Demographic History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to Introgression across *Populus* Genome

Shuyu Liu,[†,1] Lei Zhang,[†,1] Yupeng Sang,[1] Qiang Lai,[1] Xinxin Zhang,[1] Changfu Jia,[1] Zhiqin Long,[1] Jiali Wu,[1] Tao Ma ⓘ,[1] Kangshan Mao,[1] Nathaniel R. Street ⓘ,[2] Pär K. Ingvarsson ⓘ,[3] Jianquan Liu,[1] and Jing Wang ⓘ*,[1]

[1]Key Laboratory for Bio-Resources and Eco-Environment of Ministry of Education, College of Life Science & State Key Lab of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu, China
[2]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden
[3]Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden
[†]These authors contributed equally to this work.
*Corresponding author: E-mail: wangjing2019@scu.edu.cn.
Associate editor: Yuseob Kim

Abstract

Hybridization and resulting introgression are important processes shaping the tree of life and appear to be far more common than previously thought. However, how the genome evolution was shaped by various genetic and evolutionary forces after hybridization remains unresolved. Here we used whole-genome resequencing data of 227 individuals from multiple widespread *Populus* species to characterize their contemporary patterns of hybridization and to quantify genomic signatures of past introgression. We observe a high frequency of contemporary hybridization and confirm that multiple previously ambiguous species are in fact F₁ hybrids. Seven species were identified, which experienced different demographic histories that resulted in strikingly varied efficacy of selection and burdens of deleterious mutations. Frequent past introgression has been found to be a pervasive feature throughout the speciation of these *Populus* species. The retained introgressed regions, more generally, tend to contain reduced genetic load and to be located in regions of high recombination. We also find that in pairs of species with substantial differences in effective population size, introgressed regions are inferred to have undergone selective sweeps at greater than expected frequencies in the species with lower effective population size, suggesting that introgression likely have higher potential to provide beneficial variation for species with small populations. Our results, therefore, illustrate that demography and recombination have interplayed with both positive and negative selection in determining the genomic evolution after hybridization.

Key words: hybridization, genetic load, natural selection, adaptive introgression, *Populus*.

Introduction

Hybridization is a prevalent and important evolutionary process along the tree of life (Rieseberg and Carney 1998). Multiple evolutionary outcomes can arise through hybridization, for instance inducing or reversing speciation, formation of hybrid zones, transfer of small amounts of genetic material via introgression, and occasionally hybrid speciation (Abbott et al. 2013; Runemark et al. 2019). Despite the important role of hybridization in evolution, it remains to be seen how species remain distinct in most cases and which forces govern the evolutionary fate of hybrid genomes (Pease et al. 2016; Moran et al. 2021).

Following hybridization, the introgressed blocks are broken up by recombination in successive generations, and the mosaic genomes of the introgressed lineages are further shaped by the interaction of recombination and selection (Schumer et al. 2018). In regions of high recombination, neutral or mildly adaptive introgressed alleles are more likely to recombine away from deleterious alleles and escape the effects of negative linked selection, whereas a stronger localized reduction in introgression is expected within low recombination regions owing to the increased linkage between introgressed loci and neighboring selected variants (Schumer et al. 2018; Martin et al. 2019).

On the other hand, as most hybridizing species are likely to differ in their demographic histories and population sizes, the pattern and magnitude of deleterious variations and genetic load can vary dramatically among these species (Hough et al. 2013; Xue et al. 2015; Laenen et al. 2018). When there is gene flow from a donor species with a smaller effective population size and reduced efficiency of purifying selection, the recipient species may suffer from increased genetic load owing to the introduction of weakly deleterious alleles (Harris and Nielsen 2016; Juric et al. 2016). In contrast, gene flow from a donor

species with larger effective population size could alleviate the genetic load of the recipient species (Carlson et al. 2014; Edelman and Mallet 2021). Therefore, selection on deleterious introgressed variants could have a profound impact on the genome-wide patterns of introgression (Kim et al. 2018). In addition, although hybridization between divergently adapted species is largely deleterious, introgression can occasionally introduce adaptive variants that are favored by positive selection and can spread rapidly in the recipient species, a process known as adaptive introgression (Suarez-Gonzalez et al. 2018; Leroy et al. 2020). Recent studies that take a whole-genome view suggest that adaptive introgression is likely common, and can generate unique patterns of genetic diversity against genomic background (Setter et al. 2020).

Most studies have focused on each process and mechanism in isolation (Moran et al. 2021), so the relative contributions of genetic drift, selection against deleterious introgressed variants and adaptive introgression to the genomic patterns of introgression remain elusive. Additionally, multiple demographic and selective processes simultaneously operate on the evolutionary dynamics of introgression, which can potentially generate interference effects (Edelman and Mallet 2021). Therefore, it is necessary to systematically investigate how historical demographic events have influenced genetic diversity and load of deleterious mutations of different species in the first place, and how the distinct evolutionary forces further determine the genomic landscape of introgression. Assessing this issue is crucial for our understanding of the evolutionary consequences and also the potential of hybridization to produce novel adaptive variation in a changing world (Nelson et al. 2021).

Forest trees are a relevant target species for explorations of how distinct evolutionary forces interplayed to shape the distribution and evolution of neutral, deleterious, and adaptive diversity, as they are largely undomesticated with high levels of genetic diversity (Neale and Kremer 2011; Isabel et al. 2020). In addition, most tree species are outbreeding, and interspecific gene flow is particularly widespread and valuable for long-lived trees. Among forest tree species, Populus provides a well-established model system because of its economic and ecological importance and favorable attributes such as small genome size, ease of genetic transformation, availability of several well-assembled genomes and frequent interspecific hybridization (Jansson and Douglas 2007). Species of this genus are mostly deciduous, obligated outcrossing, and have widespread geographical distribution throughout the Northern Hemisphere (Eckenwalder 1996). According to the morphological traits, the Populus species are taxonomically divided into six sections: Abaso, Turanga, Leucoides, Aigeiros, Tacamahaca, Populus (Eckenwalder 1996). Interspecific hybridization is rare between sections (except for Aigeiros and Tacamahaca), but is frequent within sections (Wang et al. 2020). The frequently occurring interspecific hybrids in Populus have perplexed taxonomists, with 22 to 85 acknowledged species being classified. However, numerous described species were doubted as being varieties or hybrids of the other highly diagnostic species (Dickmann and Kuzovkina 2014).

In this study, we generated whole-genome resequencing data from 227 individuals representing seven diagnostic species (P. adenopoda, P. alba, P. davidiana, P. qiongdaoensis, P. rotundifolia, P. tremula, P. tremuloides) from portions of their native ranges, and five ambiguous species with uncertain taxonomic status (P. wulianensis, P. tibetica, P. canescens, P. ningshanica, P. tomentosa). These species are all from the section Populus of the genus Populus. We used these data to first examine patterns of genomic divergence and determine species identity and hybrid status, as the accurate delimitation of species is a fundamental first step that underlies all subsequent analyses. We then examined and compared the demographic histories of these species and investigated how distinct historical demography influenced the efficacy of selection and deleterious mutation load among species. Finally, we characterized and quantified the signatures and the genomic extent of introgression among the diagnostic species, and clarified how natural selection shaped the genome-wide patterns of introgression in terms of transfer and removal of adaptive and deleterious variants between species.

Results

Species Phylogeny, Population Structure, and Frequent Contemporary Hybridization

We collected and performed whole-genome resequencing for 227 individuals sampled from seven diagnostic and five ambiguous Populus species, with an average depth of 32.2× for each individual (supplementary fig. S1 and table S1, Supplementary Material online). Paired-end reads were mapped against P. tremula reference genome (version 2.2, https://popgenie.org) (Lin et al. 2018), and after stringent filtering standards, a final set of 8,966,513 high-quality single-nucleotide polymorphisms (SNPs) were identified within and across species. To decipher the genetic relationships among species and individuals, we constructed neighbor-joining (NJ) phylogenetic trees, performed principal component analysis (PCA), and conducted individual ancestry inference (fig. 1A–C). Our results clearly confirmed seven species: P. adenopoda, P. qiongdaoensis, P. alba, P. tremuloides, P. tremula, P. davidiana, and P. rotundifolia. The first two principal components (PC1 and PC2) explained 27% of the genome covariance and separated the seven species into different groups (fig. 1B). The optimal K value revealed by fastSTRUCTURE (Raj et al. 2014) is seven (supplementary fig. S2, Supplementary Material online), which grouped all individuals into seven genetic clusters that recapitulated the seven species represented in the PCA (fig. 1C).

Despite the clear genetic clustering of the seven species, the five ambiguous species were interspecific hybrids of these seven species. The inference of the hybrid status for these species was supported by various analyses, where these hybrids were placed intermediate to the main clusters in the PCA (fig. 1B and supplementary fig. S3, Supplementary Material online) and were found to be admixed in ancestry inference with fastSTRUCTURE (fig. 1C and supplementary figs. S4 and S5, Supplementary Material online). To determine the maternal and paternal species for each putative hybrid,
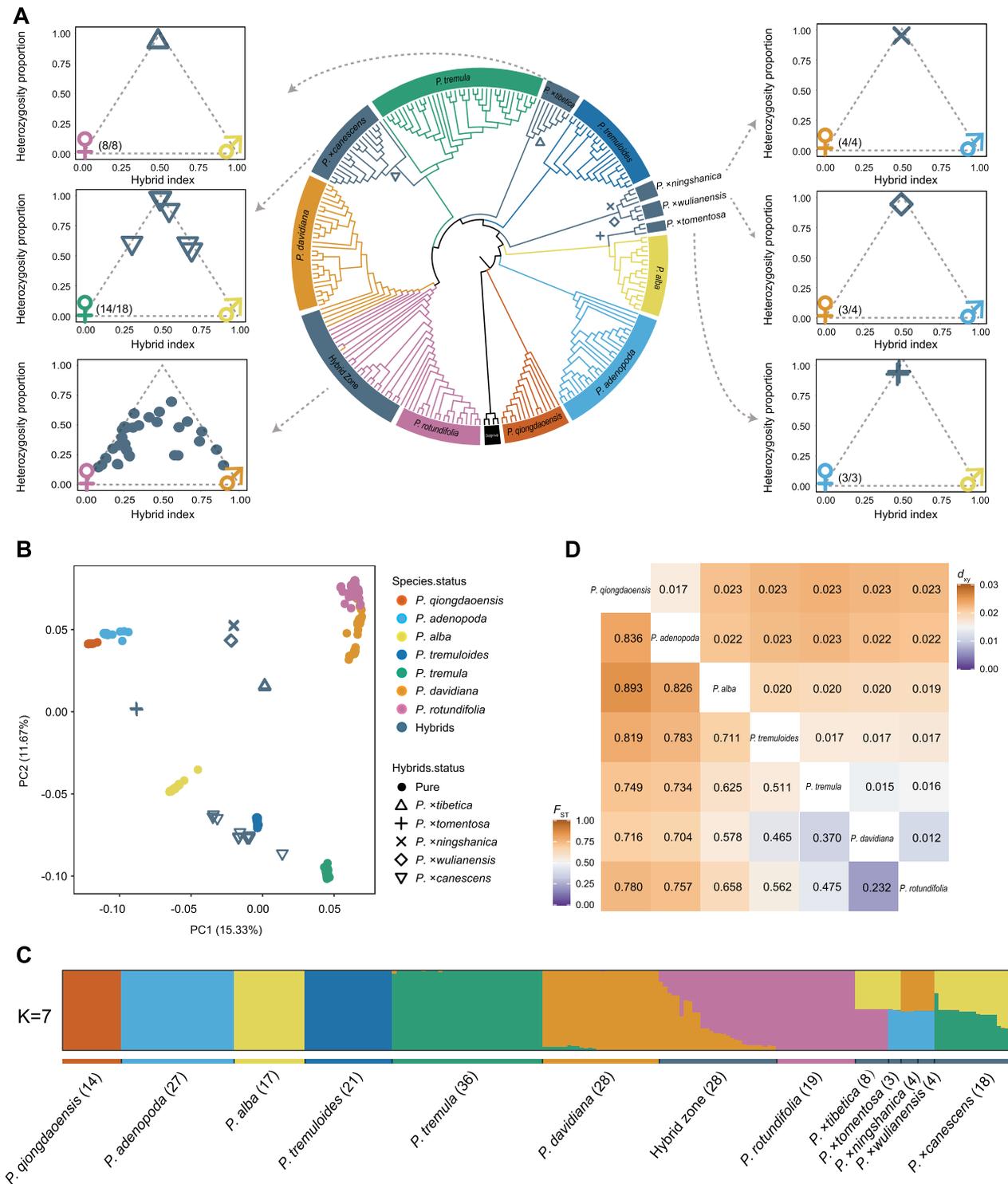
Fig. 1. Phylogenetic relationship and population structure of *Populus* species. (*A*) Neighbor-joining (NJ) phylogenetic tree of all 227 *Populus* samples constructed using whole-genome pruned SNP data, with *P. trichocarpa* and *P. euphratica* being used as outgroups (black). Different branch colors reflect different species. Gray arrows point to triangle plots for five ambiguous species (*P. ×wulianensis*, *P. ×tibetica*, *P. ×canescens*, *P. ×ningshanica*, *P. ×tomentosa*) and individuals in *P. davidiana - P. rotundifolia* hybrid zone. These triangle plots show hybrid index and heterozygosity proportion for putative hybrid individuals. The colors of the maternal and paternal species are consistent with the NJ tree, with number in the parentheses meaning the ratio of maternal species being supported among the hybrids. (*B*) Principal component analysis of 227 *Populus* samples, with the proportion of the variance explained being 15.33% for PC1 and 11.67% for PC2, respectively. (*C*) Population structure by fastSTRUCTURE analysis for the optimal $K=7$. Each colored bar represents one individual, and colored segments represent proportions of ancestral components. The number of individuals in each lineage is shown at the bottom. (*D*) Matrix of relative (F_{ST}) and absolute (d_{xy}) divergence for pairwise species comparisons (Upper: d_{xy}; Lower: F_{ST}).

we constructed the phylogenetic relationship of these individuals using the maternally inherited chloroplast and mitochondrial genomes (supplementary figs. S6 and S7, Supplementary Material online). Based on sites that are fixed for different alleles in the parental species, we further estimated the heterozygosity and hybrid indices (fig. 1A and supplementary table S2, Supplementary Material online). The results showed that P. ×tomentosa, which is widely used as an ornamental and street tree in Northern and Central China, were likely spontaneous first-generation (F_1) hybrids of P. adenopoda (maternal) and P. alba (paternal). Populus ×tibetica, which was sampled in Shannan prefecture of Tibet, were F_1 hybrids of P. rotundifolia (maternal) and P. alba (paternal). Populus ×wulianensis and P. ×ningshanica were both F_1 hybrids of P. davidiana (maternal) and P. adenopoda (paternal), although with different population groups of P. davidiana being the maternal species (supplementary fig. S4, Supplementary Material online). Populus ×canescens, well known as gray poplar, was found to be natural interspecific hybrids of P. tremula and P. alba. The individuals of P. ×canescens were found to be mostly F_1 hybrids but also have backcrossed hybrids with either of the two parental species (fig. 1A and supplementary table S2, Supplementary Material online). Lastly, we identified a natural hybrid zone between P. davidiana and P. rotundifolia where a number of individuals with admixed genotypes exist (fig. 1C and supplementary fig. S5, Supplementary Material online).

Genomic Variation, Linkage Disequilibrium, and Demographic History

To evaluate and compare the genetic diversity and demographic histories among the seven species, we removed all hybrids and only kept the 162 individuals of the seven diagnostic species for downstream analysis. We found each of the seven species to be monophyletic and genetically identifiable (supplementary figs. S8 and S9, Supplementary Material online). The level of genetic divergence (F_{ST} and d_{xy}) between pairs of species is in accordance with the phylogenetic relationship among these species (fig. 1D). We next estimated and found markedly different levels of nucleotide diversity (π) among species (fig. 2A and supplementary table S3, Supplementary Material online), although the genome-wide patterns of genetic diversity of these species were found to be highly correlated (supplementary fig. S10, Supplementary Material online). Populus davidiana, P. tremuloides, P. tremula, and P. rotundifolia displayed comparably high levels of sequence diversity, followed by P. alba and P. adenopoda. Populus qiongdaoensis, consistent with its restricted and isolated island distribution (supplementary fig. S1, Supplementary Material online), displayed extremely low sequence diversity compared with other species (fig. 2A). In addition, the number of singleton SNPs per genome and the total number of observed nonreference variants were found to differ greatly among species, being lowest in P. qiongdaoensis, followed by P. adenopoda and P. alba, and higher in other species (fig. 2B and C). Genome-wide linkage disequilibrium (LD) analysis demonstrated that the level of

LD and the decay of pairwise correlation coefficient (r^2) values also varied substantially among these species, with P. qiongdaoensis showing the slowest decay of LD compared with other species with higher genetic diversity (fig. 2D and supplementary fig. S11, Supplementary Material online).

To further explore the demographic histories of these species, we inferred the long-term effective population size (N_e) dynamics for each species using the pairwise sequential Markovian coalescent (PSMC) method (Li and Durbin 2011). The results revealed that the seven Populus species have experienced distinct demographic histories (fig. 2E; and supplementary fig. S12, Supplementary Material online). Both P. qiongdaoensis and P. adenopoda underwent a prolonged and severe decline in N_e since the Pre-Pastonian glacial period (0.8–1.3 Ma). However, compared with P. qiongdaoensis that exhibited severe declines until present-day, P. adenopoda experienced population growth after the last glacial maximum (LGM). Notably, P. alba experienced a different population history compared with all other species, with N_e being highest around 0.05–0.07 Ma followed by a continuous decline to present day. For the other four species, they differed in N_e trajectories after the Riss glaciation (0.1–0.2 Ma), with P. tremuloides experiencing a more dramatic population expansion than other species, which is in accordance with the higher number of singletons (fig. 2B), excess of low frequency variants based on both folded and unfolded site frequency spectrum (SFS) (supplementary fig. S13, Supplementary Material online), as well as more negative values of Tajima's D statistic (supplementary fig. S14, Supplementary Material online) found specifically in this species (Wang et al. 2016).

Comparison of Deleterious Mutation Load across Species

As the seven species contained highly different genomic diversity and experienced different demographic histories, we used several approaches to investigate whether the impact of purifying selection and the accumulation of mutation load varies among these species. First, we estimated the distribution of fitness effects (DFE) of new nonsynonymous mutations for each species using the software DFE-alpha (Keightley and Eyre-Walker 2007). Given that this method explicitly corrects for the effects of nonequilibrium population size changes, it is supposed to account for the differences in demographic histories experienced by different species (supplementary table S4, Supplementary Material online). The strength of purifying selection, which was defined as the product of N_e and the selection coefficient (s), was summarized in four bins, ranging from nearly neutral to strongly deleterious: $0 < N_e s < 1, 1 < N_e s < 10, 10 < N_e s < 100, N_e s > 100$ (fig. 3A and supplementary table S5, Supplementary Material online). As predicted, compared with other species, P. qiongdaoensis showed a significantly elevated proportion of effectively neutral mutations ($N_e s < 1$), followed by P. adenopoda, suggesting that these two species have likely experienced genome-wide relaxed purifying selection against weakly deleterious mutations. However, in contrast to P. adenopoda and other species that showed strong purifying selection against
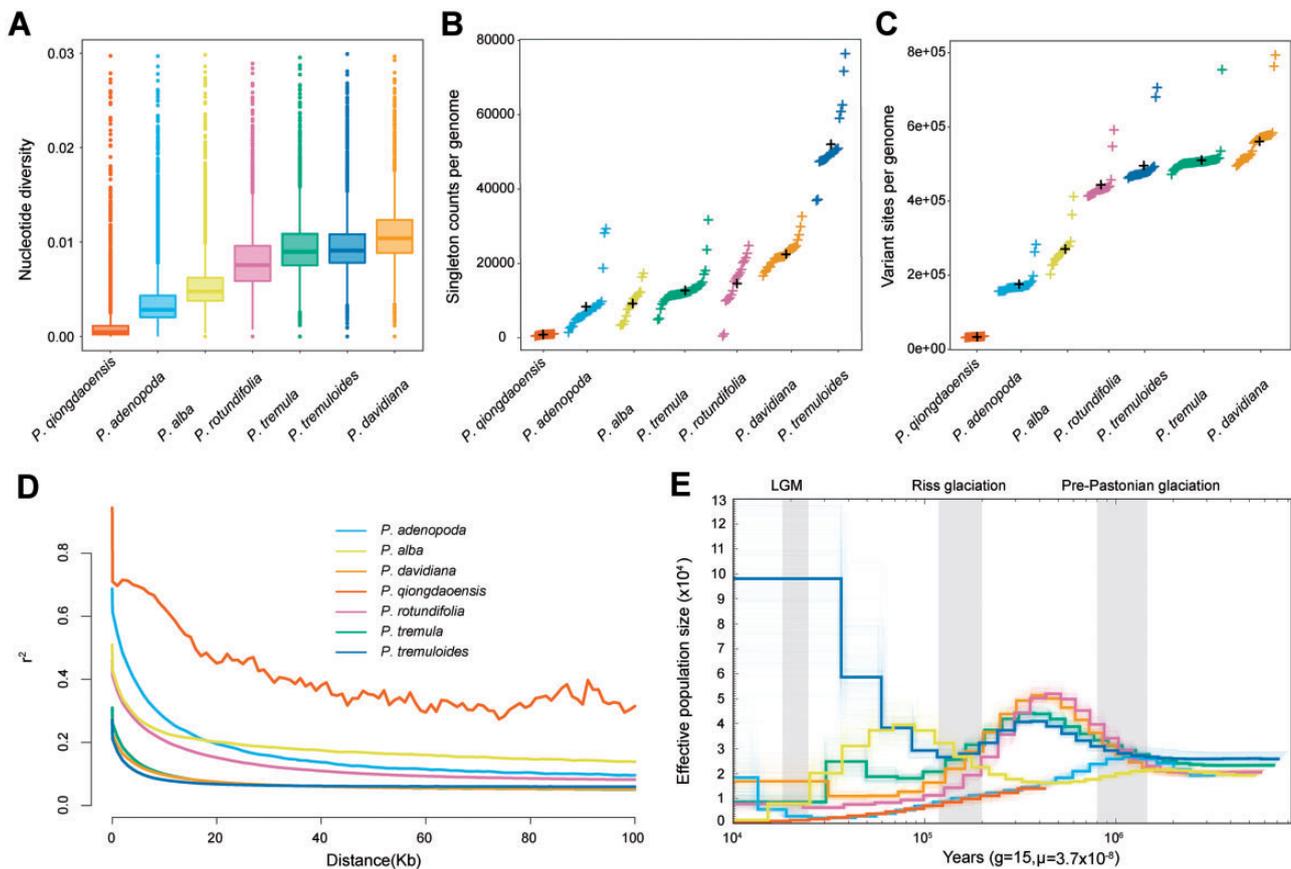
FIG. 2. Genomic features of the seven divergent *Populus* species. Boxplot of nucleotide diversity over 100-kb nonoverlapping windows (*A*), the number of singletons (*B*), and the distribution of total number of observed nonreference variants (*C*) per genome among the seven species (in *B* and *C*, the average estimates for each species were marked by black crosses). (*D*) The decay of LD measured by r^2 for the seven species. (*E*) Demographic history inferred by the PSMC model, the periods of the LGM, Riss glaciation, and Pre-Pastonian glaciation are highlighted with gray vertical bars. For each species, bold line is the median estimate for all 11 selected individuals, whereas faint lines are 110 bootstrap replicates, with ten replicates being conducted for each of the 11 samples.

strongly deleterious nonsynonymous mutations ($N_e s > 100$), *P. qiongdaoensis* still exhibited much weaker strength of selection (fig. 3A). We further estimated the inbreeding coefficient (F_{IS}) for each individual among the seven species, and found that *P. qiongdaoensis* had extremely negative F_{IS} (average = −0.569) compared with other species with F_{IS} being around zero (supplementary fig. S15, Supplementary Material online). The high heterozygosity reflected by both SFS (supplementary fig. S13, Supplementary Material online) and F_{IS} estimates, combined with the low genetic diversity of *P. qiongdaoensis*, indicates that this species may experience prevalent clonal propagation. Therefore, we need to recognize a caveat that the clonal reproduction of *P. qiongdaoensis* violates the assumption of drift–mutation–selection equilibrium, which may cause the poor fit of the demographic model to the data and may quantitatively alter its DFE predictions (supplementary fig. S16, Supplementary Material online).

To estimate the individual mutation load, we examined the ratio of the nucleotide heterozygosity of 0- to 4-fold degenerate sites, which is predicted to be elevated in species with small population size as deleterious mutations at 0-fold sites cannot be efficiently removed (Lynch et al. 1995).

Consistent with this prediction, the ratio of 0- to 4-fold heterozygosity across the seven species showed a strong negative relationship with neutral heterozygosity (Pearson's $r^2 = 0.912$, $P = 0.00082$, fig. 3B), and was significantly elevated in *P. qiongdaoensis* (Kruskal–Wallis, $P < 0.01$), followed by *P. adenopoda*, compared with other species (fig. 3B and supplementary fig. S17, Supplementary Material online).

To further explore and compare the patterns of mutation load carried by different species, we classified coding sequence variants with respect to their effect into four groups using SIFT4G (Vaser et al. 2016): synonymous, tolerated, deleterious, and loss-of-function (LoF). The alleles within each variant were polarized as ancestral or derived using the *P. trichocarpa* and *P. euphratica* as outgroups. In general, we found that most functional variants were species-specific, and relative to neutral sites, a lower proportion of functional variants were shared between species (supplementary fig. S18, Supplementary Material online). At the species level, the absolute numbers of all types of variants are consistent with the levels of genetic diversity among species (supplementary table S6, Supplementary Material online). Nonetheless, we found that the ratios of derived functional
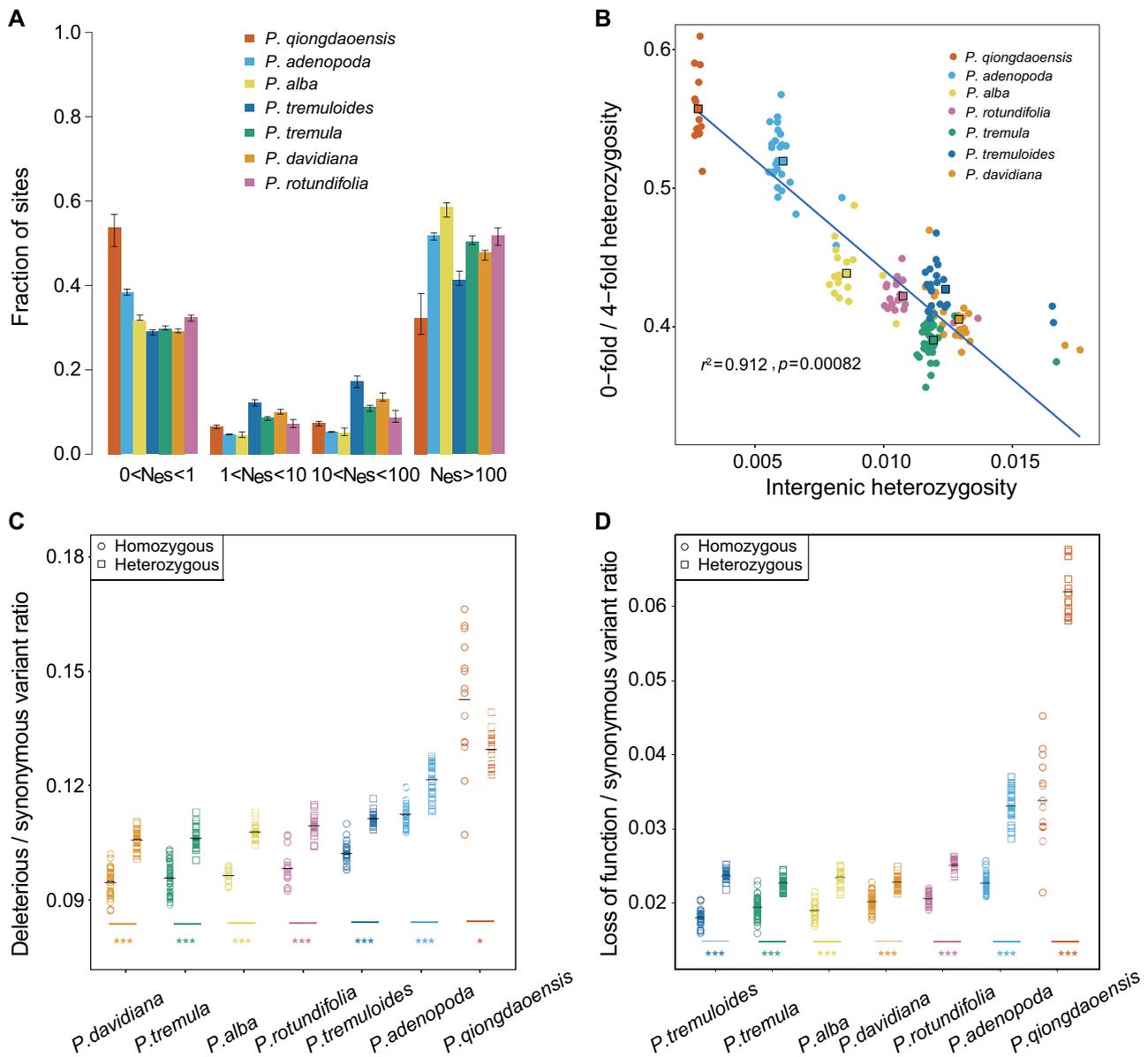
Fig. 3. Assessment of genetic load among the seven species. (A) The DFEs in bins of $N_e s$ for new 0-fold nonsynonymous mutations. Error bars represent 95% bootstrap confidence intervals. (B) The negative relationship between neutral (intergenic) nucleotide heterozygosity and the ratio of heterozygosity at 0-fold relative to 4-fold sites. Each dot represents each individual that is colored by species. The colored squares with black outline represent the average values for each species. The solid blue line denotes the best-fit linear regression line of the average ratio of 0- to 4-fold heterozygosity versus the average intergenic heterozygosity across the seven species. (C, D) Ratio of derived deleterious (C) and LoF (D) variants to synonymous variants in homozygous (circle) and heterozygous (square) tracts per individual. Horizontal bars denote average values. P values for each species were obtained from Kolmogorov–Smirnov tests comparing homozygous and heterozygous tracts, asterisks indicate degree of significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

variants (including tolerated, deleterious, or LoF variants) relative to synonymous variants were much higher in *P. qiongdaoensis*, followed by *P. adenopoda*, than in other species (fig. 3C and D; supplementary fig. S19 and table S7, Supplementary Material online). Given that LoF variants are more likely to be strongly deleterious or lethal (Glémin 2003), they occurred at much higher rates at heterozygous sites compared with homozygous sites in all species (fig. 3D). We observed a much higher ratios of heterozygous derived LoF to synonymous variants in *P. qiongdaoensis* and *P. adenopoda* compared with other species (fig. 3D), indicating that strongly

deleterious or lethal variants are more likely to act recessively and be maintained in a heterozygous state, especially in species with small population size.

Evidence for Frequent Past Introgression

Despite the seven species having reciprocal monophyly, we tested whether the speciation process among these species was accompanied with past introgression. First, we explored the phylogenetic relationships of the seven species using TWISST (Martin and Van Belleghem 2017), which quantifies the frequency of alternative phylogenetic topologies in sliding
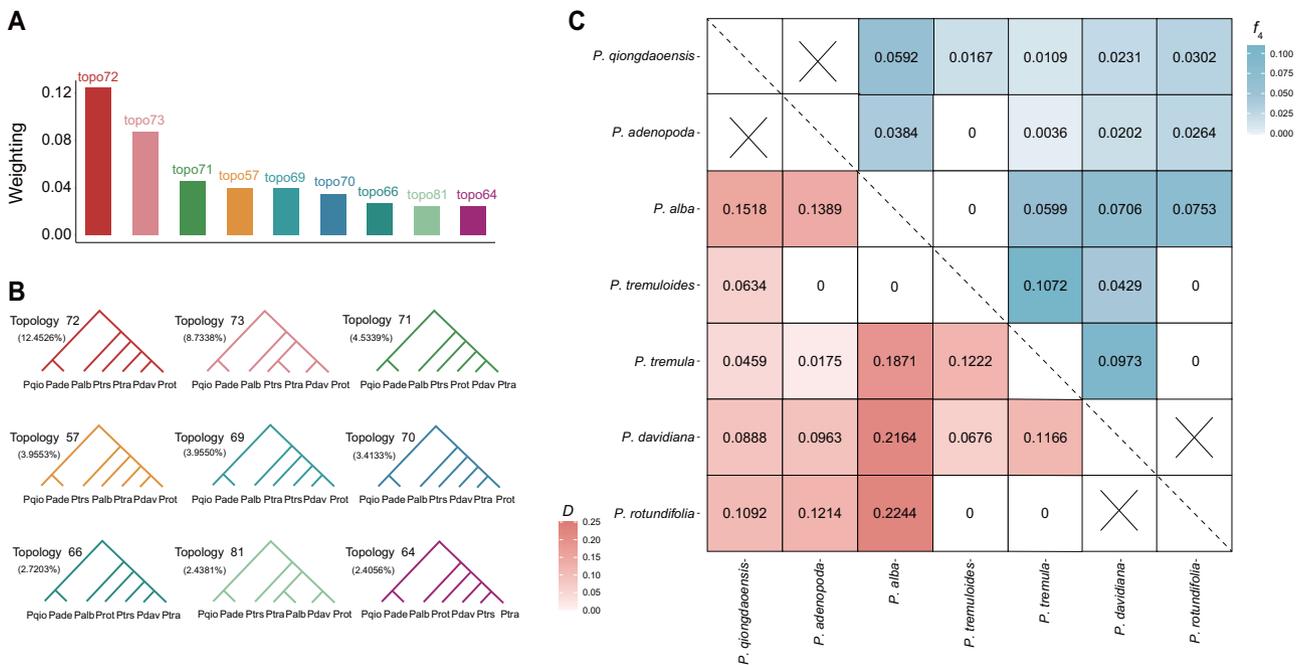
Fig. 4. Recurrent past introgression occurred during the speciation process of *Populus*. (A) The frequency of the top nine topologies reconstructed in 10-kb nonoverlapping windows along the genome. (B) Schematic representation of top nine topologies that were constructed using PhyML with the GTR model. Each topology is colored to match the bar chart on the (A). Abbreviation of species names: *Palb*, P. alba; *Pade*, P. adenopoda; *Pdav*, P. davidiana; *Pqio*, P. qiongdaoensis; *Ptra*, P. tremula; *Ptrs*, P. tremuloides; *Prot*, P. rotundifolia. (C) Heatmap indicating maximum pairwise D (below diagonal) and f_4 (above diagonal) statistics between pairs of species across all combinations of trios. Combinations marked with "x" represents sister species, introgression between these could not be assessed.

windows along the genome. For the sliding window approach, we used both a fixed window size (10 kb) and a fixed number of SNPs (1,000 SNPs), with both showing highly similar results (supplementary fig. S20 and table S8, Supplementary Material online). Among 945 possible species topologies (supplementary table S8, Supplementary Material online), the most common topology (topo72) had a genome-wide frequency of 12.45% (fig. 4A and B), which shared the same topology as the NJ tree constructed over the entire nuclear genome (supplementary fig. S9, Supplementary Material online).

To assess whether the extensive discordances among local topologies are mainly due to incomplete lineage sorting and/or interspecific gene flow, we systematically tested for signatures of introgression by calculating Patterson's D and f_4 admixture ratio (f_4-ratio) statistics (Patterson et al. 2012). The D and f_4-ratio statistics are, respectively, useful to first detect introgression and to further estimate the introgression proportion between species. Both of them measure the excess of shared derived alleles between P3 and either P1 or P2 based on a tree model of four species as (((P1, P2),P3, O) where O is the outgroup. We calculated D and f_4-ratio for all combinations of trios that are compatible with the species topology using *P. trichocarpa* and *P. euphratica* as the outgroup. 34 out of 35 trios had a significant D value at $P < 0.001$ after performing standard block-jackknife procedures (supplementary table S9, Supplementary Material online). Our results thus provide strong evidence of pervasive historical introgression among the seven *Populus* species. After ordering P1 and P2 to let nABBA≥nBABA, D and

f_4-ratio statistics were made to be always positive to reflect the introgression between P2 and P3 for each trio. As the estimates of D and f_4-ratio for the same P2–P3 species pairs varied depending on the different "control" P1 species (supplementary table S9, Supplementary Material online), we used their maximum values in the following analyses to reduce the redundancy in the data by focusing on the overall support for gene flow between P2 and P3 (fig. 4C). Overall, the extent of genome-wide introgression and the admixture proportion estimated by D and f_4-ratio varied widely among species pairs, ranging from ~0.36% between *P. tremula* and *P. adenopoda*, to 10.72% between *P. tremula* and *P. tremuloides* (fig. 4C). We did not include the trios with *P. qiongdaoensis* since the clonality of this species may bias the results, and therefore, in the downstream analysis we only focused on the ten trios without *P. qiongdaoensis* (supplementary table S10, Supplementary Material online). We further examined the relationship between the proportion of introgression estimated by f_4-ratio and the interspecific genetic divergence, and found a significantly negative relationship among the ten introgressing species pairs (supplementary fig. S21, Supplementary Material online).

Identification of Introgressed Regions and Inferring Their Functional Significance

For species pairs with significant D statistics, we further localized introgressed regions by calculating both f_d and f_{dM} statistics (Malinsky et al. 2015; Martin et al. 2015), which have been proven to be more useful to assist in locating
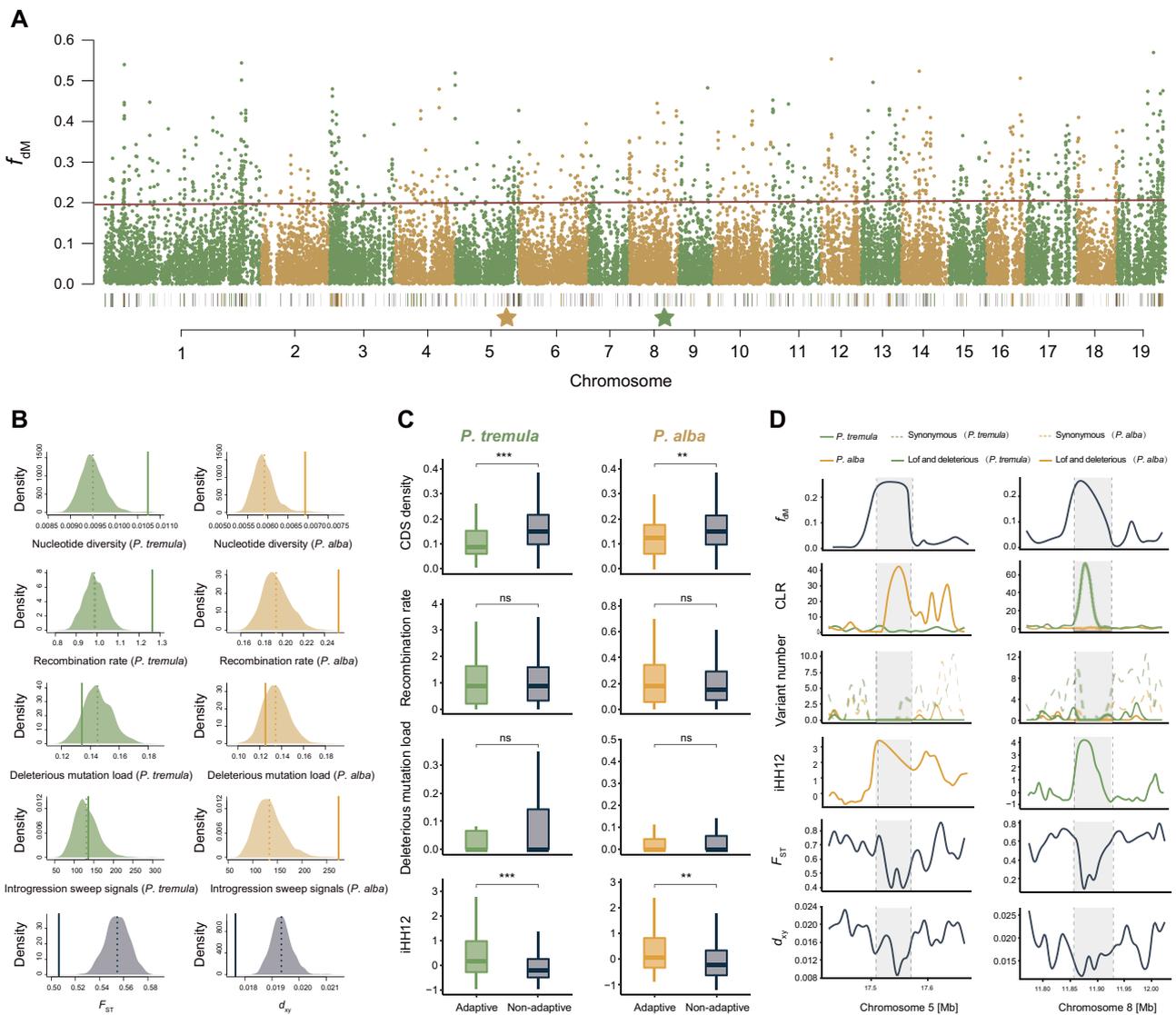
FIG. 5. Genomic characteristics of the putative introgressed regions between *Populus tremula* and *P. alba*. (A) Manhattan plot showing the f_{dM} values across genome. The black bars on bottom indicate the locations of the candidate introgressed regions. The yellow and green bars represent the adaptive introgressed regions in *P. alba* and *P. tremula*, respectively. The location of the two stars corresponds to the adaptive introgressed region shown in (D). (B) Assessing genomic features and patterns of selection acting on candidate introgressed regions. Diagram showing the comparison of nucleotide diversity (π), recombination rate, deleterious mutation load, introgression sweep signals, and genetic divergence (F_{ST} and d_{xy}) between introgressed regions (solid line) and the distribution resulting from the 1,000 randomizations (dashed line: average across randomizations) for *P. tremula* and *P. alba*, respectively. (C) Comparison of coding density, recombination rate, deleterious mutation load, and iHH12 statistics between adaptive and nonadaptive introgressed regions in *P. tremula* and *P. alba*, respectively. Significance between adaptive and nonadaptive introgressed regions within each species was evaluated using Mann–Whitney tests and the significance is indicated by asterisks above the boxplot: ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, $^{ns}P > 0.05$. (D) Magnification of f_{dM}, CLR, number for deleterious and synonymous variants, iHH12 statistic, F_{ST}, and d_{xy} across an adaptive introgression region (gray shading) specifically to *P. alba* and *P. tremula* along chromosomes 5 and 8. Genomic characteristics of the putative introgressed regions for the other nine species pairs can be found on supplementary figure S24, Supplementary Material online.

introgressed loci in small genomic regions compared with the *D* statistics. We used the results of f_{dM} statistic as it is a modified version of f_d and is symmetrically distributed around zero under the null hypothesis of no introgression (supplementary figs. S22 and S23, Supplementary Material online). Introgressed regions were defined as the top f_{dM} windows that summed to the genomic proportion estimated from the f_4-ratio for each trio (supplementary table S10, Supplementary Material online).

With the candidate introgressed regions identified, we assessed genomic characteristics and patterns of selection acting on these regions (fig. 5A and supplementary fig. S24A, Supplementary Material online). As expected for the occurrence of interspecific introgression, the introgressed regions, in general, showed significantly lower interspecific genetic divergence (F_{ST} and d_{xy}) but higher intraspecific nucleotide diversity (π) in both P2 and P3 species than expected at random (fig. 5B and table 1). In accordance with multiple

studies showing that high recombination regions tend to be more permissive to introgression compared with regions with low recombination rates (Aeschbacher et al. 2017), we found that the introgressed regions were significantly concentrated in genomic regions with high recombination (fig. 5B and supplementary fig. S24B, Supplementary Material online). In addition, we assessed the deleterious mutation load by calculating the ratio of deleterious (including LoF) to synonymous variants, and in most trios, we found a reduced burden of deleterious mutations in introgressed regions in both P2 and P3 species than expected at random (fig. 5B and supplementary fig. S24B, Supplementary Material online). We further evaluated gene ontology (GO) enrichment to find out whether genes of introgressed origin were enriched for specific biological functions. Our results showed that in seven trios with more divergent P2 and P3 species (e.g., between *P. alba*, *P. adenopoda*, and the other four species), GO terms of "pollen–pistil interaction," "recognition of pollen," "pollination," "reproduction," and "cell recognition" were significantly enriched in the introgressed regions (supplementary table S11, Supplementary Material online). Whereas GO terms of "cellular response to stress," "double-strand break repair," "response to jasmonic acid," "response to brassinosteroid" were enriched in introgressed regions between more recently divergent species, although most of them were not significant after accounting for multiple comparisons (supplementary table S11, Supplementary Material online).

Finally, we sought to determine whether the candidate introgressed regions overlap with the inferred adaptive introgression sweeps. Based on the assumption that the action of adaptive introgression can leave a "volcano" pattern with narrowly reduced genetic diversity at the selected target sites but flanked by broad genomic region with high diversity (Moest et al. 2020; Setter et al. 2020), we used VolcanoFinder to calculate composite-likelihood ratio (CLR) statistics to detect signatures of introgression sweeps within each species. The adaptive introgressed regions were defined as those introgressed regions that overlap with the inferred introgression sweeps of the highest 5% CLR support in either species (supplementary fig. S25, Supplementary Material online). On an average, 9.26% of introgressed regions had at least one putative sweep signal across the trios. In P2–P3 species pairs with largely different genetic diversity and N_e, we found that selective sweeps were significantly enriched within introgressed regions in species with low genetic diversity and N_e (e.g., *P. adenopoda* and *P. alba*) (fig. 5B; supplementary fig. S24B, Supplementary Material online; table 1). Introgressed regions were not significantly enriched for sweeps in the species with high genetic diversity and N_e. In addition, in three trios where P2 and P3 species have similar genetic diversity and N_e, inconsistent patterns were found with introgressed regions being enriched for sweeps in some trio (Prot_Pdav_Ptra) but not in others (Prot_Pdav_Ptrs, Prot_Ptra_Ptrs) (table 1 and supplementary fig. S24B, Supplementary Material online).

Recent studies revealed that false-positive signals of adaptive introgression in genomic regions with high coding density

and low recombination rate are more likely to be generated owing to the stronger influence of recessive deleterious mutations in these regions (Zhang et al. 2020). To account for the confounding effects of coding density and recombination rate in the detection of adaptive introgression, we separated the introgressed regions into adaptive and nonadaptive introgressed regions. Compared with nonadaptive introgressed regions, we found significantly lower coding density and similar recombination rates in adaptive introgressed regions in most trios (fig. 5C and supplementary fig. S24C, Supplementary Material online), suggesting that the confounding factors such as recessive deleterious mutations should not have much influence on the detection of adaptive introgression in this study. Moreover, we calculated and compared the deleterious mutation load and the haplotype homozygosity-based statistic (iHH12) between adaptive and nonadaptive introgressed regions. In most trios, we did not find significant differences of deleterious mutation load between adaptive and nonadaptive regions. In contrast, compared with nonadaptive introgressed regions, adaptive introgressed regions showed significantly higher iHH12 values in both P2 and P3 species in most trios (fig. 5C and D; supplementary fig. S24C, Supplementary Material online). These results further support for the strong signatures of positive selection in these candidate regions and suggest that they are likely to be the genuine targets of adaptive introgression. Lastly, we investigated whether genes located in the candidate regions of adaptive introgression were enriched for specific biological functions. We found that the GO categories with the most significant enrichment are related to protein kinase activity, defense response, NAD+ nucleosidase activity, and cellular protein metabolic process (supplementary fig. S26, Supplementary Material online). Moreover, we identified a set of adaptive introgressed genes for which orthologs have been shown to be involved in responses to circadian rhythm, photoperiodic flowering, and diverse kinds of environmental stresses (supplementary table S12, Supplementary Material online).

Discussion

Genome-scale data have revealed that hybridization and introgression are important evolutionary processes. Although the prevalence of hybridization and introgression have been broadly accepted, we still lack an understanding of their genomic consequences (Moran et al. 2021). In this study, we have used whole-genome resequencing data to identify seven genetically divergent *Populus* species and to certify that several previously ambiguous species are all hybrids. Owing to the morphological distinct cohort of hybrids, especially for long-lived trees, hybrid individuals can often be recognized and misclassified as separate species, which have been reported in *Eucalyptus*, *Juglans*, and *Quercus* (Burgarella et al. 2009; Robins et al. 2021; Zhang et al. 2022). We found that except for *P. ×canescens* where a few individuals were found to be backcrossed hybrids, all other previous ambiguous species are all first-generation hybrids between different pairs of distinct species. The low frequency of post-F₁ hybrids

Table 1. Genomic Characteristics of Putative Introgressed Regions among the Ten Trios (corresponding to fig. 5B and supplementary fig. S24B, Supplementary Material online).

Trio	P1	P2	P3	f_4	Nucleotide Diversity[a]		Recombination Rate[a]		Deleterious Mutation Load[a]		Introgression Sweep Signals[a]		F_{ST}[a]	d_{xy}[a]
					P2	P3	P2	P3	P2	P3	P2	P3		
Ptrs_Prot_Palb	Populus tremuloides	P. rotundifolia	P. alba	7.53%	1.066*	1.091*	1.231**	1.240**	0.886*	0.773**	1.159	1.549**	1.009	0.943**
Ptrs_Pdav_Palb	P. tremuloides	P. davidiana	P. alba	7.06%	1.089**	1.117*	1.332**	1.292**	0.914	0.825*	0.885	1.560**	0.971*	0.946**
Ptrs_Ptra_Palb	P. tremuloides	P. tremula	P. alba	5.99%	1.130**	1.172*	1.283**	1.308**	0.925	0.930	1.038	2.051**	0.913**	0.913**
Ptrs_Pdav_Pade	P. tremuloides	P. davidiana	P. adenopoda	2.02%	1.236**	2.139**	1.908**	0.825	1.058	0.999	0.885	4.883**	0.643**	0.834**
Prot_Pdav_Ptra	P. rotundifolia	P. davidiana	P. tremula	9.73%	1.070**	1.043*	1.389**	1.314**	0.946	0.972	1.347*	1.762**	0.854**	0.952**
Prot_Pdav_Ptrs	P. rotundifolia	P. davidiana	P. tremuloides	4.29%	1.070**	1.027	1.087	1.094*	0.999	0.974	0.924	0.622	0.911**	0.936**
Ptrs_Prot_Pade	P. tremuloides	P. rotundifolia	P. adenopoda	2.64%	1.376**	2.053**	1.755**	0.868	1.085	1.054	1.301	4.640**	0.654**	0.801**
Ptrs_Palb_Pade	P. tremuloides	P. alba	P. adenopoda	3.84%	1.229*	1.392**	1.516**	1.124*	0.891	0.875	1.454	2.513**	0.943**	0.895**
Ptrs_Ptra_Pade	P. tremuloides	P. tremula	P. adenopoda	0.36%	1.242*	1.999**	1.260	0.731	0.996	1.137	0.239	2.947**	0.627**	0.740**
Prot_Ptra_Ptrs	P. rotundifolia	P. tremula	P. tremuloides	10.72%	1.002	0.989	1.148**	1.252**	0.930	0.959	0.633	0.515	0.992	0.915**

NOTE.—[a]The estimates of the observed values for each statistic relative to the averages of 1,000 genomic randomizations, with asterisks denoting the assessed significance (*$P < 0.05$; **$P < 0.01$).

indicates the presence of postzygotic isolation barriers (Coughlan and Matute 2020), such as the nonviability and sterility of F_1 hybrids, may prevent ongoing hybridization between these distinct species. However, future more explicit studies are needed to investigate the genetic architecture of intrinsic postzygotic barriers in these divergent species. In comparison, we found low numbers of fixed nucleotide differences between P. davidiana and P. rotundifolia, and moreover, advanced-generation hybrids are found in the hybrid zone between the two species, implying that strong reproductive and genetic divergence have not yet had enough time to evolve in these younger species (Abbott 2017).

After removing individuals of recent hybrid origin, we further identified distinct monophyletic clades and a well-supported phylogeny for the seven species. The population genomic data showed that the seven species displayed dramatically different demographic histories and varied extent of the accumulation of deleterious mutations. Compared with the other four species, P. qiongdaoensis, P. adenopoda, and P. alba display much lower levels of genetic diversity and experience long-term historical population declines. The efficacy of selection is substantially reduced in these species, resulting in an increased burden of deleterious mutations, especially for P. qiongdaoensis and P. adenopoda (fig. 3). However, some caution should be applied here, because the multiple-species comparisons are based on the results using a single P. tremula reference genome for read mapping and variant calling, which inevitably raises the concern of reference bias (Lachance and Tishkoff 2013). Nevertheless, several studies have shown that most population genetic analyses and demographic inferences yield consistent results regardless of the reference genome choice (Gopalakrishnan et al. 2017; Günther and Nettelblad 2019). In particular, all individuals used in this study are sequenced at high coverage (average coverage $>30\times$) and have high mapping rates to the reference genome (average mapping rate $>95\%$), suggesting that the effects of reference bias are supposed to be limited in this study.

On the other hand, our results demonstrate that hybridization and introgression have been pervasive throughout the evolutionary and speciation history of these *Populus* species, despite species-pairs being diverged for millions of years (Wang et al. 2016). The f_4-ratio statistics show that the history of introgression has shaped an appreciable proportion of extant genomes, ranging from 0.36% to 10.72% between various pairs of species. We next investigated how various evolutionary forces acts together to determine the extent to which introgressed regions are retained across the genome (Martin and Jiggins 2017; Moran et al. 2021). First, we found there was a decrease in the fraction of introgressed genome as the divergence age between hybridizing species increased. This is expected as the density of genetic incompatibilities increases as species divergence, which further affect the amount of residual introgression between species after hybridization (Coyne and Orr 1989; Matute et al. 2010). Second, variation in recombination rates along the genome is inferred to play a key role in determining the genomic landscape of introgression (Nachman and Payseur 2012; Aeschbacher et al. 2017).

Recent studies have suggested that if barriers to introgression are polygenic and made up of many loci across the genome, a positive correlation between introgression and recombination is expected because natural selection is more efficient to separate neutral or mutually beneficial foreign loci from linked deleterious loci in regions with high recombination rates (Schumer et al. 2018; Martin et al. 2019). In accordance with this expectation, introgressed regions in most trios exhibited significantly higher recombination rates and genetic diversity compared with the genomic background. Third, differences in the efficiency of purifying selection between the two parental species could also determine the patterns of introgression (Harris and Nielsen 2016; Kim et al. 2018). If the donor species harbors fewer deleterious mutations, the hybridization could alleviate the genetic load of the recipient species, resulting in a lower load in introgressed compared with nonintrogressed genomic regions of the recipient species (Juric et al. 2016). In fact, we found a trend of reduced burden of deleterious mutations in introgressed regions in most of our species trios (fig. 5B), indicating that the introgressed regions with higher genetic load have been purged by purifying selection and those retained thus carry fewer deleterious mutations. Finally, recent studies suggest that adaptive introgression is another important evolutionary process that result in the persistence and spread of introgressed alleles in a number of plant species (Hufford et al. 2013; Suarez-Gonzalez et al. 2018; Leroy et al. 2020; Nelson et al. 2021). In accordance with these findings, our results show strong signatures of adaptive introgression sweeps in some introgressed regions. Interestingly, we found that in pairs of species with substantial variation in N_e, introgressed regions were inferred to have undergone selective sweeps at greater than expected frequencies in the species with lower N_e. This finding reflects the fact that adding foreign genetic variation from species with large N_e to species of small N_e would more likely to create a fitness advantage (Pfennig et al. 2016). This is because species with large N_e usually have greater efficiency of natural selection to improve the use of weakly beneficial alleles, and if a large donor species sends migrants into a small recipient species, individuals with hybrid ancestry may have more adaptive alleles and thus higher relative fitness (Olson-Manning et al. 2012). In addition, introgression from species with large N_e can also increase the heterozygosity and supply novel genetic variation on which positive selection can act and can further accelerate adaptation (Allendorf et al. 2010). As the effective population size differed greatly across many hybridizing species, investigating how demography and recombination interacts with positive and negative selection will lead to a better understanding of how distinct evolutionary forces shape the genome-wide patterns of introgression across the tree of life.

In conclusion, we performed a comprehensive genomic investigation of nucleotide diversity, LD, demographic history, and genetic load for seven widespread and keystone *Populus* species. Coupled with the evidence of widespread past and contemporary hybridization among species, our results reveal that both purifying selection against deleterious mutations and positive selection favoring the adaptive introgressed loci have played fundamental roles in shaping the genomic landscape of introgression. These findings add to our understanding of how the various evolutionary mechanisms have acted together to shape the admixed genomes and to determine the extent and fate of introgressed genetic variants. The newly generated genome-wide data are also valuable resources for future breeding and biodiversity programs of these species to improve tree health and productivity in the face of climate change.

Materials and Methods

Sample Collection and Sequencing

We collected whole-genome resequencing data of 227 individuals from seven diagnostic and five ambiguous species from section *Populus* (including aspens and white poplars) of the genus *Populus*, including 27 *P. adenopoda*, 17 *P. alba*, 34 *P. davidiana*, 14 *P. qiongdaoensis*, 41 *P. rotundifolia*, 36 *P. tremula*, 21 *P. tremuloides*, four *P. wulianensis*, eight *P. tibetica*, 18 *P. canescens*, four *P. ningshanica*, and three *P. tomentosa*. All individuals were sampled from natural populations with at least 100 m apart. Among them, 119 individuals were collected from our previous studies (Wang et al. 2016; Li et al. 2021), and the other 108 individuals were sequenced in this study (supplementary table S1, Supplementary Material online). Genomic DNA was extracted from leaf samples with the Qiagen DNeasy plant kit. Whole-genome paired-ends reads with a target coverage of 20× were generated using Illumina platforms (HiSeq 2000 and Hiseq X Ten).

Read Mapping and SNP Calling

Prior to read mapping, we used Trimmomatic v.0.38 (Lohse et al. 2012) to remove adapters and to trim low-quality bases. Bases with quality lower than 20 were trimmed from the start or the end of reads, and the entire reads were discarded if shorter than 36 after trimming. Read quality was assessed with FastQC v.0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed October 23, 2019) before and after trimming. Remaining read pairs were mapped to a new version of high-quality *P. tremula* reference genome (v2.2, available at http://popgenie.org, last accessed October 12, 2019) with BWA-MEM v.0.7.17 (Li 2013) and sorted with SAMtools v.1.9 (Li et al. 2009). Then, MarkDuplicates tool from Picard v.2.18.21 (http://broadinstitute.github.io/picard/, last accessed March 16, 2019) was used to mark PCR duplicates, and GATK v.3.8.1 (DePristo et al. 2011) was used to perform realignment around indels. Finally, we used GATK v.3.8.1 HaplotypeCaller to do per-sample variant detection and used GenotypeGVCFs tool to perform joint genotyping. Because our data set contains individuals from multiple species, we used the "EMIT_ALL_SITES" flag to generate "all sites" VCF (including nonvariant sites) when doing SNP calling and genotyping.

To further retain high quality SNPs and minimize genotype calling bias, we used several stringent filtering criteria as following steps: 1) We used Heng Li's SNPable tool (http://lh3lh3.users.sourceforge.net/snpable.shtml, last accessed January 9, 2020) to mask genomic regions where reads were not uniquely mapped. To do so, we first divided the reference

genome into overlapping 100-mers and then aligned these back to the genome (bwa aln -R 1,000,000 -O 3 -E 3). Only the sites (328,895,043 out of 361,795,894 sites) in which all 100-mers mapped uniquely and without 1-mismatch were kept for downstream analyses; 2) SNPs with multialleles (>2) and with read depth (DP) lower than half or higher than 2-fold of the average sequencing depth were removed; 3) SNPs with quality by depth (QD) < 2.0, FS > 60, MQ < 20, MQRankSum < -12.5 and ReadPosRankSum < -8.0 were filtered; 4) SNPs located within 5-bp distance from any indels were filtered; 5) after treating genotypes with genotype quality score (GQ) <30 as missing, SNPs with $>20\%$ missing rate were filtered. Finally, a total of 8,966,513 SNPs were retained.

Population Structure and Phylogenetic Analyses

To analyze population structure, we only keep SNPs with missing rate $<10\%$ and minor allele frequency (MAF) $> 5\%$. We further excluded highly correlated SNPs by performing an LD-based SNP pruning in PLINK v.1.9.0 (Purcell et al. 2007). We scanned the genome with sliding windows of 50 SNPs and advancing steps of ten SNPs, and any SNP with a correlation coefficient (r^2)>0.2 with other SNP within the window was removed, which yielded 170,245 independent SNPs for population structure analyses. First, we used the smartpca program in EIGENSOFT v.7.2.1 (Patterson et al. 2006) to perform the PCA on the pruned SNPs (fig. 1B). We then used fastSTRUCTURE (Raj et al. 2014) to investigate the population structure across all individuals, with the number of clusters (K) being set from 1 to 12. The model complexity that maximizes the marginal likelihood was evaluated by the script chooseK.py, which further determine the optimal K value in the data set (fig. 1C). Finally, in order to construct the phylogenetic relationship of these individuals and species, we first obtain publicly available short read Illumina data for two outgroup species: P. trichocarpa (SRA ID: SRR1569480, SRR1569515) and P. euphratica (SRA ID: SRR5712977, SRR5713030) (Evans et al. 2014; Ma et al. 2018). Populus trichocarpa and P. euphratica are from section Tacamahaca and Turanga of the genus, respectively, which represent the proper outgroup species for the studied species as which are all from section Populus. We individually aligned the reads from the four individuals of the two outgroup species to the P. tremula assembly, and used UnifiedGenotyper tool of GATK v.3.8.1 (DePristo et al. 2011) to call SNPs at all sites with mode of "EMIT_ALL_SITES." We then intersected these sites with previously pruned SNP variants and allowed no missing in any of the four outgroup individuals, which leaved 46,034 pruned SNPs with outgroup information. To quantify the relatedness between individuals and species, the pairwise identify-by-state (IBS) genetic distance matrix was calculated using PLINK v.1.90 with the parameter-distance 1-ibs. Based on the distance matrix, we constructed a NJ phylogenetic tree using MEGAX (Kumar et al. 2018) and displayed the tree using FigTree v.1.4.4 (fig. 1A).

Classification of Parental and Hybrids Individuals

As some individuals were shown to be hybrids, we further calculated the proportion of ancestry and the interspecific heterozygosity for these individuals to further classify their hybrid status. First, we determined the ancestry-informative sites as those being fixed ($F_{ST} = 1$) SNPs between the assumed parental species. In total, we identified 634,428, 473,523, 326,131, 302,710, and 888 fixed SNPs between the five candidate species pairs: P. adenopoda and P. alba, P. adenopoda and P. davidiana, P. tremula (only use China population) and P. alba, P. rotundifolia and P. alba, P. davidiana and P. rotundifolia. By calculating the proportion of one-parent alleles over all alleles and the proportion of heterozygous sites for the putative hybrids, we assessed the hybrid index and interspecific heterozygosity using custom python script.

We determined the maternal and paternal species of each verified F_1 or backcrossed hybrid, and also quantified the proportion of the genome derived from each of the parental species. To do so, we inferred the phylogenetic relationships of all individuals by constructing the NJ chloroplast and mitochondrial phylogenetic tree, which are all maternally inherited in Populus and can help us to infer the maternal species. We first mapped the filtered reads from all 227 Populus individuals and four outgroup individuals (the same as above) against the P. tremula chloroplast and mitochondrial genome (Kersten et al. 2016) separately using BWA-MEM v.0.7.17. UnifiedGenotyper tool of GATK v.3.8.1 was then used to call SNPs using the haploid option (-ploidy 1) at all sites with mode of "EMIT_ALL_SITES." After treating genotypes with DP<100 and GQ<30 as missing data, a total of 749 and 2,589 biallelic SNPs with QD≥10 without missing genotypes were retained to construct the chloroplast and mitochondrial NJ tree (supplementary figs. S6 and S7, Supplementary Material online).

Estimation of Genetic Variation, LD, and Species Demographic History

After removing the identified contemporary hybrids, we assessed and compared the genomic characteristics and demographic histories among the seven divergent species. First, following the recommendation of (Korunes and Samuk 2021), we calculated interspecific genetic divergence (F_{ST} and d_{xy}) and intraspecific nucleotide diversity (π) over 100-kb nonoverlapping window after taking into account of both the polymorphic and monomorphic sites using the program pixy (Korunes and Samuk 2021) (figs. 1D and 2A). Second, for each species, the number of variants and the number of singleton SNPs in each individual as well as Tajima's D statistics over 100-kb nonoverlapping window were calculated using VCFtools v.0.1.15 (Danecek et al. 2011) (fig. 2B and C; supplementary fig. S14, Supplementary Material online). Third, we choose the same number of individuals (ten) from each species to infer the folded and unfolded SFS (supplementary fig. S13, Supplementary Material online). The derived versus ancestral allelic state was inferred through comparison with P. trichocarpa and P. euphratica sequences using the est-sfs software (Keightley and Jackson 2018). Fourth, LD decay of each species (fig. 2D and supplementary fig. S11, Supplementary Material online) was estimated for all pairs of SNPs with MAF ≥ 0.05 within a 200-kb window using PopLDdecay v.3.40 (Zhang et al. 2019). Finally, PSMC was used to infer historical dynamics of N_e with parameters -

N25 -t15 -r5 -p "4 + 25×2 + 4 + 6" for each species. Assuming a mutation rate of 3.75×10^{-8} mutations per site per generation and a generation time of 15 years (Ingvarsson 2008), we converted scaled population parameters into years and N_e (fig. 2E). For each species, we selected 11 individuals to run the PSMC analyses and 50 bootstrap estimates were made per individual (supplementary fig. S12, Supplementary Material online). Bootstrapping was conducted by breaking the consensus genome sequence into 0.5-Mb sequence segments and then randomly sampling segments with replacement to let the total size of sampled segments close to the reference genome.

Assessment of Genetic Load among Species

We calculated a number of metrics to quantify and compare the selection efficacy and the deleterious genetic load carried by each species. First, based on the comparison of folded SFS for 0-fold nonsynonymous and 4-fold synonymous sites, we used DFE-alpha v.2.16 (Keightley and Eyre-Walker 2007) to estimate the DFE for new 0-fold nonsynonymous mutations. After fitting a demographic model with a stepwise change in population size to the neutral SFS and incorporating the estimated parameters from the demographic model, fitness effects of new deleterious mutations and the strength of purifying selection ($N_e s$) were estimated for each species (fig. 3A). To generate the 95% confidence intervals for each parameter, we generated 200 bootstrap replicates by randomly resampling across all sites (both variant and invariant) in each site class and excluded the top and bottom 2.5% bootstrap replicates. Second, the genome-wide ratio of 0-fold nonsynonymous to 4-fold synonymous nucleotide heterozygosity were calculated for each individual over 100-kb nonoverlapping windows using pixy (Korunes and Samuk 2021) (fig. 3B). Third, the effects of SNP variants on protein-coding gene sequences were further annotated and classified as sites of loss of function (LoF), missense, and synonymous variants using SnpEff v.5.0 (Cingolani et al. 2012). LoF variant denote those with gain and/or loss of stop codon, or those with loss of start codon. Missense SNPs were further predicted as deleterious (score $\leq$ 0.05) or tolerated (score $>$ 0.05) based on the SIFT score computed by the program SIFT 4G (Vaser et al. 2016). At each SNP position, we determined the derived versus ancestral allelic state using the est-sfs software through comparison with *P. trichocarpa* and *P. euphratica* sequences. The relative proportion of homozygous and heterozygous derived alleles for LoF, deleterious, tolerated, and synonymous variants were estimated for each individual (fig. 3C and D). Finally, we estimated the inbreeding coefficient (F_{IS}) with VCFtools v.0.1.15 to determine the genomic extent of inbreeding for each individual among the seven species (supplementary fig. S15, Supplementary Material online).

Detection of Past Introgression across the Genome

To evaluate and test for the signals of past introgression among the seven *Populus* species, we implemented two statistical analyses for the 162 nonhybrid individuals. The first approach was to examine and quantify the variation in local topological relationships of the seven species throughout the genome. We used TWISST (Martin and Van Belleghem 2017) to calculate the weightings of all possible tree topologies through 100,000 iterative samplings of subtrees along the genomes. The filtered SNPs were phased and imputed with Beagle v.4.1 (Browning and Browning 2009). The nonoverlapping windows of both a fixed absolute window size (10 kb) and a fixed number of SNPs (1,000 SNPs) were, respectively, used for inferring local genealogies using PhyML with the GTR model (fig. 4A and B; supplementary table S8, Supplementary Material online). The second approach was to calculate the Patterson's D and f_4 admixture ratio statistics (Patterson et al. 2012) for all 35 possible combinations of trios of the seven *Populus* species using *P. trichocarpa* and *P. euphratica* as the outgroup (fig. 4C and supplementary table S9, Supplementary Material online). Allowing nonmissing genotypes in the two outgroup species, the analysis was performed based on the data set of 7,128,956 SNPs using the Dtrios program in Dsuite (Malinsky et al. 2021). The significance of the D-statistics was determined by performing a block jackknifing of the statistic. In Dsuite, the calculation of the f_4 admixture ratio requires that P3 be split into two subsets, P3a and P3b, which was done by randomly sampling alleles from P3 at each SNP.

To further locate the introgressed genomic regions for the trios of species with significant D values, we computed the f_d and its extension f_{dM} statistics for sliding windows of 50 SNPs with steps of 20 SNPs throughout the whole genome (fig. 5A and supplementary fig. S24A, Supplementary Material online). Following the approach used by (Morales-Cruz et al. 2021), we defined the putative introgressed regions as windows with highest X% of f_{dM} values, where X was determined by the proportion of introgression estimated based on the f_4 ratio statistic after merging top windows separated by <1 kb in distance. We further performed GO enrichment analysis using topGO 2.42.0 (Alexa and Rahnenführer 2009) to identify the biological processes of genes in introgression regions for each trio.

We next characterized and compared the genomic features between the introgressed regions and genomic-wide averages for each trio, we used "bedtools shuffle" (Quinlan and Hall 2010) to generate 1,000 random sets of genomic regions, each having the same size as the introgressed regions, to represent genome-wide scale. We compared π, F_{ST}, d_{xy}, population-scaled recombination rate, deleterious mutation load, and signatures of adaptive introgression sweeps between the introgressed regions and the 1,000 randomly chosen genomic regions of the same size (fig. 5B and supplementary fig. S24B, Supplementary Material online). For the estimation of population-scaled recombination rate, we used the Interval program of LDhat v2.2 (McVean et al. 2004) to perform the calculation, with 1,000,000 MCMC iterations sampling every 2,000 iterations and a block penalty parameter of five for each species. Deleterious mutation load was determined as the ratio of the derived deleterious (also including the loss-of-function) variants to the derived synonymous variations. The genome-wide scans of introgression sweeps within each species were implemented using VolcanFinder v.1.0 (Setter et al. 2020) with the Model1 over

10-kb nonoverlapping windows. Only CLR peaks exceeding a threshold defined as 95th percentile of the distribution of CLR values across all windows were considered as evidence for sweeps. The overlapped introgressed regions identified by both f_{dM} statistic and VolcanFinder approach were defined as those under adaptive introgression. For each statistic, we compared the distribution of 1,000 randomizations with the observed estimate to compute the P value.

To further identify and compare the signatures of selection between adaptive and nonadaptive introgressed regions (fig. 5C; and supplementary fig. S24C, Supplementary Material online), we calculated the haplotype homozygosity-based statistic iHH12 (Torres et al. 2019), which was certified to have good power to detect both hard and soft sweeps, for SNPs with MAF $> 5\%$ using the software *selscan* with default parameters (Szpiech and Hernandez 2014). After calculation, the iHH12 values were then normalized across the whole genome by the *norm* function within the *selscan* package for each species. Finally, to determine whether any functional classes of genes were over-represented among the candidate adaptive introgression regions, we performed a functional enrichment analysis using web-based platform g: Profiler (https://biit.cs.ut.ee/gprofiler, last accessed December 10, 2021). The interaction network of the GO categories with Benjamini–Hochberg false discovery rate <0.05 was visualized using the EnrichmentMap tool implemented in the software Cytoscape (Reimand et al. 2019) (supplementary fig. S26, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Author Contributions

J.W. and J.L. conceived and designed the research. J.W. supervised the study. L.Z. and K.M. performed the sampling and collected the materials. T.M. prepared DNA for sequencing. J.W., S.L., Y.S., Q.L., X.Z., C.J., Z.L., and J.L.W. conducted all bioinformatic analyses. J.W. wrote the manuscript, with input from S.L., J.L., N.R.S., and P.K.I. All authors approved the final manuscript.

Data Availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Material on-line. The newly generated genome sequencing data of the samples produced in this study have been deposited in the National Genomics Data Center (NGDC) under the accession number PRJCA006056. The VCF file and all other relevant summary statistics files are available at https://zenodo.org/record/5726696. All scripts used in this study will be available at https://github.com/jingwanglab/Populus-Genomic-Consequences-of-Hybridization.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. *J Evol Biol.* 26(2):229–246.

Abbott RJ. 2017. Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *J Syst Evol.* 55(4):238–258.

Aeschbacher S, Selby JP, Willis JH, Coop G. 2017. Population-genomic inference of the strength and timing of selection against gene flow. *Proc Natl Acad Sci U S A.* 114(27):7061–7066.

Alexa A, Rahnenführer J. 2009. Gene set enrichment analysis with topGO. *Bioconductor Improv.* 27:1–26.

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. *Nat Rev Genet.* 11(10):697–709.

Browning BL, Browning SR. 2009. A unified approach to genotype im-putation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet.* 84(2):210–223.

Burgarella C, Lorenzo Z, Jabbour-Zahab R, Lumaret R, Guichoux E, Petit R, Soto A, Gil L. 2009. Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity* 102(5):442–452.

Carlson SM, Cunningham CJ, Westley PA. 2014. Evolutionary rescue in a changing world. *Trends Ecol Evol.* 29(9):521–530.

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 ge-nome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.

Coughlan JM, Matute DR. 2020. The importance of intrinsic postzygotic barriers throughout the speciation process. *Philos Trans R Soc Lond B Biol Sci.* 375(1806):20190533.

Coyne JA, Orr HA. 1989. Patterns of speciation in *Drosophila*. *Evolution* 43(2):362–381.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al.; 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.

Dickmann DI, Kuzovkina J. 2014. Poplars and willows of the world, with emphasis on silviculturally important species. In: Isebrands JG, Richardson J, editors. Poplars and willows: trees for society and the environment. Wallingford: CABI. p. 8–91.

Eckenwalder JE. 1996. Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckler TM, editors. Biology of *Populus* and its implications for management and conservation Ottawa: NRC Research Press. p. 7–32.

Edelman NB, Mallet J. 2021. Prevalence and adaptive impact of intro-gression. *Annu Rev Genet.* 55:265–283.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet.* 46(10):1089–1096.

Glémin S. 2003. How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution* 57(12):2678–2687.

Gopalakrishnan S, Castruita JAS, Sinding M-HS, Kuderna LF, Räikkönen J, Petersen B, Sicheritz-Ponten T, Larson G, Orlando L, Marques-Bonet T, et al. 2017. The wolf reference genome sequence (*Canis lupus*

lupus) and its implications for *Canis* spp. population genomics. *BMC Genomics* 18(1):1–11.

Günther T, Nettelblad C. 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* 15(7):e1008302.

Harris K, Nielsen R. 2016. The genetic cost of Neanderthal introgression. *Genetics* 203(2):881–891.

Hough J, Williamson RJ, Wright SI. 2013. Patterns of selection in plant genomes. *Annu Rev Ecol Evol Syst.* 44(1):31–49.

Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. 2013. The genomic signature of crop-wild introgression in maize. *PLoS Genet.* 9(5):e1003477.

Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* 180(1):329–340.

Isabel N, Holliday JA, Aitken SN. 2020. Forest genomics: advancing climate adaptation, forest health, productivity, and conservation. *Evol Appl.* 13(1):3–10.

Jansson S, Douglas CJ. 2007. *Populus*: a model system for plant biology. *Annu Rev Plant Biol.* 58:435–458.

Juric I, Aeschbacher S, Coop G. 2016. The strength of selection against Neanderthal introgression. *PLoS Genet.* 12(11):e1006340.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.

Keightley PD, Jackson BC. 2018. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics* 209(3):897–906.

Kersten B, Faivre Rampant P, Mader M, Le Paslier M-C, Bounon R, Berard A, Vettori C, Schroeder H, Leplé J-C, Fladung M. 2016. Genome sequences of *Populus tremula* chloroplast and mitochondrion: implications for holistic poplar breeding. *PLoS One* 11(1):e0147209.

Kim BY, Huber CD, Lohmueller KE. 2018. Deleterious variation shapes the genomic landscape of introgression. *PLoS Genet.* 14(10):e1007741.

Korunes KL, Samuk K. 2021. pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour.* 21(4):1359–1368.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* 35(6):1547–1549.

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35(9):780–786.

Laenen B, Tedder A, Nowak MD, Toräng P, Wunder J, Wötzel S, Steige KA, Kourmpetis Y, Odong T, Drouzas AD, et al. 2018. Demography and mating system shape the genome-wide impact of purifying selection in *Arabis alpina*. *Proc Natl Acad Sci U S A*. 115(4):816–821.

Leroy T, Louvet JM, Lalanne C, Le Provost G, Labadie K, Aury JM, Delzon S, Plomion C, Kremer A. 2020. Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytol.* 226(4):1171–1182.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv. Available from: http://arxiv.org/abs/1303.3997.*

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–496.

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(16):2078–2079.

Li JL, Zhong LL, Wang J, Ma T, Mao KS, Zhang L. 2021. Genomic insights into speciation history and local adaptation of an alpine aspen in the Qinghai–Tibet Plateau and adjacent highlands. *J Syst Evol.* 59(6):1220–1231.

Lin Y-C, Wang J, Delhomme N, Schiffthaler B, Sundström G, Zuccolo A, Nystedt B, Hvidsten TR, De la Torre A, Cossu RM, et al. 2018. Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proc Natl Acad Sci U S A*. 115(46):E10970–E10978.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 40(Web Server Issue):W622–W627.

Lynch M, Conery J, Burger R. 1995. Mutation accumulation and the extinction of small populations. *Am Nat.* 146(4):489–518.

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. *Proc Natl Acad Sci U S A*. 115(2):E236–E243.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* 350(6267):1493–1498.

Malinsky M, Matschiner M, Svardal H. 2021. Dsuite-Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour.* 21(2):584–595.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol Biol Evol.* 32(1):244–257.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.* 17(2):e2006288.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev.* 47:69–74.

Martin SH, Van Belleghem SM. 2017. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* 206(1):429–438.

Matute DR, Butler IA, Turissini DA, Coyne JA. 2010. A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science* 329(5998):1518–1521.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–584.

Moest M, Van Belleghem SM, James JE, Salazar C, Martin SH, Barker SL, Moreira GR, Mérot C, Joron M, Nadeau NJ, et al. 2020. Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. *PLoS Biol.* 18(2):e3000597.

Morales-Cruz A, Aguirre-Liguori J, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. 2021. Extensive introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. *Genome Biol.* 22(1):254.

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. 2021. The genomic consequences of hybridization. *eLife* 10:e69016.

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos Trans R Soc Lond B Biol Sci.* 367(1587):409–421.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nat Rev Genet.* 12(2):111–122.

Nelson TC, Stathos AM, Vanderpool DD, Finseth FR, Yuan Y-W, Fishman L. 2021. Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation (*Mimulus* section *Erythranthe*). *PLoS Genet.* 17(2):e1009095.

Olson-Manning CF, Wagner MR, Mitchell-Olds T. 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. *Nat Rev Genet.* 13(12):867–877.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192(3):1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2(12):e190.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* 14(2):e1002379.

Pfennig KS, Kelly AL, Pierce AA. 2016. Hybridization as a facilitator of species range expansion. *Proc R Soc B.* 283(1839):20161329.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set

for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197(2):573–589.

Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, et al. 2019. Pathway enrichment analysis and visualization of omics data using g: profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc*. 14(2):482–517.

Rieseberg LH, Carney SE. 1998. Plant hybridization. *New Phytol*. 140(4):599–624.

Robins T, Binks R, Byrne M, Hopper S. 2021. Landscape and taxon age are associated with differing patterns of hybridization in two *Eucalyptus* (Myrtaceae) subgenera. *Ann Bot*. 127(1):49–62.

Runemark A, Vallejo-Marin M, Meier JI. 2019. Eukaryote hybrid genomes. *PLoS Genet*. 15(11):e1008404.

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, et al. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* 360(6389):656–660.

Setter D, Mousset S, Cheng X, Nielsen R, DeGiorgio M, Hermisson J. 2020. VolcanoFinder: genomic scans for adaptive introgression. *PLoS Genet*. 16(6):e1008867.

Suarez-Gonzalez A, Lexer C, Cronk QC. 2018. Adaptive introgression: a plant perspective. *Biol Lett*. 14(3):20170688.

Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 31(10):2824–2827.

Torres R, Szpiech ZA, Hernandez RD. 2019. Correction: human demographic history has amplified the effects of background selection across the genome. *PLoS Genet*. 15(1):e1007898.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc*. 11(1):1–9.

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. *Mol Biol Evol*. 33(7):1754–1767.

Wang M, Zhang L, Zhang Z, Li M, Wang D, Zhang X, Xi Z, Keefover-Ring K, Smart LB, DiFazio SP, et al. 2020. Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. *New Phytol*. 225(3):1370–1382.

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, et al. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* 348(6231):242–245.

Zhang C, Dong S-S, Xu J-Y, He W-M, Yang T-L. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35(10):1786–1788.

Zhang W-P, Cao L, Lin X-R, Ding Y-M, Liang Y, Zhang D-Y, Pang E-L, Renner SS, Bai W-N. 2022. Dead-end hybridization in walnut trees revealed by large-scale genomic sequence data. *Mol Biol Evol*. 39(1):msab308. doi: 10.1093/molbev/msab308.

Zhang X, Kim B, Lohmueller KE, Huerta-Sánchez E. 2020. The impact of recessive deleterious variation on signals of adaptive introgression in human populations. *Genetics* 215(3): 799–812.