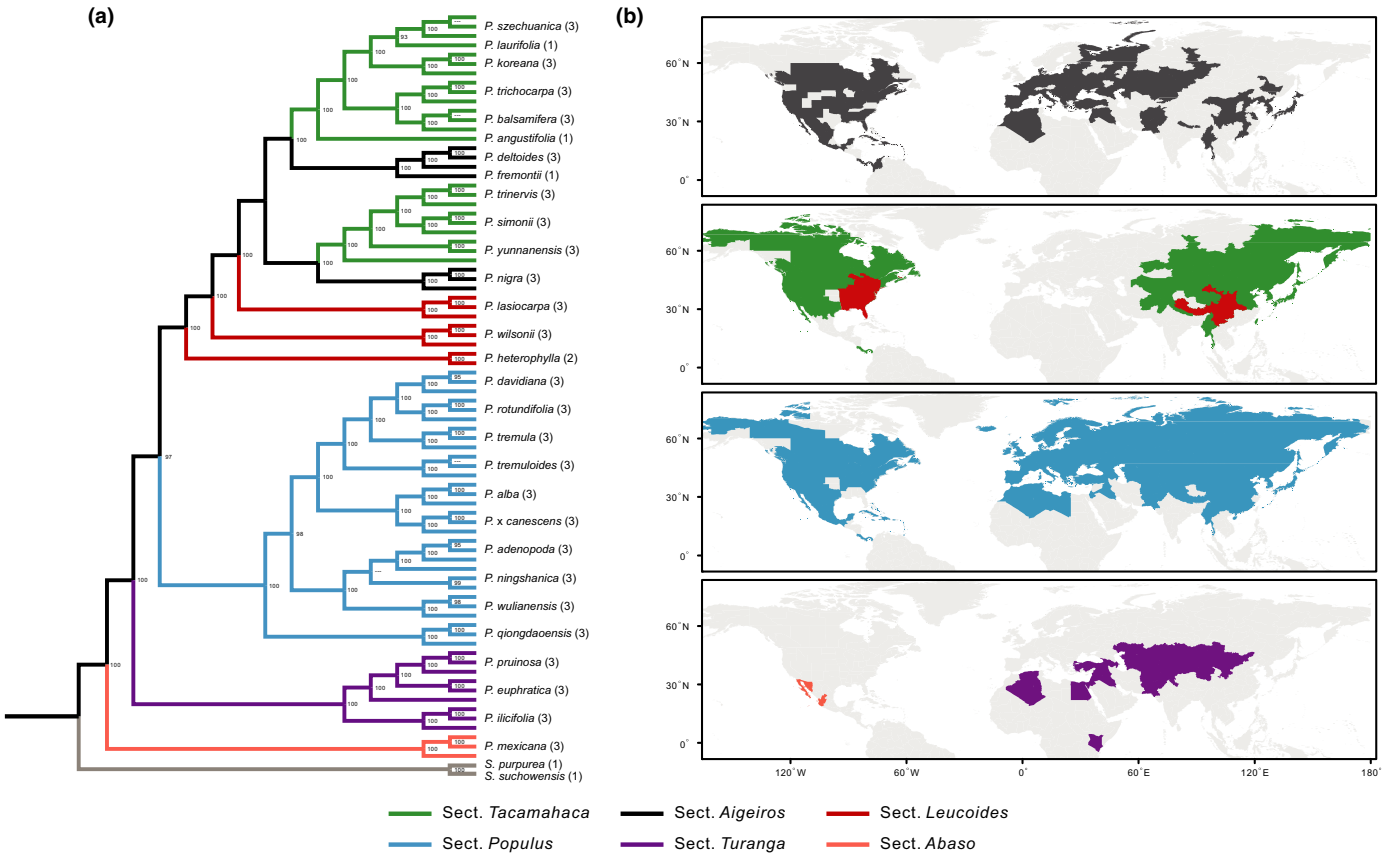

Fig. 1 (a) Phylogenetic relationships among 29 *Populus* taxa (80 samples) and two *Salix* species rooted on *S. purpurea* and *S. suchowensis* based on the first and second codon positions of 5305 single-copy genes analysed with ASTRAL species tree methods. Numbers at each node represent bootstrap values. The numbers in parentheses next to the taxa names represent the number of samples for each taxon. (b) The geographical distributions of six intragenic sections of genus *Populus*.

maximum = 1.53 Mb) of sect. *Tacamahaca* (Fig. S11b), between *P. rotundifolia* and *P. davidiana* (median = 14.1 kb, maximum = 1.15 Mb) of sect. *Populus* (Fig. S11c), and between *P. euphratica* and *P. pruinosa* (median = 7.3 kb, maximum = 454.1 kb) of sect. *Turanga* (Fig. S11d). All of these closely related species showed evidence of extensive interspecific gene flow in previous studies (Meirmans *et al.*, 2010; Zheng *et al.*, 2017; Ma *et al.*, 2018). Moreover, we also found that the hybrid aspen, *P.* × *canescens*, shared much longer haplotypes (median = 17.2 kb, maximum = 1.37 Mb) with its parents, *P. alba* and *P. tremula* (Fig. S12a), as expected. A similar length distribution of shared haplotypes was found for *P. wulianensis* and *P. ningshanica* (Fig. S12b), both of which might be interspecific hybrids between *P. adenopoda* and *P. davidiana*. We observed a negative relationship between IBD length and recombination rates for closely related species, but not for distantly diverged species (Fig. S13; Table S3), supporting theoretical predictions that the IBD blocks will degrade over time due to recombination, even in regions where recombination is rare. In addition, we used ABBA–BABA tests to examine gene flow within each clade. Our results suggested that most pairs of species within each clade showed obvious gene flow and those that there was a positive relationship between the extent of IBD and the level of gene flow (Fig. S14; Table S4). All of these results indicated that frequent hybridisation occurred among poplar species within the same section or clade.

To further examine the possible occurrence of gene flow across the deep clades comprising the four major sections of *Populus*, we performed additional ABBA–BABA test based on the phylogenetic relationships recovered above. We found that *P. mexicana* was more closely related to the ATL clade than to any other species of sects. *Turanga* and *Populus* (Fig. 2b). Further detailed analyses revealed gene flow between *P. mexicana* and *P. heterophylla*, which was statistically significant regardless of the ATL species used for the statistical comparison (Z-score > 3 and P-value < 0.0027; Fig. 2c). Both *P. mexicana* and *P. heterophylla* are found in North America, but their range limits do not currently overlap (Isebrands & Richardson, 2014). Similarly, we also found that *P. pruinosa* was more closely related to the common ancestor of *P. trinervis*, *P. simonii*, *P. yunnanensis* and *P. nigra* than to other species of sects. *Aigeiros* and *Tacamahaca* (Fig. S15a), while the common ancestor of *P. davidiana* and *P. rotundifolia* was more closely related to *P. lasiocarpa* than to any other species of the ATL clade (Fig. S15b), suggesting ancient gene flow between these species (Fig. S16). These complex admixture histories were also supported by the widespread incongruence between the chloroplast tree and the species tree (X. Liu *et al.*, 2017; Zhang *et al.*, 2018; here), and therefore could be

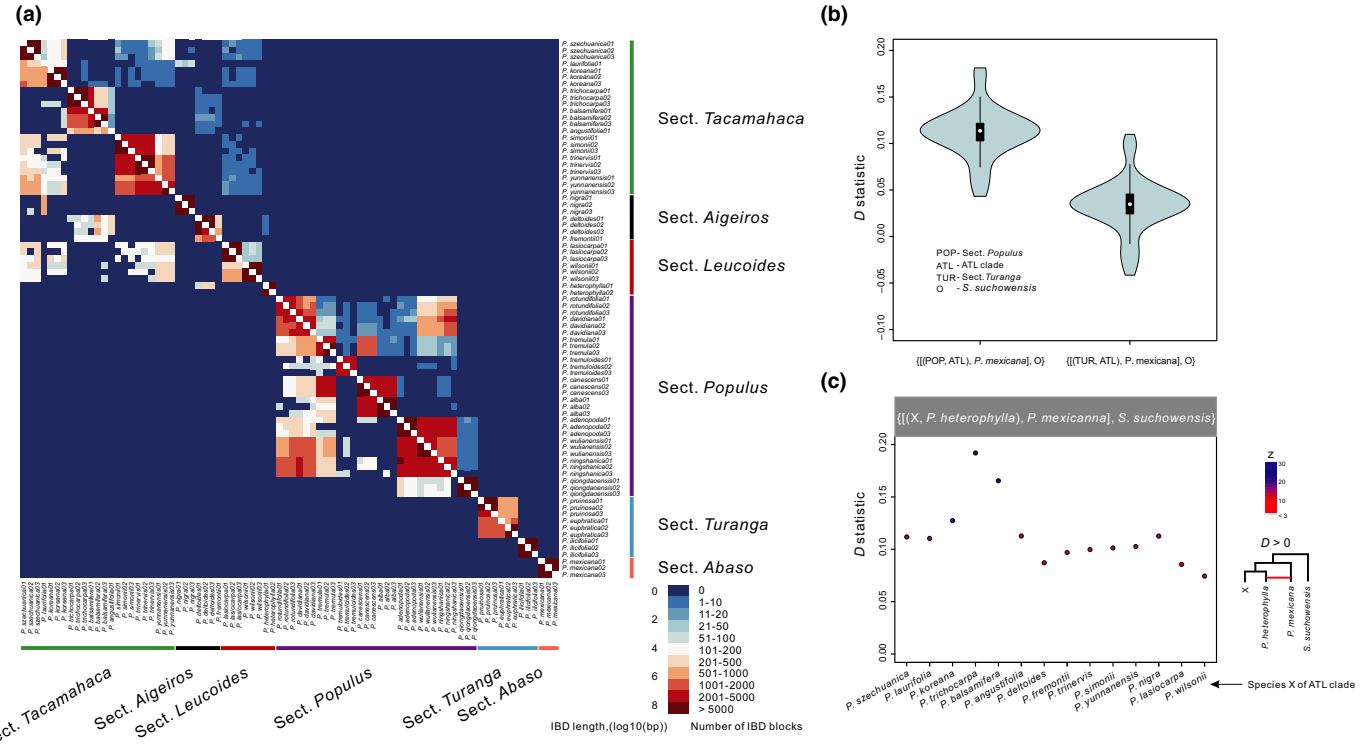

Fig. 2 (a) Estimated haplotype sharing in the genus *Populus*. Heatmap colours represent the total length (below the diagonal) and the total number (above the diagonal) of identity-by-descent (IBD) blocks for each pairwise comparison. (b, c) ABBA–BABA tests provide evidence of gene flow between *P. mexicana* and *P. heterophylla*.

partially responsible for the recovery of paraphyly for sect. *Leucoides* with respect to sects. *Aigeiros* and *Tacamahaca* (Fig. 1a).

Trans-specific polymorphisms under long-term balancing selection

Shared polymorphisms among species can be maintained not only by ILS and interspecific gene flow, but also by long-term balancing selection (Charlesworth, 2006; Fijarczyk & Babik, 2015). To investigate the trans-specific polymorphisms under long-term balancing selection in the genus *Populus*, we further performed a whole-genome scan across 72 individuals from six species (Table S5), *P. trichocarpa*, *P. balsamifera*, *P. tremula*, *P. tremuloides*, *P. euphratica* and *P. pruinosa*, representing three clades with sufficient divergence (6–11 Myr, Zhang *et al.*, 2018). These demographic requirements ensure that observed trans-specific polymorphisms are more likely to have been maintained under balancing selection in each lineage, rather than just drift. Because only three individuals from the same small population of *P. mexicana* were sampled, we excluded this clade for our analyses. The genome-wide averages of genetic divergence (F_{ST}) between species ranged from 0.29 to 0.41 for comparisons within the same clade and from 0.75 to 0.86 for comparisons between different clades (Table S6), indicating high genetic differentiation among these species, especially among the three clades that we analysed. This was also supported by the lack of correlation among allele frequencies for polymorphisms shared across species (Fig. S17). Despite their clear divergence, we found abundant trans-specific polymorphisms (Table S6; Fig. S18). We found *c.* 2.17–2.82 million trans-specific polymorphisms among species pairs within the same clade, and *c.* 0.21–0.54 million among species pairs from different clades (Table S6). However, it is not clear whether these trans-specific polymorphisms are maintained by balancing selection, ILS or gene flow.

To investigate this further, we focused on 7711 SNVs that segregated in all six species (Fig. S18a). Among these, we observed 2925 SNVs located in the genic regions of 1007 genes (Table S7). After excluding genes that showed evidence of deletion or duplication by copy number filtering, 484 genes containing 1031 SNVs were retained (Fig. 3a). To prevent any shared SNVs due to repeated mutations, we selected only genes with two or more shared SNVs, at least one of which was in an exon. This identified 100 genes containing 562 shared SNVs. As expected for ancestral polymorphisms maintained by selection, these genes showed substantially higher values of Tajima's *D* than for all genes in the genome (Fig. 3b). Finally, we focused on genes with at least two shared SNVs that were in strong linkage disequilibrium ($r^2 > 0.3$) across species from all three deeply diverged clades to reduce false positives. Using these strict criteria, we ended up with 45 genes containing 150 shared SNVs (Table S8). These genes with shared haplotypes showed much higher nucleotide diversity (π) and intermediate allele frequency relative to the genome (Table S8), and thus were more likely to be maintained by long-term balancing selection. We determined the haplotypes in the identified regions for each species and found that these sequences clustered by allele rather than species

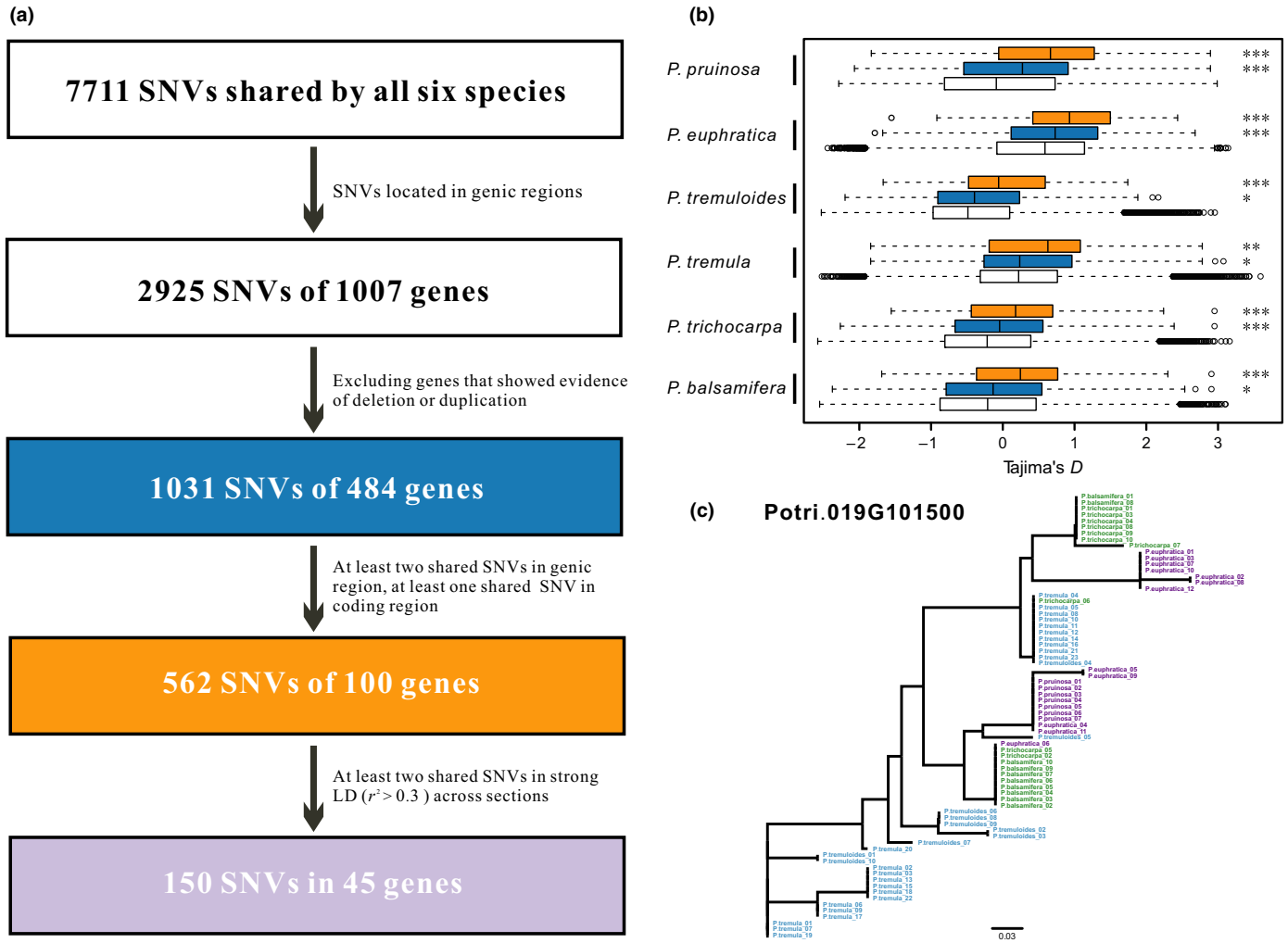

Fig. 3 (a) Pipeline of the SNV filtering process to identify candidate trans-specific polymorphisms under balancing selection. (b) The Tajima's *D* estimator is higher for candidate genes showing no evidence of deletion or duplication (blue) and genes with at least two shared SNVs in genic region and at least one shared SNV in coding region (orange) compared to all genes with variation (white). *, $P < 0.05$ and > 0.01; **, $P < 0.01$ and > 0.001; ***, $P < 0.001$ (Mann–Whitney test). (c) Candidate regions in the gene Potri.019G101500 produce an allelic tree, rather than a species tree. Text colours indicate sections, as in Fig. 1. Additional examples of trans-specific polymorphism can be seen in Supporting Information Fig. S19.

(Figs 3c, S19). Therefore, we considered these genes with shared haplotypes as candidates for long-term balancing selection.

Gene ontology enrichment analyses revealed that these genes were mainly associated with plant development, reproduction and response to biotic and abiotic stress (Table S9). For example, *BODYGUARD 3* (*BDG3*) encodes an epidermis-specific extracellular α/β-hydrolase fold–containing protein that may function in the formation of the epidermal cell wall and cuticle (Kurdyukov *et al.*, 2006); *MTN1* encodes one of the 5-methylthioadenosine nucleosidases that are essential for normal vascular development and reproduction in *A. thaliana* (Waduwara-Jayabahu *et al.*, 2012), whereas *FORMIN HOMOLOGY 5* (*FH5*) encodes a protein with similarity to formins that is involved in cytokinesis (Ingouff *et al.*, 2005) and plays pivotal roles in the regulation of endosperm development (Fitz Gerald *et al.*, 2009) and in the establishment of actin polarity during pollen germination (Cheung *et al.*, 2010; Liu *et al.*, 2018). Moreover, *ANK1* is a member of the ankyrin (*ANK*) gene cluster in *A. thaliana*, which encodes ANK transmembrane proteins that play a wide variety of roles in protein–protein interactions, signal transduction and defence responses (Lu *et al.*, 2003; Becerra *et al.*, 2004). Extremely high diversity and potential signals of balancing selection were also observed for ANK transmembrane proteins in *A. thaliana* (Du *et al.*, 2007), suggesting that genes from this family may be common targets of balancing selection in numerous species. We did not find that the well known *S* (sterility) genes, *SRK* and *SCR* in Arabidopsis (Kusaba *et al.*, 2001), in our scan for shared haplotypes were under balancing selection in *Populus*. However, we found evidence for long-term balancing selection in a homologous gene of Arabidopsis *AT4G21390*, which is tightly linked to the S-locus and encodes a S-locus lectin protein kinase family protein (Kamau & Charlesworth, 2005; Kamau *et al.*, 2007). In addition, we also found that several genes subjected to balancing selection encoded proteins involved in the response to biotic and abiotic stress, including *CYCLOPHILIN 38* (*CYP38*) (Fu *et al.*, 2007), *RESPONSE REGULATOR 22* (*RR22*) (Kang *et al.*,

2012), *Filamentous temperature sensitive H 11* (*FtsH11*) (Chen *et al.*, 2006), *MALE DISCOVERER 1-INTERACTING RECEPTOR LIKE KINASE 2* (*MIK2*) (Wang T. *et al.*, 2016; Van der Does *et al.*, 2017), *Drought-induced protein 19* (*Di19-3*) (Qin *et al.*, 2014) and others (Table S8).

Discussion

Our phylogenetic analyses of genomic data recovered four clades in the tree model genus *Populus*, by contrast with the six sections previously acknowledged based on morphological traits (Eckenwalder, 1996). One species, *P. mexicana* from the southern part of North America, was identified as a basal monotypic clade. Within each clade, we identified frequent gene flow and hybridisation between different species. We also found that gene flow occurred between the four clades during their early diversification. We confirmed that numerous ancient polymorphisms persisted across different species of three major clades through balancing selection. Both gene flow and ILS of ancient polymorphisms obviously violate a model of strict bifurcating divergence of the genus *Populus* between and within the major clades.

Phylogenetic relationships, shared polymorphism and gene flow

Our analyses based on nuclear genomic data recovered four well supported clades: sects. *Abaso*, *Turanga* and *Populus*, and ATL (Figs 1, S1, S2, S5–S9). The three previously identified sections within the ATL clade, sects. *Aigeiros*, *Tacamahaca* and *Leucoides*, were found to be paraphyletic with respect to one another (Figs S1, S5–S9). Most phylogenetic analyses suggested that the monospecific sect. *Abaso* diverged first, followed by sect. *Turanga* and then sect. *Populus* and ATL (Fig. 1). However, in analyses of some gene partitions, sect. *Turanga* is sister to ATL while sect. *Populus* diverged following sect. *Abaso* (Figs S1, S5–S9). The conflict among these trees can be partially explained by the strong base compositional bias and high mutation rates for the third codon position (Jarvis *et al.*, 2014; L. Liu *et al.*, 2017). In any case, phylogenetic inconsistencies among different datasets of the nuclear genome are mainly related to relationships among sects. *Turanga*, *Populus* and the ATL clade while sect. *Abaso* always diverged first. Therefore, the recovered interclade relationships based on nuclear genomic data are consistent with an origin of this genus in North America and then further dispersal to other regions of the Northern Hemisphere (Fig. 1). This is generally consistent with fossil evidence, which suggests that sect. *Abaso* appeared first in North America (Eckenwalder 1996).

By contrast with the nuclear results, our phylogenetic analyses of plastomes only recovered monophyly of sect. *Turanga*, and supported its basal position in the genus, whereas the other sections did not show corresponding monophyletic clustering (Fig. S10). Both gene flow and ILS of ancient polymorphism are likely to have contributed to these inconsistent histories among the four major clades, but gene flow was likely to have played a more important role for the following two reasons. First, all four clades diverged from one another between 6 and 11 Ma based on the fossil calibrations of the plastome phylogeny (Zhang *et al.*, 2018). Therefore, sufficient generations have passed that, in the absence of selection, ancient polymorphisms should have been fixed by genetic drift. In addition, we found that the internodes between these major clades were relatively short and the random fixation of ancient polymorphisms across the radiative polytomy also likely led to the phylogenetic inconsistencies (Wu, 1991). Under such a scenario, it is difficult to discern ILS from gene flow. Second, our plastome phylogeny recovered a close relationship of sect. *Abaso* with *P. heterophylla* and related species from the ATL clade (Fig. S10), which was also supported by our ABBA–BABA tests that detected significant gene flow between *P. mexicana* and the ATL clade, especially with *P. heterophylla* (Fig. 2c). It should be noted, however, that we failed to detect shared IBD blocks between these groups, suggesting that this gene flow would have occurred very early, and that subsequent recombination has erased the long IBD haplotypes. Therefore, these results may be best interpreted as chloroplast captures during the early hybridisation of two ancient lineages when reproductive isolation was not yet complete, although we cannot completely exclude the possibility of ILS. Under this hypothetical scenario, after the ancient hybridisation between the ancestral *P. mexicana* and the ancestor that gave rise to *P. heterophylla* and related species (referred here as the *heterophylla*-like ancestor), repeated backcrosses to the ancestral *P. mexicana* led to capture and fixation of the *heterophylla* chloroplast in *P. mexicana* (Fig. S20; X. Liu *et al.*, 2017). This scenario is consistent with the fossil record, in which sect. *Leucoides* is the first to appear in North America following sect. *Abaso* (Eckenwalder 1996).

Ancestral gene flow between pairs of species within each clade was apparent from the numerous shared IBD haplotypes between most pairs of species (Fig. 2a) and our ABBA–BABA tests results. Moreover, the quantity and extent of IBD haplotypes is proportional to the level of gene flow between pairs of species (Fig. S14). *P.* × *canescens*, which is a hybrid (primarily F1s) between *P. alba* and *P. tremula* (Rajora & Dancik, 1992), had the longest IBD haplotypes, supporting the use of shared IBD blocks as an indicator of past gene flow. ILS of ancient polymorphisms should also exist across many pairs of *Populus* species because of the short divergence times among many species pairs within the four main clades (Ingvarson, 2010). These factors combine to confound the reconstruction of the bifurcating relationships among the current species within and among clades. Nonetheless, gene flow may have been an important contributor to the early local adaptation and divergence of *Populus* species (Suarez-Gonzalez *et al.*, 2016, 2018), which was also indicated in other species groups (Sun *et al.*, 2018; Wu *et al.*, 2018). Further detailed studies based on more sampling of closely related species may be a fruitful avenue towards understanding the historical influences of gene flow in *Populus*.

Trans-specific polymorphisms mediated by balancing selection

Balancing selection can maintain ancient polymorphisms over long time frames and across species boundaries (Ségurel *et al.*,

2012). In such cases, phylogenetic analysis of orthologous genes may not reflect true species relationships because they will cluster by allele rather than by species. To detect such polymorphisms, we sampled 72 individuals from six species across three of the major clades in *Populus*. We identified 45 genes with polymorphisms that segregated in all six species across these three deeply diverged clades and exhibited patterns consistent with balancing selection contributing to their long-term maintenance. These three clades are sufficiently diverged that trans-clade polymorphisms resulting from ILS of ancient polymorphisms were unlikely. As previously suggested, ancestral gene flow has occurred among these clades, so we cannot totally exclude the possibility that ancient hybridisation and introgression was the source of these shared polymorphisms. However, these genes all contain species-specific polymorphisms as well, so recent hybridisation is not likely to have accounted for the shared polymorphisms; otherwise, long haplotypes should be shared across the three sections of *Populus*. In addition, introgression from ancient hybridisation also would have been subject to lineage sorting, and polymorphisms would not be likely to persist in the absence of selection. Introgression should cause the sharing of similar alleles across species but would not explain the presence of divergent alleles coexisting in both species. Therefore, it is more likely that the detected trans-specific polymorphisms arose from ancient polymorphisms maintained by balancing selection, a scenario that is also supported by elevated Tajima's *D* values (Fig. 3b). It should be noted that our stringent filtering criteria is likely to have seriously underestimated the number of trans-specific polymorphisms mediated by balancing selection. More sites would have been identified if we had relaxed the filtering criteria or focused on the polymorphisms shared by species from two clades (Fig. S18). However, as the criteria become less stringent, a larger proportion of the trans-specific polymorphisms are likely to have resulted from persistence after hybridisation instead of being maintained by balancing selection. In any case, our results revealed that the persistence of selection-mediated ancestral polymorphisms is likely to have been prevalent across the long evolutionary history of the genus *Populus*, which will increase the number of loci that are inconsistent with the true species tree because of increased ILS (Guerrero & Hahn, 2018).

In animals, genes identified to have experienced balancing selection are mainly suggested to be responsible for host–pathogen interactions (Leffler *et al.*, 2013). Similarly, disease resistance (R) genes as well as self-incompatibility (S) genes have been found to be under balancing selection in plants (Takebayashi *et al.*, 2003; Roux *et al.*, 2013; Karasov *et al.*, 2014). A recent study found that the genes under balancing selection are responsible for environmental adaptation, and that the distribution of divergent alleles of the same species are correlated with divergent niches (Wu *et al.*, 2017). The genes we have identified in *Populus* encompass all of these functions, including mating compatibility, development, and resistance to biotic and abiotic stress (Table S8). Because the sampled individuals of each species cover only a portion of its total distributional range, we were unable to evaluate whether divergent alleles were correlated with different habitats. It therefore remains unknown whether these genes may have contributed to the widespread distributions of each species. Future studies should use more extensive geographic sampling to uncover the correlations of the divergent alleles with habitat, thereby providing possible mechanisms for local adaptation (Wu *et al.*, 2017). Overall, our population genomic data across the long-term diversification history of the genus identified numerous genes that were likely to be influenced by balancing selection, which would not have been detected by traditional forward or reverse genetic approaches. These approaches should be widely employed in the future to reveal how these divergent alleles of the same species at the same locus contribute to functional adaptation and how these divergent alleles are maintained over vast evolutionary distances.

Acknowledgements

This research was supported by National Natural Science Foundation of China (31590821, 31561123001, 31922061, 31500502, 41871044), National Key Research and Development Program of China (2017YFC0505203, 2016YFD0600101), National Key Project for Basic Research (2012CB114504), National Science Foundation grants (DEB-1542599, NSF-1542509, ISO-1542479, 1542486), and Fundamental Research Funds for the Central Universities (2018CDDY-S02-SCU, SCU2019D013).

Author contributions

TM, JL and MW planned and designed the research. LZ, JL and MW conducted fieldwork. MW, ZZ, ML, DW and XZ analysed the data. ZX designed the phylogenetic analyses. MW, TM and JL wrote the manuscript. KK-R, LBS, SPD, MSO and TY revised the manuscript. MW and LZ contributed equally to this work.

ORCID

Stephen P. DiFazio https://orcid.org/0000-0003-4077-1590
Jianquan Liu https://orcid.org/0000-0002-4237-7418
Tao Ma https://orcid.org/0000-0002-7094-6868
Matthew S. Olson https://orcid.org/0000-0002-0798-145X
Mingcheng Wang https://orcid.org/0000-0002-3631-9174
Zhenxiang Xi https://orcid.org/0000-0002-2851-5474
Zhiyang Zhang https://orcid.org/0000-0002-9466-9439

References

Arnold ML. 2006. *Evolution through genetic exchange*. Oxford, UK: Oxford University Press.

Baute GJ, Owens GL, Bock DG, Rieseberg LH. 2016. Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild *Helianthus* diversity, genetic structure, and interspecies gene flow. *American Journal of Botany* **103**: 2170–2177.

Becerra C, Jahrmann T, Puigdomènech P, Vicient CM. 2004. Ankyrin repeat-containing proteins in *Arabidopsis*: characterization of a novel and abundant group of genes coding ankyrin-transmembrane proteins. *Gene* **340**: 111–121.

BIG Data Center Members. 2019. Database resources of the BIG Data Center in 2019. *Nucleic Acids Research* **47**: D8–D14.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**: 1372–1373.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**: 459–471.

Busing FM, Meijer E, Van der Leeden R. 1999. Delete-m jackknife for unequal m. *Statistics and Computing* **9**: 3–8.

Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Boerjan W. 2005. Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. *Theoretical and Applied Genetics* **111**: 1440–1456.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**: e64.

Chen J, Burke JJ, Velten J, Xin Z. 2006. FtsH11 protease plays a critical role in Arabidopsis thermotolerance. *The Plant Journal* **48**: 73–84.

Cheung AY, Niroomand S, Zou Y, Wu HM. 2010. A transmembrane formin nucleates subapical actin assembly and controls tip-focused growth in pollen tubes. *Proceedings of the National Academy of Sciences, USA* **107**: 16390–16395.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**: 80–92.

Coyne JA, Orr HA. 2004. *Speciation*. Sunderland, MA, USA: Sinauer.

Crowl AA, Manos PS, McVay JD, Lemmon AR, Lemmon EM, Hipp AL. 2019. Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). *New Phytologist*. doi: 10.1111/nph.15842.

Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, 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 Research* **24**: 1274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Delph LF, Kelly JK. 2014. On the importance of balancing selection in plants. *New Phytologist* **201**: 45–56.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* **24**: 332–340.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Du J, Wang X, Zhang M, Tian D, Yang Y. 2007. Unique nucleotide polymorphism of ankyrin gene cluster in *Arabidopsis*. *Journal of Genetics* **86**: 27–35.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**: 2239–2252.

Eaton DA, Hipp AL, González-Rodríguez A, Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* **69**: 2587–2601.

Eckenwalder JE. 1996. *Systematics and evolution of* Populus. National Research Council of Canada. Ottawa, ON, Canada: NRC Research Press.

Feng S, Ru D, Sun Y, Mao K, Milne R, Liu J. 2019. Trans-lineage polymorphism and non-bifurcating diversification of the genus *Picea*. *New Phytologist* **221**: 576–587.

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Molecular Ecology* **24**: 3529–3545.

Fitz Gerald JN, Hui PS, Berger F. 2009. Polycomb group-dependent imprinting of the actin regulator *AtFH5* regulates morphogenesis in *Arabidopsis thaliana*. *Development* **136**: 3399–3404.

Fu A, He Z, Cho HS, Lima A, Buchanan BB, Luan S. 2007. A chloroplast cyclophilin functions in the assembly and maintenance of photosystem II in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **104**: 15947–15952.

Gao Z, Przeworski M, Sella G. 2015. Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution* **69**: 431–446.

Geraldes A, Hefer CA, Capron A, Kolosova N, Martinez-Nuñez F, Soolanayakanahally RY, Stanton B, Guy RD, Mansfield SD, Douglas CJ *et al.* 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Molecular Ecology* **24**: 3243–3256.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y *et al.* 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–722.

Guerrero RF, Hahn MW. 2018. Quantifying the risk of hemiplasy in phylogenetic inference. *Proceedings of the National Academy of Sciences, USA* **115**: 12787–12792.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: e1000695.

Hamzeh M, Dayanandan S. 2004. Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast trnT-trnF region and nuclear rDNA. *American Journal of Botany* **91**: 1398–1408.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics* **10**: 551.

Huerta-Cepas J, Dopazo J, Gabaldon T. 2010. ETE: a python environment for tree exploration. *BMC Bioinformatics* **11**: 24.

Ingouff M, Fitz Gerald JN, Guerin C, Robert H, Sorensen MB, Van Damme D, Geelen D, Blanchoin L, Berger F. 2005. Plant formin AtFH5 is an evolutionarily conserved actin nucleator involved in cytokinesis. *Nature Cell Biology* **7**: 374–380.

Ingvarson PK. 2010. Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in *Populus*. In: Jansson S, Bhalerao RP, Groover AT, eds. *Genetics and genomics of* Populus. New York, NY, USA: Springer, 91–112.

Isebrands J, Richardson J, eds. 2014. *Poplars and willows: trees for society and the environment*. Boston, MA, USA: FAO&CABI.

Jansson S, Douglas CJ. 2007. *Populus*: a model system for plant biology. *Annual Review of Plant Biology* **58**: 435–458.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT *et al.* 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320–1331.

Kamau E, Charlesworth B, Charlesworth D. 2007. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. *Genetics* **176**: 2357–2369.

Kamau E, Charlesworth D. 2005. Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant *Arabidopsis lyrata*. *Current Biology* **15**: 1773–1778.

Kang NY, Cho C, Kim NY, Kim J. 2012. Cytokinin receptor-dependent and receptor-independent pathways in the dehydration response of *Arabidopsis thaliana*. *Journal of Plant Physiology* **169**: 1382–1391.

Karasov TL, Kniskern JM, Gao L, DeYoung BJ, Ding J, Dubiella U, Lastra RO, Nallu S, Roux F, Innes RW *et al.* 2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**: 436–440.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**: 487–493.

Klein J, Sato A, Nikolaidis N. 2007. MHC, TSP, and the origin of species: from immunogenetics to evolutionary genetics. *Annual Review of Genetics* **41**: 281–304.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**: 356.

Kurdyukov S, Faust A, Nawrath C, Bär S, Voisin D, Efremova N, Franke R, Schreiber L, Saedler H, Metraux J-P *et al.* 2006. The Epidermis-specific extracellular BODYGUARD controls cuticle development and morphogenesis in *Arabidopsis*. *Plant Cell* **18**: 321–339.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. 2001. Self-incompatibility in the genus *Arabidopsis*: characterization of the *S* locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. *Plant Cell* **13**: 627–643.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N *et al.* 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**: 371–375.

Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology* **63**: 17–30.

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G *et al.* 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**: 1578–1582.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178–2189.

Liu C, Zhang Y, Ren H. 2018. Actin polymerization mediated by AtFH5 directs the polarity establishment and vesicle trafficking for pollen germination in *Arabidopsis*. *Molecular Plant* **11**: 1389–1399.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* **10**: 302.

Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J *et al.* 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proceedings of the National Academy of Sciences, USA* **114**: E7282–E7290.

Liu X, Wang Z, Shao W, Ye Z, Zhang J. 2017. Phylogenetic and taxonomic status analyses of the Abaso section from multiple nuclear genes and plastid fragments reveal new insights into the North America origin of *Populus* (Salicaceae). *Frontiers in Plant Science* **7**: 2022.

Liu Y, Li D, Zhang Q, Song C, Zhong C, Zhang X, Wang Y, Yao X, Wang Z, Zeng S *et al.* 2017. Rapid radiations of both kiwifruit hybrid lineages and their parents shed light on a two-layer mode of species diversification. *New Phytologist* **215**: 877–890.

Lu H, Rate DN, Song JT, Greenberg JT. 2003. ACD6, a novel ankyrin protein, is a regulator and an effector of salicylic acid signaling in the *Arabidopsis* defense response. *Plant Cell* **15**: 2408–2420.

Ma J, Wan D, Duan B, Bai X, Bai Q, Chen N, Ma T. 2019. Genome sequence and genetic transformation of a widely distributed and cultivated poplar. *Plant Biotechnology Journal* **17**: 451–460.

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. *Nature Communications* **4**: e2797.

Ma T, Wang K, Hu Q, Xi Z, Wan D, Wang Q, Feng J, Jiang D, Ahani H, Abbott RJ *et al.* 2018. Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proceedings of the National Academy of Sciences, USA* **115**: E236–E243.

Mayr E. 1966. *Animal species and evolution*. Cambridge, MA, USA: Belknap Press of Harvard University Press.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The Genome Analysis Toolkit: a Map Reduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Meirmans PG, Lamothe M, Gros-Louis MC, Khasa D, Perinet P, Bousquet J, Isabel N. 2010. Complex patterns of hybridization between exotic and native North American poplar species. *American Journal of Botany* **97**: 1688–1697.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**: i541–i548.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM *et al.* 2016. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**: 1077–1082.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**: 568–583.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genetics* **2**: 2074–2093.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* **14**: e1002379.

Qin L, Li Y, Li D, Xu W, Zheng Y, Li X. 2014. *Arabidopsis* drought-induced protein Di19-3 participates in plant response to drought and high salinity stresses. *Plant Molecular Biology* **86**: 609–625.

Rajora OP, Dancik BP. 1992. Genetic characterization and relationships of *Populus alba*, *P. tremula*, and *P.* × *canescens*, and their clones. *Theoretical and Applied Genetics* **84**: 291–298.

Roux C, Pauwels M, Ruggiero MV, Charlesworth D, Castric V, Vekemans X. 2013. Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. *Molecular Biology and Evolution* **30**: 435–447.

Schluter D. 2000. *The ecology of adaptive radiation*. Oxford, UK: OUP.

Schluter D. 2001. Ecology and the origin of species. *Trends in Ecology & Evolution* **16**: 372–380.

Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, Moyse J, Ross S, Gamble K, Sella G *et al.* 2012. The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences, USA* **109**: 18493–18498.

Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA *et al.* 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist* **196**: 713–725.

Solís-Lemus C, Yang M, Ané C. 2016. Inconsistency of species tree methods under gene flow. *Systematic Biology* **65**: 843–851.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Suarez-Gonzalez A, Hefer CA, Christe C, Corea O, Lexer C, Cronk QC, Douglas CJ. 2016. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). *Molecular Ecology* **25**: 2427–2442.

Suarez-Gonzalez A, Hefer CA, Lexer C, Douglas CJ, Cronk QC. 2018. Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P. trichocarpa*. *New Phytologist* **217**: 416–427.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**: 1569–1571.

Sun Y, Abbott RJ, Lu Z, Mao K, Zhang L, Wang X, Ru D, Liu J. 2018. Reticulate evolution within a spruce (*Picea*) species complex revealed by population genomic analysis. *Evolution* **72**: 2669–2881.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Takebayashi N, Brewer PB, Newbigin ED, Uyenoyama MK. 2003. Patterns of variation within self-incompatibility loci. *Molecular Biology and Evolution* **20**: 1778–1794.

Tuskan GA, 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 der Does D, Boutrot F, Engelsdorf T, Rhodes J, McKenna JF, Vernhettes S, Koevoets I, Tintor N, Veerabagu M, Miedes E *et al.* 2017. The Arabidopsis leucine-rich repeat receptor kinase MIK2/LRR-KISS connects cell wall integrity sensing, root growth and response to abiotic and biotic stresses. *PLoS Genetics* **13**: e1006832.

Waduwara-Jayabahu I, Oppermann Y, Wirtz M, Hull ZT, Schoor S, Plotnikov AN, Hell R, Sauter M, Moffatt BA. 2012. Recycling of methylthioadenosine is essential for normal vascular development and reproduction in *Arabidopsis*. *Plant Physiology* **158**: 1728–1744.

Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Molecular Biology and Evolution* **33**: 1754–1767.

Wang T, Liang L, Xue Y, Jia P, Chen W, Zhang M, Wang Y, Li H, Yang W. 2016. A receptor heteromer mediates the male perception of female attractants in plants. *Nature* **531**: 241–244.

Wang Z, Du S, Dayanandan S, Wang D, Zeng Y, Zhang J. 2014. Phylogeny reconstruction and hybrid analysis of *Populus* (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. *PLoS ONE* **9**: e103645.

Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf E, Allan GJ, DiFazio SP, Potts BM *et al.* 2006. A framework for

community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7: 510–523.

Wu CI. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127**: 429–435.

Wu D, Ding X, Wang S, Wójcik JM, Zhang Y, Tokarska M, Li Y, Wang M, Faruque O, Nielsen R *et al.* 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature Ecology & Evolution* **2**: 1139–1145.

Wu Q, Han T, Chen X, Chen J, Zou Y, Li Z, Guo Y. 2017. Long-term balancing selection contributes to adaptation in *Arabidopsis* and its relatives. *Genome Biology* **18**: 217.

Yang W, Wang K, Zhang J, Ma J, Liu J, Ma T. 2017. The draft genome sequence of a desert tree *Populus pruinosa*. *GigaScience* **6**: 1–7.

Zhang L, Xi Z, Wang M, Guo X, Ma T. 2018. Plastome phylogeny and lineage diversification of Salicaceae with focus on poplars and willows. *Ecology and Evolution* **8**: 7817–7823.

Zheng H, Fan L, Milne RI, Zhang L, Wang Y, Mao K. 2017. Species delimitation and lineage separation history of a species complex of aspens in China. *Frontiers in Plant Science* **8**: 375.

Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evans LM, Schmutz J, Jenkins JW, Yan J, Tuskan GA *et al.* 2018. Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics* **293**: 1437–1452.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Comparison of tree topologies estimated using the concatenation method RAxML from the whole-genome SNVs and SNVs at four-fold degenerate sites.

Fig. S2 Plots of the first two principal components for SNV data from 80 individuals from six sections of the genus *Populus*.

Fig. S3 The number of SNVs shared among the four major clades revealed by phylogenetic analyses and the number of SNVs fixed between clades.

Fig. S4 Distribution of mean bootstrap support values of individual gene trees and all possible topologies of gene trees with their observed frequencies.

Fig. S5 The percentage of gene trees supporting the split of the four major clades and overlapped ML trees based on C_{12} dataset of 5305 single-copy genes.

Fig. S6 Tree topologies estimated using the concatenation method of RAxML from the C_{12}, C_3 and CDS datasets of 5305 single-copy genes.

Fig. S7 Tree topologies estimated using the coalescent method ASTRAL from the C_{12}, C_3 and CDS datasets of 5305 single-copy genes.

Fig. S8 Tree topologies estimated using the coalescent method MP-EST from the C_{12}, C_3 and CDS datasets of 5305 single-copy genes.

Fig. S9 Tree topologies estimated using the coalescent method ASTRAL after collapsing all branches with support values > 50% in the gene trees and tree topologies estimated based on gene trees with mean bootstrap value ≥ 50.

Fig. S10 A maximum-likelihood chloroplast DNA phylogeny based on 77 plastome protein-coding genes.

Fig. S11 The length distribution and frequency distribution of the shared IBD blocks.

Fig. S12 The length distribution of the shared IBD blocks identified between *P.* × *canescens* and its parents, *P. alba* and *P. tremula* (a) and between *P. wulianensis*, *P. ningshanica* and their probable parents, *P. adenopoda* and *P. davidiana* (b).

Fig. S13 The negative correlations between the IBD block length and the population-scaled recombination rate.

Fig. S14 The positive correlations between the *D* statistics and the length of IBD blocks shared between species in 'ATL' clade and sect. *Populus*.

Fig. S15 *D* statistics from ABBA–BABA tests for two different tree topologies.

Fig. S16 The distribution of average *D* statistics over 100 kb nonoverlapping windows for three different tree topologies.

Fig. S17 Joint allele frequency spectra between the six species used in the analyses of trans-species polymorphisms and balancing selection.

Fig. S18 The number of SNVs shared among the six species for SNV datasets with maximum missing genotype rate of 20% (a) and 50% (b) in all six species.

Fig. S19 The candidate regions in the 45 candidate genes under balancing selection clustered by allele rather than species.

Fig. S20 Hypothetical crossing of two species indicated by red and blue, respectively, proposed to have led to the chloroplast captures (small circles).

Table S1 Summary statistics of genome resequencing data for 80 individuals.

Table S2 Statistics of SNVs classified by their physical locations in the *Populus trichocarpa* genome.

Table S3 Statistics of IBD blocks shared between species within ATL clade, sect. *Populus* and sect. *Turanga*, and the relationship between IBD block length and recombination rate.

Table S4 Detail information of ABBA–BABA analyses for all comparisons. This table was uploaded as a separate file.

Table S5 The sources and summary statistics of 72 published genome sequences used in the analyses of trans-specific polymorphisms and balancing selection.

Table S6 Pairwise F_{ST} values (above the diagonal) and numbers of shared SNVs (below the diagonal) between each pair of species.

Table S7 Statistics of SNVs that segregate in all six species.

Table S8 Information on candidate genes and trans-specific polymorphisms under long-term balancing selection. This table was uploaded as a separate file.

Table S9 GO term enrichment of 45 candidate genes under balancing selection.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.