PROF. YONGSHUAI SUN (Orcid ID : 0000-0002-6926-8406)

PROF. JIANQUAN LIU (Orcid ID : 0000-0002-4237-7418)

Population genomic analysis reveals that homoploid hybrid speciation can be a lengthy process

Dafu Ru[1] *, Yongshuai Sun[1,2]*, Donglei Wang[1]*, Yang Chen[1], Tianjing Wang[1], Quanjun Hu[1], Richard J. Abbott[3], Jianquan Liu[1]#

[1]Key Laboratory for Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, P. R. China

[2]CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, P. R. China

[3]School of Biology, Mitchell Building, University of St Andrews, St Andrews, Fife KY16 9TH, UK

*equal contributions to this work; #corresponding author (liujq@lzu.edu.cn)

Abstract

An increasing number of species are thought to have originated by homoploid hybrid speciation (HHS), but in only a handful of cases are details of the process known. A previous study indicated that *Picea purpurea*, a conifer in the Qinghai-Tibet Plateau (QTP), originated through HHS from *P. likiangensis* and *P. wilsonii*. To investigate this origin in more detail we analyzed transcriptome data for 114 individuals collected from 34 populations of the three *Picea* species from their core distributions in the QTP. Phylogenetic, principal component and admixture analyses of nuclear SNPs showed the species to be delimited genetically and that *P. purpurea* was admixed with approximately 60% of its ancestry derived from *P. wilsonii* and 40% from *P. likiangensis*. Coalescent simulations revealed the best-fitting model of origin involved formation of an intermediate hybrid lineage between *P. likiangensis* and *P. wilsonii* approximately 6 million years ago (mya), which backcrossed to *P. wilsonii* to form *P. purpurea* approximately one mya. The intermediate hybrid lineage no longer exists and is referred to as a 'ghost' lineage. Our study emphasises the power of population genomic analysis combined with coalescent analysis for reconstructing the stages involved in the

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.14909

origin of a homoploid hybrid species over an extended period. In contrast to other studies, we show that these stages can in some instances span a relatively long period of evolutionary time.

Key words: Coalescent analysis, homoploid hybrid speciation, hybridization, *Picea*, population genomics, Qinghai-Tibet Plateau

Introduction

Homoploid hybrid speciation (HHS), the origin without change in ploidy level of a hybrid species that is reproductively isolated from its parent species, is considered rare (Coyne & Orr, 2004; Gross & Rieseberg, 2005; Ma *et al*., 2006; Abbott *et al*., 2010; Mao & Wang, 2011; Schumer *et al.*, 2014). However, an increasing number of homoploid hybrid species have been reported for plants and animals in recent years (Abbott *et al*., 2013; Yakimowski & Rieseberg, 2014; Lamichhaney *et al*., 2018), though only a handful of these currently meet the strict criterion of Schumer *et al*., (2014) that hybridization was the cause of reproductive isolation of the hybrid from its parents. It was initially proposed that HHS can occur rapidly, within a few generations, based on artificial synthesis of reproductively isolated hybrids (Stebbins, 1957; Grant, 1966a,b; Rieseberg, 1997) and computer simulation studies (McCarthy *et al*., 1995; Buerkle *et al*., 2000), and this has been confirmed very recently in the wild for a hybrid finch (*Geospiza*), which originated in only three generations in the Galapagos Islands (Lamichhaney *et al*., 2018). However, in many cases it is thought that genome stabilization of homoploid hybrid species can take hundreds rather than a few generations to occur, even though ecological and/or intrinsic isolation originates early in the formation of such species (Buerkle & Rieseberg 2008).

Recognising and categorizing putative examples of HHS in the wild is valuable even when evidence that hybridization caused reproductive isolation is missing (Nieto Feliner *et al.*, 2017). In fact, the latter evidence is very difficult to obtain for species with long generation times, such as trees. For example, the type of QTL mapping studies that were crucial in showing that hybridization was a cause of reproductive isolation of homoploid hybrid species of sunflowers and *Heliconius* butterflies (Rieseberg *et al.*, 2003; Salazar *et al.*, 2010; Heliconius Genome Consortium, 2012), would take many years to complete in trees. Recording putative homoploid hybrid species and discovering how they originated and over

what period of time remains valuable because it targets such species for further analysis of the causes of reproductive isolation of a hybrid lineage from its parents. If the formation of a hybrid species is rapid, with complete reproductive isolation established within a few generations (e.g. Lamichhaney *et al*., 2018), the genomic composition of such a lineage will not be affected by subsequent gene flow from its parents. However, if reproductive isolation remains leaky over a lengthy period of time, the genomic composition of the hybrid lineage could possibly be affected by subsequent bouts of hybridization and the action of selection and drift on resulting progeny. Opportunities for further hybridization and backcrossing might occur long after the intial hybrid lineage was formed, for example during a period of secondary contact after a long period of geographical isolation. Rather surprisingly, such occurrences of later-stage hybridization in the evolution of a putative homoploid hybrid species have yet to be reported.

In this study, we aimed to examine the diploid hybrid origin of a spruce species, *Picea purpurea*, based on phylogenetic and coalescent analyses of population genomic data. This species is endemic to the Qinghai-Tibet Plateau (QTP) where it is a dominant tree species of some spruce forests. Initially, it was considered closely related to *P. likiangensis* based on the joint possession of flat or nearly flat leaves, though the two species differ for seed cone size and the development of stomatal lines on the abaxial surface of leaves (Cheng & Fu, 1978). More recently, a phylogenetic analysis of chloroplast (cp) DNA variation indicated that *P. purpurea* was more closely related to *P. wilsonii* (Ran *et al.*, 2006). All three species appear to have different ecological niches: *P. likiangensis* occurs at high elevations (2500-4100 m) in the southeastern QTP; *P. wilsonii* occurs mainly at lower elevations (1400-2800 m) in the northeastern QTP with a few populations scattered in neighbouring areas of north China; *P. purpurea* occurs at higher elevations (2600-3800 m) than *P. wilsonii* in northeastern QTP where its distribution partially overlaps that of *P. wilsonii*. A population genetic analysis utilising 11 nuclear loci (Li *et al.*, 2010; Sun *et al.*, 2014) later indicated that *P. purpurea* originated through diploid hybridization between *P. wilsonii* and *P. likiangensis* with approximately 70% of its nuclear genome coming from *P. likiangensis* and the remainder from *P. wilsonii*. Surprisingly, however, both its mtDNA (maternally inherited via seed) and cpDNA (paternally inherited through pollen) were shown to be derived from *P. wilsonii* (Sun *et al.*, 2014; Lookwood et al. 2013), though in areas of sympatry *P. purpurea* shares some mt- and cpDNA haplotypes with *P. likiangensis* (Du et al. 2011; Zou et al. 2013). These findings indicated that *P. purpurea* is a diploid hybrid species of *P. wilsonii* and *P.*

likiangensis, but that backcrossing of an initial hybrid lineage to *P. wilsonii* may have occurred resulting in *P. purpurea* containing both the cpDNA and mtDNA of *P. wilsonii*.

Theoretically, backcrossing of a hybrid to *P. wilsonii* should result in a higher nuclear contribution from *P. wilsonii* to the hybrid, which is the opposite to what was found based on the population genetic analysis of 11 nuclear loci (Li *et al.*, 2010; Sun *et al.*, 2014). An alternative scenario is that *P. purpurea* diverged from *P. wilsonii* through bifurcation divergence, and subsequently was subjected to a high level of nuclear gene flow (introgression) from *P. likiangensis*, while retaining the organelle genomes of *P. wilsonii*. Introgression occurs commonly between many conifer species (e.g., Petit & Excoffier, 2009; Du *et al.*, 2011; Bodare *et al.*, 2013; Li *et al.*, 2013; Zou *et al.*, 2013; Li *et al.*, 2015; Sun *et al.*, 2015; Ru *et al.*, 2016; Suarez-Gonzalez *et al.*, 2016, 2018), lending support to this hypothesis. Coalescent ancestry analysis and modelling of whole-genome sequence diversity can help distinguish between the alternative scenarios of hybrid speciation and introgressive admixture after bifurcating divergence (Nice *et al.*, 2013). We therefore sequenced transcriptomes of individuals of *P. wilsonii*, *P. purpurea*, and *P. likiangensis*, and conducted population genetic and coalescent analyses to determine which of the above two scenarios most likely accounts for the origin of *P. purpurea*. An analysis of transcriptomes rather than genomes was conducted as the *Picea* nuclear genome is very large and contains many repetitive sequences (Nystedt *et al.*, 2013) making analysis difficult. In contrast, transcriptome analysis is relatively easy to perform and has been proved to be a powerful tool for population genomic analysis in other studies (e.g. Chapman *et al.* 2013; Ru *et al.* 2016).

Material and methods

Material and RNA sequencing

Fresh, mature leaf needles were collected from first-year branches of 34, 40 and 40 individuals of *P. wilsonii*, *P. purpurea*, and *P. likiangensis*, respectively, sampled allopatrically from spatially separated forest stands across the core distributions of each species (Fig. 1 and Table 1). Because of incomplete reproductive isolation in areas of overlap, hybridization occurs between *P. purpurea* and *P. likiangensis* (Du *et al.*, 2011), and between *P. wilsonii* and other spruce species (Zou *et al.*, 2013). However, our sampling excluded all forest stands where two or more spruce species co-occurred and where there might be a legacy, therefore, of recent hybridization and introgression. Sampling locations

were recorded using an eTrex GIS (German, Germany). In addition to sampling the three species above, one individual of *P. breweriana* was sampled as outgroup.

Needles were frozen immediately in liquid nitrogen and kept at -80° prior to RNA extraction. Total RNA was extracted using TRIzol® Reagent (Life Technologies, Thermo Fisher Scientific, Waltham, MA, USA) followed by a DNase treatment using the TURBO DNA-free™ Kit (Life Technologies, Thermo Fisher Scientific), and quantified on an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA). Sequencing libraries for each individual were prepared using a NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) and examined following standard RNA-seq methodology (Erich *et al.*, 2008; Hansen *et al.*, 2010; Jiang *et al.*, 2011; Wang *et al*., 2009). We used an Illumina Hiseq 2500 platform to generate 150bp paired-end raw reads and deposited data sets for all individuals in BioSample (average number of raw bases > 6 Gb; Table S1).

Read trimming and assembly

We used Trimmomatic (Bolger *et al*., 2014) to trim and filter raw reads. This involved (1) trimming adapter sequences and bases from either the start or the end of reads when base quality was < 20, and (2) discarding reads with fewer than 36 bases after trimming.

The RNA-seq *de novo* assembly for each species was performed using Trinity ver. 2.6.6 (Grabherr *et al.*, 2011) with default parameters based on pooled libraries, respectively. A set of non-redundant, representative sequences was retained by CD-HIT ver. 4.6.1 (Huang *et al.*, 2010) for all Trinity assemblies with a threshold value of 0.95. Coding and peptide sequences in the open read frame were predicted by the TransDecoder ver. 2.0.1 (Haas *et al.*, 2013). We ran "TransDecoder.LongOrfs" to search for ORFs at least 100 amino acids long and identified them based on homology to known proteins via blast and Pfam (Finn *et al*., 2016) searches. Blast and Pfam search results were integrated using "Transdecoder.Predict".

Read mapping and variant calling

Because reference genomes for *P. wilsonii*, *P. purpurea*, and *P. likiangensis* are not available it was not possible to verify the assembled transcriptomes of each tree surveyed through *de novo* gene prediction. Instead, we compared our assembly results with the published reference transcriptome of *P. abies* (Nystedt *et al.*, 2013) from which bacterial contaminants were removed using BLAST ver. 2.2.30+ (Camacho et al., 2009; Ru *et al.*, 2016). Our assembled transcriptomes for *P. wilsonii*, *P. purpurea*, and *P. likiangensis* had lower N50 values than the reference *P. abies* transcriptome (Table S2) and consequently we mapped quality-filtered reads (Table S1) to a revised transcriptome of *P. abies* (Ru *et al*. 2016) after removing the reported fungal transcripts (Ru *et al*. 2016; Delhomme *et al*. 2015). To generate cpDNA sequences for each of the 114 trees sampled we used the published chloroplast (cp) genome sequence for *P. abies* (Nystedt *et al*., 2013) as reference.

For each individual, filtered clean reads were mapped to both the reference transcriptome and cp genome reference, respectively, using BWA-MEM (Li & Durbin, 2009) algorithm with default parameters. Duplicates were excluded from further analyses with PICARD-TOOLS ver. 1.92 (http://broadinstitute.github.io/picard/), followed by local re-alignment around indels using GATK (DePristo *et al*., 2011).

Variant calling was conducted using the "mpileup" of SAMTOOLS ver. 0.1.19 (Li *et al*., 2009) based on the uniquely mapped reads of all trees. The base quality and mapping quality were set to 20 and 30, respectively. The maximum posterior probability was used in genotyping each locus based on genotype likelihoods in the VCF file generated by SAMTOOLS. To obtain high quality SNPs, we filtered variant sites using the following criteria, respectively: no INDEL within a 5 bps window, root mean square quality >= 20, and proportion of missing bases within each species < 50%. Bases with depth of coverage (DP) <10 were considered missing for each tree (Li *et al.*, 2013; Wang *et al.*, 2013; Li *et al.*, 2014; Chapman *et al.*, 2013). Variant sites with significant deviation from Hardy–Weinberg equilibrium ($P < 0.001$) or with minimum allele frequency <0.05 were removed using VCFTOOLS ver. 0.1.14 (Danecek *et al*., 2011) to reduce false discovery rate. Sites with high observed heterozygosity (Ho >0.6) were removed from further analysis to decrease the effects of paralogous variants (Renaut *et al.*, 2013). The same pipeline in variant calling was used to generate the cpDNA sequence for each tree. For convenience, we denoted the nuclear transcriptomic sequences and cpDNA sequences as N-RNA-seq dataset and C-RNA-seq dataset, respectively.

Genetic diversity and structure

Transcriptome-wide distributions of nucleotide diversity π (Nei & Li, 1979), and Tajima's *D* (Tajima, 1989) were calculated using VCFTOOLS (Danecek *et al*., 2011) on the N-RNA-seq dataset. To investigate genetic differentiation between the three species, we calculated F_{ST} (Weir & Cockerham, 1984) per locus using VCFTOOLS ver. 0.1.14 (Danecek *et al.*, 2011). Negative values were reassigned to zero for the calculation of mean genome-wide F_{ST}. In addition, the average number of nucleotide substitutions d_{xy} per locus (Foote *et al.*, 2016) were calculated using a custom Perl script (Supplementary Material).

We constructed phylogenies for all 114 individuals based on N-RNA-seq datasets using TreeBeST (http://treesoft.sourceforge.net/treebest.shtml) with *P. breweriana* used as outgroup. Branch reliability was assessd using 1,000 bootstrap replicates. Maximum likelihood (ML)-based phylogenetic trees of C-RNA-seq datasets were further constructed using RAxML ver.8.1.20 (Stamatakis, 2014), with a single substitution model of GTR+GAMMA and 1000 bootstrap replicates.

Genetic structure across the three species was examined using principal component analysis (PCA) and ADMIXTURE analysis on the N-RNA-seq dataset. We used PLINK ver. 1.07 (Purcell *et al*., 2007; Danecek *et al*., 2011) with parameter --indep-pairwise 50 5 0.4 to reduce the linkage disequilibrium effect. SmartPCA program from EIGENSOFT package version 6.0.1 (Price *et al*., 2006) was used to perform PCA, followed by a Tracy–Widom test performed in R to determine the significance of principal components. ADMIXTURE ver. 1.23 (Alexander & Lange, 2011) was used to estimated the ancestry of each individual, with varying group number (*K*) from 1 to 10. In ADMIXTURE, 200 bootstrap replicates were used to calculate cross-errors. The optimal *K* was indicated by a lower cross-error.

PhyloNet test of HHS

We performed PhyloNet analysis to test HHS based on the orthologous sequences identified by OrthoMCL from the N-RNA-seq dataset (Loytynoja & Goldman, 2008). We generated the 1:1:1:1 orthologous gene datasets for *P. likiangensis*, *P. purpurea*, *P. wilsonii* and *P. breweriana* (outgroup) with a custom perl script. Amino acid sequences of each ortholog group were aligned using the L-INS-i algorithm in MAFFT (Katoh & Standley, 2013). The poorly aligned regions were trimmed using TrimAL v1.2 (Capella-Gutiérrez *et al.*, 2009).

The protein-coding nucleotide sequences for each ortholog group were aligned based on corresponding amino acid alignments using PAL2NAL v14 (Suyama *et al*., 2006) to ensure the correct reading frame. We constructed phylogenies per gene using RAxML's rapid bootstrap algorithm under the GTRGAMMA model to find the best-scoring ML tree with the groups satisfying the following cut-offs for filtering: sequence length > 300 bps, and '-' character (gap) in each sequence less than 50%. Each gene phylogeny was examined with 100 bootstrap replicates and gene trees with less than 70% bootstrap support were excluded from further analysis. A custom R script was used to count the number of resulting phylogenies showing different topologies.

The 3782 orthologous gene trees with more than 70% bootstrap support for branches were used to infer interspecific relationships using PhyloNet ver. 3.6.1 (Than *et al*., 2008; Yu *et al.*, 2014). Rooted trees were converted into the required input format with a custom Perl script. Maximum likelihood with parametric bootstrap networks (using the command InferNetwork_ML_Bootstrap) in a coalescent framework with both incomplete lineage sorting and gene flow taken into account, was inferred using PhyloNet allowing 0, 1 and 2 reticulations in 50 runs to return the best network.

Testing HHS using coalescent simulations

We used *fastsimcoal2* (Excoffier *et al*., 2013) to test HHS hypotheses by comparing predefined demographic models using coalescent simulations. For each pair of species, two-dimensional joint site frequency spectra (2D-SFS) were constructed from posterior probabilities of sample allele frequencies generated by ngsTools (Fumagalli *et al*., 2014) with the transcriptome of *P. abies* as reference (Nystedt *et al.*, 2013). Only the four-fold Degenerate Synonymous Sites (4DTV) were used to generate 2D-SFS. In total, we used 15 different evolutionary models (Fig. S4), of which 8 (model1 – model8) represented dichotomous or radiative topologies with or without gene flow after divergence, 5 models (model9–model13) represented classical models of homoploid hybrid speciation via a single hybridization event with or without migration/size-change after divergence, and 2 models (model14–model15) represented models of hybrid speciation involving a ghost intermediate hybrid lineage in the origin of *P. purpurea*.

For each model, we performed 100,000 coalescent simulations to estimate the expected 2D-SFS and computed log-likelihoods based on simulated and observed 2D-SFS matrixes. Global maximum likelihood estimates for each model were obtained from 60 independent runs, with 20-60 conditional maximization algorithm cycles. The relative fit of the different demographic models to the data was evaluated using the Akaike Information Criterion (AIC), and the model with the minimum AIC value was determined as the optimal. We assumed a mutation rate of 4.01×10^{-8} per site per generation and a generation time of 50 years (De La Torre *et al*., 2017; Li et al. 2010). A parametric bootstrapping approach was used to construct 95% confidence intervals with 50 independent runs for each bootstrap.

Results

Filtering, assembly, mapping and genotyping

We generated short Illumina-reads for 114 individuals of *P. likiangensis, P. wilsonii* and *P. purpurea,* and one individual of *P. breweriana* as outgroup. For each sample, after quality control, an average of 40.81M quality-filtered reads (about 89.9% of raw reads) was retained (Table S1). Using pool-assembly strategy, we generated *de novo* transcriptome assemblies for *P. likiangensis, P. wilsonii* and *P. purpurea*. After redundancy reduction and open reading frame (ORF) prediction for assemblies for *P. likiangensis*, *P. wilsonii* and *P. purpurea*, we produced 263279, 252680, and 127119 contigs with an N50 value of 588, 633, and 663 respectively (Table S2).

Because the total number of contigs in our assemblies was much larger and the N50 was lower than the transcriptome of *P. abies,* we mapped quality-filtered reads (Table S1) to the revised transcriptome of *P. abies* (Ru et al. 2016) and called SNPs for each individual. The average effective depth for our dataset was 47.22-fold, with average mapping rate of 60.90% coverage of the revised transcriptome reference (Table S1). After removing loci with missing rate >50% and deviation from HWE (P <0.001) in any of our three species (*P. likiangensis*, *P. purpurea* and *P. wilsonii*) 12,276 loci containing 670,146 single nucleotide polymorphisms (SNPs) remained. Of these, 107,301 SNPs showing MAF > 0.05 and *Ho* < 0.6 across all individuals (Table 2) were retained to form the N-RNA-seq dataset.

Mapping of quality-filtered reads from each individual (Table 1) to the chloroplast genome of *P. abies* resolved 1452 SNPs, of which 192 were retained after further filtering to form the C-RNA-seq dataset used for phylogenetic analysis.

Genetic diversity and population genetic structure

In the N-RNA-seq dataset, we identified 113,188 SNPs from 40 individuals of *P. likiangensis*, 96,329 SNPs from 34 individuals of *P. wilsonii*, and 108,913 SNPs from 40 individuals of *P. purpurea* (Table 2). Of these, 1,613 SNPs were specific to *P. purpurea*, 5,719 to *P. wilsonii* and 22,213 to *P. likiangensis*. About half of all SNPs (65,765 of 145,135) were shared among the three species (Fig. S1). Consistent with the expectation of hybrid speciation, the number of shared SNPs between *P. purpurea* and either *P. likiangensis* (86,715) or *P. wilsonii* (86,350) was higher than that between *P. likiangensis* and *P. wilsonii* (70,025) (Fig. S1). *Picea purpurea* ($\pi_{pururea}$ = 0.00392 ± 0.00305) contained a similar level of genetic diversity to *P. wilsonii* ($\pi_{wilsonii}$ = 0.00392 ± 0.00302) and *P. likiangensis* ($\pi_{likiangensis}$ = 0.00411 ± 0.00304). Tajima's *D* exhibited a strong skew toward low-frequency variants in all three species ($D_{likiangensis}$ = -0.890 ± 0.761; $D_{wilsonii}$ = -1.111 ± 0.728; $D_{pururea}$ = -1.216 ± 0.706).

Three species-specific clusters were identified by the PCA of nuclear SNPs (Fig. 5a), with the first two components (significant based on the Tracy-Widom test; Table S3) explaining 15.763% and 7.466% of the total variance, respectively. Such low proportions of variance explained by the first two PCs have been reported in other population genomic analyses (e.g. Malaspinas et al. 2016). ADMIXTURE further showed that three genetic groups (*K*=3) was optimal for the N-RNA-seq data set (highest log-likelihood value = -7705873.06, and lowest CV-error = 0.43). When *K* = 2, individuals of the two assumed parental species, *P. likiangensis* and *P. wilsonii*, clustered into two distinct groups, while individuals of the hybrid species, *P. purpurea*, exhibited mixed ancestry derived from *P. likiangensis* and *P. wilsonii* (Fig. 5b). ADMIXTURE showed that approximately 40% of the ancestry of *P. purpurea* was derived from the group represented by *P. likiangensis*, while approximately 60% was derived from *P. wilsonii*. When *K*=3, trees of each species clustered into their respective group, corresponding to the three species identified from morphological traits. Mean genome-wide differentiation between *P. likiangensis* and *P. wilsonii* (F_{ST} =0.207) was higher than that between *P. purpurea* and either *P. likiangensis* (F_{ST} =0.116) or *P. wilsonii* (F_{ST} =0.127) (Table 3, Fig. 3). Similarly, differentiation measured in terms of pairwise nucleotide divergence between species (d_{xy}) showed that d_{xy} between *P. likiangensis*

and *P. wilsonii* was significantly higher than between *P. purpurea* and either of its parents (Table 3, Fig. 4), consistent with the expectation of hybrid speciation model.

The three delimited genetic groups detected by PCA and ADMIXTURE were also evident in a NJ phylogeny constructed from the concatenated N-RNA-seq dataset (Fig. S3) with the *P. purpurea* lineage more closely related to the *P. wilsonii* lineage than to the *P. likiangensis* lineage. A ML phylogeny constructed from the C-RNA-seq dataset showed individuals of *P. likiangensis* constituted a clade with high bootstrap support, while individuals of *P. purpurea* and *P. wilsonii* clustered into another clade, which comprised several weakly supported subclades represented by *P. purpurea* and *P. wilsonii* respectively (Fig. S2), consistent with previous results based on fewer loci (Lookwood et al. 2013; Sun et al. 2014).

PhyloNet test of HHS

A total of 7762 orthologous gene groups across four species was identified and after filtering 7281 of these were used to generate gene trees. A total of 3782 gene trees with >=70% bootstrap support for all branches was subjected to PhyloNet testing. Of these, 1750 clustered *P. purpurea* with *P. wilsonii*, 1332 clustered *P. purpurea* with *P. likiangensis* and 700 showed *P. purpurea* as an isolated clade with *P. wilsonii* and *P. likiangensis* clustered together (Fig. 2a-c). The resulting inferred phylogenetic network by PhyloNet with an assumption of two past hybrid events (Fig. 2d) indicated a hybrid origin for *P. purpurea*. The reticulate relationship between *P. purpurea* and *P. wilsonii* had a higher inheritance probability (59.19%) than that between *P. purpurea* and *P. likiangensis* (40.81%), indicating a greater genomic contribution of *P. wilsonii* to *P. purpurea*.

Coalescent analysis of alternative speciation patterns

The best fitting model for the origin of *P. purpurea* (with the lowest AIC value, Table S4) was one involving hybridization and backcrossing (model 15) rather than bifurcation followed by introgression (Fig. S4, S5). The model indicates that a hybrid 'ghost' lineage was originally formed through hybridization between *P. likiangensis* and *P. wilsonii*, which through backcrossing to *P. wilsonii* gave rise to *P. purpurea* (Fig. 6). The estimated parameters of the model suggest that *P. purpurea* originated very recently ~1 Ma (95%HPDI: 0.56-1.42 Ma), whereas the hybrid 'ghost' lineage between *P. likiangensis* and *P. wilsonii* originated much earlier, ~6.42Ma (95%HPDI: 1.67-6.45 Ma). The estimated model

parameters also indicate that 61.10% of the nuclear genome of the ghost lineage was derived from *P. likiangensis*, and 38.90% from *P. wilsonii*. Importantly, it was further indicated that 62.84% of the nuclear genome of the initial population of *P. purpurea* was derived from the ghost lineage through backcrossing to *P. wilsonii*, while 37.16% came from *P. wilsonii*. In total, therefore, 38.4% of the nuclear genome of the intial population of *P. purpurea* was indicated to be derived from *P. likiangensis*, and 61.6% from *P. wilsonii*. Under this scenario, divergence between *P. likiangensis* and *P. wilsonii* was estimated to have occurred ~10.59Ma (95%HPDI: 4.47-14.07 Ma). The current effective population sizes (*Ne*) of *P. likiangensis*, *P. wilsonii* and *P. purpurea* were estimated to be ~70597, 72700 and 77423, respectively, and gene flow from *P. wilsonii* to *P. purpurea* was estimated to be higher than that in the opposite direction, or between *P. purpurea* and *P. likiangensis*, and between *P. likiangensis* and *P. wilsonii* (Table 4).

Discussion

Our analyses of population genomic data support previous findings based on a population genetic analysis of 11 loci (Sun *et al.*, 2014) that *P. purpurea* originated through diploid hybrid speciation from *P. likiangensis* and *P. wilsonii*. However, whereas our previous study indicated that approximately 70% of *P. purpurea*'s genome came from *P. likiangensis*, with the remainder inherited from *P. wilsonii*, the present extensive nuclear genomic analysis indicated that the majority (60%) of *P. purpurea's* genome came from *P.wilsonii* and a minority (40%) from *P. likiangensis*. The difference between the current estimates and Sun et al.'s estimates maybe caused by the number of loci (12095 vs 11 loci) used in each analysis. And we think, the current estimates are more accurate than before. This, together with the fact that *P. purpurea* contains both the mtDNA and cpDNA of *P. wilsonii* (Sun *et al*., 2014), indicates that following formation of a hybrid between the two parent species, backcrossing to *P. wilsonii* occurred, resulting in the origin of *P. purpurea*. Coalescent analysis supports this hypothesis and indicates that a hybrid lineage (now extinct and therefore referred to as a 'ghost' lineage) existed for a long period before *P. purpurea* originated through backcrossing to *P. wilsonii*. Our study emphasises the power of population genomic analysis combined with coalescent analysis for reconstructing the origin of a homoploid hybrid species and illuminating the different stages involved in such an origin and over what period they occurred.

Hybrid origin of *Picea purpurea* as an independent lineage

Our analyses showed that *P. purpurea* is genetically delimited from its two close relatives, *P. likiangensis* and *P. wilsonii*, containing 1,613 species-specific SNPs and forming a distinct genetic group in phylogenetic, PCA, and ADMIXTURE analyses of genome-scale nuclear variation. Phylogenetic analyses of orthologous genes showed that ~81% of 3782 gene trees examined clustered *P. purpurea* with either *P. wilsonii* or *P. likiangensis,* respectively. Furthermore, PhyloNet analyses of these trees indicated a hybrid origin of *P. purpurea* with approximately 60% of its nuclear genome derived from *P. wilsonii* and the remainder from *P. likiangensis*. ADMIXTURE analysis, conducted on genome-wide SNPs, further indicated a hybrid origin of *P. purpurea* with individuals having, on average, 60% of their ancestry in common with *P. wilsonii* and 40% shared with *P. likiangensis.*

To determine the likely evolutionary scenario for the origin of *P. purpurea,* particularly homoploid hybrid speciation versus bifurcation divergence followed by secondary gene flow, coalescent simulations were conducted on the nuclear SNP dataset of all three species using *fastsimcoal2.* A total of 15 alternative evolutionary models were tested and the best fitting one indicated that a hybrid lineage (now extinct) originally formed between *P. likiangensis* and *P. wilsonii*, and later backcrossed to *P. wilsonii* leading to the origin of *P. purpurea*. We cannot say, of course, whether the initial hybridization event and/or backcrossing to *P. wilsonii* was instrumental in causing *P. purpurea* to become reproductively isolated from its two parental species, i.e., to satisfy one of Schumer *et al.*'s (2014) three criteria for homoploid hybrid speciation. However, our results indicate that Schumer *et al.*'s two other criteria are satisfied, namely evidence of hybridization in the genome of *P. purpurea* and extrinsic ecological isolation from both parent species (Sun et al. 2014; Wang et al. 2017). With regard to the latter, *P. purpurea* is (eco)geographically isolated from *P. likiangensis* and ecologically isolated from *P. wilsonii* (Fig. 1; Sun et al. 2014; Wang et al. 2017). It should be noted that most spruce species can be intercrossed to produce viable seeds, indicating that incomplete intrinsic postzygotic reproductive isolation is widespread in the genus (Wright, 1955) and that extrinsic barriers are key to preventing hybridization and homogenization of species in the wild. It is feasible that transgressive segregation following hybridization between *P. likiangensis* and *P. wilsonii* and later backcrossing to *P. wilsonii* resulted in the formation of a hybrid lineage (*P. purpurea*) adapted to arid environments at high altitudes

from which both parent species are excluded (Sun et al., 2014). However, proving this will be very difficult, given the requirement to undertake a very long-term study involving crossing and detailed analysis of the fitness of both parent species and their hybrid progeny at locations where *P. purpurea* occurs.

Formation of a 'ghost' hybrid lineage and the lengthy origin of *P. purpurea*

According to our *fastsimcoal2* best-fitting coalescent model of origin (Fig. 6), two main stages were involved in the origin of *P. purpurea.* The first stage involved the origin of an intermediate hybrid lineage, which we might assume contained the mtDNA of one parent and the cpDNA of the other. This origin is dated to approximately 6 mya. *Picea wilsonii* and *P. likiangensis* are currently allopatrically distributed and no hybrids between them have been observed in the wild. Thus, it is feasible that as a result of tectonic and climatic oscillations in the eastern and northeastern QTP during the late Miocene and early Pliocene (Mulch & Chamberlain, 2006; Deng & Ding, 2015) the two species came into contact during this period and hybridized to form a hybrid lineage. This hybrid lineage must have existed for a lengthy period before the second stage occurred in the origin of *P. purpurea,* involving backcrossing of the hybrid lineage to *P. wilsonii,* dated to approximately one mya. This resulted in the origin of *P. purpurea,* which contained both the mtDNA and cpDNA of *P. wilsonii* as well as a greater proportion of the nuclear genome of *P. wilsonii*. Subsequently, the initial intermediate hybrid lineage became extinct during the Pleistocene, possibly because of climatic change occurring in the QTP (Deng & Ding, 2015). Detailed analysis of the genomes of *P. purpurea, P. likiangensis* and *P. wilsonii*, and coalescent analysis of the evolutionary history of these species, have, thus, shown that the origin of *P. purpurea* involved two phases of hybridization separated by approximately 5 million years, and was completed over a period ~6 million years, which is considerably longer than estimated for other homoploid hybrid species noted for their very rapid origins (Buerkle & Rieseberg, 2008; Abbott et al., 2010; Lamichhaney *et al*., 2018).

Although our analysis shows that *P. purpurea* contains a greater proportion of the genome of *P. wilsonii* relative to its other parent, *P. likiangensis*, *P. purpurea* is morphologically more similar to *P. likiangensis* (Cheng & Fu, 1978; Farjon, 1990; Fu *et al.*, 1999; Farjon, 2001). Its morphology is therefore presumably more influenced by genes inherited from *P. likiangensis*, which according to our simulations continued to exchange genes with *P. purpurea* following the latter's origin.

Acknowledgements This work was supported by grants from National key research and development program (2017YFC0505203), National Natural Science Foundation of China (grant numbers 31590821, 31670665, 91731301), National Key Project for Basic Research (2014CB954100), CAS "Light of West China" Program and Graduate Student's Research and Innovation Fund of Sichuan University (2018YJSY007).

Author contributions J.L. planned and designed the research. D.R., Y.S., D.W., Y.C., T.W. and Q.H. performed experiments, conducted fieldwork, analysed data etc. J.L., R.J.A., Y.S. and D.R. wrote the manuscript. D.R, Y.S. and D.W. contributed equally.

Data accessibility

The sequencing data have been deposited in GenBank under the bioproject: PRJNA401149 and PRJNA301093.

Supplementary Material

Supplementary figures S1-S5, Supplementary tables S1-S4 and Supplementary perl scripts can be found in the online version of this article.

References

Abbott RJ, Hegarty MJ, Hiscock SJ, Brennan AC (2010) Homoploid hybrid speciation in action. *Taxon*, **59**, 1375-1386.

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. *Journal of Evolutionary Biology*, **26**, 229-246.

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Bodare S, Stocks M, Yang JC, Lascoux M (2013) Origin and demographic history of the endemic Taiwan spruce (*Picea morrisonicola*). *Ecology and Evolution*, **3**, 3320-3333.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

Buerkle CA, Morris RJ, Asmussen MA & Rieseberg LH (2000) The likelihood of homoploid hybrid speciation. *Heredity*, **84**, 441–451.

Buerkle CA & Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266–275.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Capella-Gutiérrez S, Silla-Martínez J, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Chapman MA, Hiscock SJ, Filatov DA (2013) Genomic diver-gence during speciation driven by adaptation to altitude. *Molecular Biology and Evolution*, **30**, 2553–2567.

Cheng W, Fu L (1978) *Flora of China*. Science Press, Beijing, China.

Coyne JA, Orr HA (1998) The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London, Series B*, **353**, 287-305.

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.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK (2017) Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution*, **34**, 1363-1377.

Delhomme N, Sundstrom G, Zamani N € *et al*. (2015) Serendipitous meta-transcriptomics: the fungal community of Norway spruce (Picea abies). *PLoS ONE*, **10**, e0139080

Deng T, Ding L (2015) Paleoaltimetry reconstructions of the Tibetan Plateau: progress and contradictions. *National Science Review*, **2**, 417-437.

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. *Nature Genetics*, **43**, 491-498.

Du FK, Peng XL, Liu JQ, Lascoux M, Hu FS, Petit RJ (2011) Direction and extent of organelle DNA introgression between two spruce species in the Qinghai-Tibetan Plateau. *New Phytologist*, **192**, 1024-1033.

Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. *Nature Methods*, **5**, 679–682.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

Farjon A (1990) *Pinaceae. Drawings and descriptions of the genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and Picea.* Koeltz Scientific Books, Koenigstein, Germany.

Farjon A (2001) *World Checklist and Bibliography of Conifers.* 2nd edn. Royal Botanic Gardens, Kew, UK.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44**, D279–285

Fu L, Li N, Mill R (1999) *Pinaceae. Flora of China*. Science Press, Beijing, China.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) NGSTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486-1487.

Foote AD, Vijay N, Ávila-Arcos MC, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson MB, Korneliussen TS, Martin MD, *et al.* (2016) Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, **7**, 11693.

Gompert Z, Fordyce JA, Forister ML, Shapiro AM, Nice CC (2006) Homoploid hybrid speciation in an extreme habitat. *Science*, **314**, 1923-1925.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, **19**, 2455-2473.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644-652.

Grant V (1966a) Selection for vigor and fertility in a species cross in *Gilia. Genetics*, **53**, 757–775.

Grant V (1966b) The origin of a new species of *Gilia* in a hybridization experiment. *Genetics*, **54**, 1189–1199.

Gross BL, Rieseberg LH (2005) The ecological genetics of homoploid hybrid speciation. *Journal of Heredity*, **96**, 241-252.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494-1512.

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38, e131.

Heliconius Genome C (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94-98.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680-682.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, **21**, 1543-1551.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772-780.

Lamichhaney S, Han F, Webster MT, Andersson L, Grant BR, Grant PR (2018) Rapid hybrid speciation in Darwin's finches. *Science*, **359**, 224-228.

Li J, Fang X (1999) Uplift of the Tibetan Plateau and environmental changes. *Chinese Science Bulletin*, **44**, 2117-2124.

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, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078-2079.

Li J, Shi Y, Li B, Li SD (1995) *Uplift of the Qinghai–Xizang (Tibet) Plateau and global change.* Lanzhou University Press, Lanzhou, China.

Li L, Abbott RJ, Liu B, Sun Y, Li L, Zou J, Wang X, Miehe G, Liu J (2013) Pliocene intraspecific divergence and Plio-Pleistocene range expansions within *Picea likiangensis* (Lijiang spruce), a dominant forest tree of the Qinghai-Tibet Plateau. *Molecular Ecology*, **22**(20), 5237-5255.

Li L, Sun Y, Zou J, Yue W, Wang X, Liu J (2015) Origin and speciation of *Picea schrenkiana* and *Piceasmithiana* in the Center Asian Highlands and Himalayas. *Plant Molecular Biology Reporter*, **33**, 661-672.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, *et al.* (2013) Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics*, **45**, 1431–1438.

Li M, Tian S, Yeung CK, Meng X, Tang Q, Niu L, *et al.* (2014) Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Scientific reports*, **4**, 4678.

Li Y, Stocks M, Hemmila S, Kallman T, Zhu H, Zhou Y, Chen J, Liu J, Lascoux M (2010) Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution*, **27**, 1001-1014.

Lisiecki LE, Raymo ME (2007) Plio–Pleistocene climate evolution: trends and transitions in glacial cycle dynamics. *Quaternary Science Reviews*, **26**, 56-69.

Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632-1635.

Ma XF, Szmidt AE, Wang XR (2006) Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Molecular Biology and Evolution*, **23**, 807–816.

Mao JF, Wang XR (2011) Distinct niche divergence characterizes the homoploid hybrid speciation of *Pinus densata* on the Tibetan Plateau. *The American Naturalist*, **177**, 424–439.

Malaspinas A, Westaway MC,Craig M et al (2016). A genomic history of Aboriginal Australia. Nature, **538**, 207–214

McCarthy EM, Asmussen MA & Anderson WW (1995) A theoretical assessment of recombinational speciation. *Heredity*, 74, 502–509.

Mulch A, Chamberlain CP (2006) The rise and growth of Tibet. *Nature*, **439**, 670-671.

Nei M (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York, USA.

Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA (2013) Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution*, **67**, 1055-1068.

Nieto Feliner G, Álvarez I, Fuertes- Aguilar J, Heuertz M, Marques I, Moharrek F, Piñeiro R, Riina R, Rosselló JA, Soltis PS, Villa- Machío I (2017) Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, **118**, 513–516.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579-584.

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends Ecology Evolution*, **24**, 386-393.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904-909.

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. *American Journal of Human Genetics*, **81**, 559-575.

Ran JH, Wei XX, Wang XQ (2006) Molecular phylogeny and biogeography of *Picea* (Pinaceae): implications for phylogeographical studies using cytoplasmic haplotypes. *Molecular Phylogenetics and Evolution*, **41**, 405-419.

Renaut S, Grassa C, Yeaman S (2013) Genomic islands ofdivergence are not affected by geography of speciation insunflowers. *Nature Communications*, **4**, 1827.

Rieseberg LH (1997) Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, **28**, 359-389.

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211-1216.

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014) Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evolutionary Biology*, **14**, 254.

Ru D, Mao K, Zhang L, Wang X, Lu Z, Sun Y (2016) Genomic evidence for polyphyletic origins and interlineage gene flow within complex taxa: a case study of *Picea brachytyla* in the Qinghai-Tibet Plateau. *Molecular Ecology*, **25**, 2373-2386.

Salazar C, Baxter SW, Pardo-Diaz C, Wu G, Surridge A, Linares M, Bermingham E, Jiggins CD (2010) Genetic evidence for hybrid trait speciation in Heliconius butterflies. PLoS Genetics, 6, 1–12.

Schumer M, Rosenthal GG, Andolfatto P (2014) How common is homoploid hybrid speciation? *Evolution*, **68**, 1553-1560.

Shi Y, Li J, Li B, Li L (1998) *Uplift and environmental changes of Qinghai-Tibetan Plateau in the Late Cenozoic*. Guangdong Science and Technology Press, Guangzhou, China.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313**.**

Stebbins GL (1957) The hybrid origin of microspecies in the *Elymus glaucus* complex. *Cytologia*, Suppl. Vol., 336–340.

Suarez-Gonzalez A, Hefer CA, Christe C, *et al.* (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 QCB (2018) Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P.trichocarpa*. *New Phytologist*, **217**, 416-427.

Sun Y, Abbott RJ, Li L, Li L, Zou J, Liu J (2014) Evolutionary history of purple cone spruce (*Picea purpurea*) in the Qinghai-Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Molecular Ecology*, **23**, 343-359.

Sun Y, Li L, Li L, Zou J, Liu J, Carine M (2015) Distributional dynamics and interspecific gene flow in *Picea likiangensis* and *P. wilsonii* triggered by climate change on the Qinghai-Tibet Plateau. *Journal of Biogeography*, **42**, 475-484.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, W609–W612.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.

Wang GD, Zhai WW, Yang HC, Fan RX, Cao X, Zhong L, Wang L, Liu F, Wu H, Cheng LG, *et al.* (2013) The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nature Communications*, **4**, 1860.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57-63.

Wang GH, Li H, Zhao HW, Zhang WK (2017) Detecting climatically driven phylogenetic and morphological divergence among spruce (*Picea*) species worldwide. *Biogeosciences*, **14**, 2307-2319.

Weir BS, Cockerham CC (1984) Estimating *F- statistics* for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Wright JW (1955). Species crossability in spruce in relation to distribution and taxonomy. *Forest Science*, **1**, 319–340

Yakimowski SB, Rieseberg LH (2014) The role of homoploid hybridization in evolution: a century of studies synthesizing genetics and ecology. *American Journal of Botany*, **101**, 1247-1258.

Yu Y, Dong J, Liu KJ, and Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 16448–16453.

Zou J, Sun Y, Li L, Wang G, Yue W, Lu Z, Wang Q, Liu JQ (2013) Population genetic evidence for speciation pattern and gene flow between *Picea wilsonii*, *P. morrisonicola* and *P. neoveitchii*. *Annals of Botany*, **112**, 1829-1844.

Table 1. Locations of *Picea* individuals sampled for transcriptome analysis (including one individual of *P. breweriana*, 40 of *P. likiangensis*, 40 of *P. purpurea* and 34 of *P. wilsonii*)

species	latitude	longitude	Altitude(m)	n
P. breweriana	35.93306	104.15008	1774	1
P. likiangensis				40
MSZ-01	30.12058	101.75611	4221	1
MSZ-02	30.29497	101.60944	3978.51	3
MSZ-03	30.27436	101.522	3589.66	1
MSZ-04	29.99958	100.87139	4178.94	1
MSZ-05	30.28694	99.519222	4252	3
MSZ-06	29.72983	98.62975	4026.58	3
MSZ-07	29.61128	98.156944	4104.2	2
MSZ-08	29.68322	97.931917	4122.95	2
MSZ-09	29.55578	96.777333	4186.15	1
MSZ-35	29.49017	96.672278	3920.11	2
MSZ-15	29.18475	93.978556	2988.12	1
MSZ-25	29.46339	94.61775	2913.14	1
MSZ-30	29.56961	94.557972	3421.91	1
MSZ-31	29.67392	94.720028	3663.2	3
MSZ-33	29.8905	95.523278	2698.53	1
MSZ-34	29.82383	95.711528	3262.82	2
MSZ-38	28.40706	98.987944	3882.86	1
MSZ-40	27.93083	99.616472	3511.8	2
MSZ-42	27.569	100.02383	3025.85	1
MSZ-46	27.19836	100.27886	3260.41	1
MSZ-47	27.13161	100.23303	2947.51	3
MSZ-48	27.14214	100.2335	3197.45	2
MSZ-50	27.02508	100.20897	2845.13	2
P. purpurea				40
ZR_07	33.07804	102.85164	3568	5
ZR_08	34.02296	102.73741	3526	5
ZR_10	34.27802	103.00059	3556	7
ZR_11	34.45132	102.69788	3132	7

ZR_14	35.53111	102.24462	3085	8
ZR_25	33.0443	103.72414	3497	8
P. wilsonii				34
ZR_09	34.04897	103.2207	2390	7
ZR_15	36.95562	102.46394	2306	6
ZR_16	35.78197	104.05484	2304	7
ZR_24	33.29427	104.47862	2389	7
ZR_26	34.1588	102.90655	2769	7

n indicate the number of samples.

Table 2. Summary statistics for N-RNA-seq and C-RNA-seq datasets

	All samples[a]	All samples[b]	All samples[c]	All samples[d]	*P. likiangensis*[d]	*P. purpurea*[d]	*P. wilsonii*[d]
Number of loci	1	1	12276	11410	14107	12237	12156
Number of SNPs	1452	192	670146	107301	113188	108913	96329
$\pi \pm$ SD					0.00411 ± 0.00304	0.00392 ± 0.00305	0.00392 ± 0.00302
Tajima's D ± SD					-0.890 ± 0.761	-1.216 ± 0.706	-1.111 ± 0.728

[a]Based on raw C-RNA-seq dataset without MAF and *H*o filtering.

[b]Based on final C-RNA-seq dataset after MAF and *H*o control.

[c]Based on the raw N-RNA-seq without MAF and *H*o filtering.

[d]Based on the final N-RNA-seq after MAF and *H*o control.

Table 3. Summary statistics for F_{ST} and d_{xy}

Species pair	$F_{ST} \pm$ SD	$d_{xy} \pm$ SD
P. likiangensis vs. *P. wilsonii*	0.207 ± 0.158	0.319 ± 0.134
P. likiangensis vs. *P. purpurea*	0.116 ± 0.104	0.276 ± 0.108
P. wilsonii vs. *P. purpurea*	0.127 ± 0.145	0.269 ± 0.130

Table 4. Inferred demographic parameters of the best-fitting demographic model shown in Fig. 6

Parameters	Point estimation	95% CI Lower bound	95% CI Upper bound
$N_{e\text{-}ANC}$	65342	2,104	96,147
$N_{e\text{-}P.likiangensis}$	70,597	29,918	75,635
$N_{e\text{-}P.purpurea}$	77,423	32,917	94,648
$N_{e\text{-}P.wilsonii}$	72,700	19,881	77,336
$N_{e\text{-}ghost}$	48,999	16,238	137,627
$m^1_{P.likiangensis\text{->}P.wilsonii}$	0.00366	9.19e-9	0.0167
$m^1_{P.wilsonii\text{ ->}P.\ likiangensis}$	7.65e-6	5.53e-8	0.0160
$m_{ghost\text{->}P.likiangensis}$	0.0344	7.02e-9	0.0494
$m_{P.likiangensis\text{->}ghost}$	0.00108	1.06e-8	0.00861
$m^2_{P.likiangensis\text{->}P.wilsonii}$	1.76e-7	3.78e-8	2.57e-5
$m^2_{P.wilsonii\text{ ->}P.\ likiangensis}$	8.80e-6	1.43e-8	5.27e-5
$m_{P.likiangensis\text{->}P.purpurea}$	1.68e-5	7.29e-8	1.64e-4
$m_{P.purpurea\text{->}P.likiangensis}$	6.95e-5	4.07e-8	1.62e-4
$m_{P.purpurea\text{->}P.wilsonii}$	4.84e-5	1.93e-8	2.21e-4
$m_{P.wilsonii\text{->}P.purpurea}$	1.51e-4	2.04e-8	4.62e-4
T_{ADM1}	1,001,200	564,100	1,416,350
T_{ADM2}	6,422,150	1,673,800	6,447,150
T_{DIV}	10,586,400	4,465,050	14,068,850

$N_{e\text{-}P.likiangensis}$, $N_{e\text{-}P.purpurea}$, $N_{e\text{-}P.wilsonii}$, $N_{e\text{-}ghost}$ and $N_{e\text{-}ANC}$ indicate the effective population sizes of *P. likiangensis*, *P. purpurea*, *P. wilsonii*, the ghost intermediate lineage and ancestral population respectively. $m^1_{P.likiangensis\text{->}P.wilsonii}$, $m^1_{P.\ wilsonii\text{->}p.\ likiangensis}$, $m^2_{P.likiangensis\text{->}P.wilsonii}$ and $m^2_{P.\ wilsonii\text{->}p.\ likiangensis}$ indicate migration per generation before and after hybridization between *P. likiangensis* and *P. wilsonii* respectively; $m_{P.\ likiangensis\text{->}ghost}$: migration per generation from *P. likiangensis* to the ghost lineage and $m_{ghost\text{->}P.likiangensis}$ migration per generation from ghost lineage to *P. likiangensis*. $m_{P.likiangensis\text{->}P.purpurea}$, $m_{P.purpurea\text{->}P.likiangensis}$, $m_{P.wilsonii\text{->}P.purpurea}$, and $m_{P.purpurea\text{->}P.wilsonii}$ indicate migration per generation between *P. purpurea* and *P. likiangensis* or *P. wilsonii* in both directions, respectively. T_{ADM1} indicates time (years) of backcrossing of the ghost lineage to *P. wilsonii* that gave rise to *P. purpurea* while T_{ADM2} indicates time (years) of formation of the ghost lineage between *P. likiangensis* and *P. wilsonii*. T_{DIV} indicates the estimated divergence time (years) between *P. likiangensis* and *P. wilsonii* obtained from *fastsimcoal2*.