MOLECULAR ECOLOGY

Molecular Ecology (2017) **26**, 59–76 doi: 10.1111/mec.13765

SPECIAL ISSUE: THE MOLECULAR MECHANISMS OF ADAPTATION AND SPECIATION: INTEGRATING GENOMIC AND MOLECULAR APPROACHES

Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow

CAMILLE CHRISTE,*[1] KAI N. STÖLTING,*[1] MARGOT PARIS,* CHRISTELLE FRAÏSSE,†‡ NICOLAS BIERNE†‡ and CHRISTIAN LEXER*§

*Department of Biology, University of Fribourg, Chemin du Musée 10, CH-1700 Fribourg, Switzerland, †Institut des Sciences de l'Evolution (UMR 5554), CNRS-UM2-IRD, Place Eugene Bataillon, F-34095 Montpellier, France, ‡Station Méditerranéenne de l'Environnement Littoral, Université Montpellier 2, 2 Rue des Chantiers, F-34200 Séte, France, §Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria

Abstract

Speciation often involves repeated episodes of genetic contact between divergent populations before reproductive isolation (RI) is complete. Whole-genome sequencing (WGS) holds great promise for unravelling the genomic bases of speciation. We have studied two ecologically divergent, hybridizing species of the 'model tree' genus *Populus* (poplars, aspens, cottonwoods), *Populus alba* and *P. tremula*, using >8.6 million single nucleotide polymorphisms (SNPs) from WGS of population pools. We used the genomic data to (i) scan these species' genomes for regions of elevated and reduced divergence, (ii) assess key aspects of their joint demographic history based on genome-wide site frequency spectra (SFS) and (iii) infer the potential roles of adaptive and deleterious coding mutations in shaping the genomic landscape of divergence. We identified numerous small, unevenly distributed genome regions without fixed polymorphisms despite high overall genomic differentiation. The joint SFS was best explained by ancient and repeated gene flow and allowed pinpointing candidate interspecific migrant tracts. The direction of selection (DoS) differed between genes in putative migrant tracts and the remainder of the genome, thus indicating the potential roles of adaptive divergence and segregating deleterious mutations on the evolution and breakdown of RI. Genes affected by positive selection during divergence were enriched for several functionally interesting groups, including well-known candidate 'speciation genes' involved in plant innate immunity. Our results suggest that adaptive divergence affects RI in these hybridizing species mainly through intrinsic and demographic processes. Integrating genomic with molecular data holds great promise for revealing the effects of particular genetic pathways on speciation.

Keywords: divergence, pool-seq, *Populus*, secondary contact, selection, speciation

Received 26 January 2016; revision received 9 June 2016; accepted 14 July 2016

Introduction

Speciation often involves repeated, spatially variable episodes of genetic contact before reproductive isolation (RI) between diverging populations is complete (Coyne & Orr 2004; Arnold 2006). It is now widely recognized that the traditional sorting into different biogeographic modes of speciation (allo-, sym- and parapatric) is sometimes overly simplistic, as populations may alternate between these different biogeographic states during different stages of divergence (Smadja & Butlin 2011; The Marie Curie SPECIATION Network 2012).

Correspondence: Christian Lexer, Fax: +43 1 4277 9541; E-mail: christian.lexer@univie.ac.at

[1]These authors contributed equally.

Thus, 'divergence with gene flow' (DWGF) has been suggested as a synthetic term to address all those cases of divergence that entail gene exchange (i.e. sym- or parapatry) during some stage of the process (Smadja & Butlin 2011; Feder *et al.* 2012). These concepts have already proven useful for understanding patterns and drivers of DWGF along the divergence continuum in many groups of animals and plants, including numerous cases of adaptive population divergence and ecological speciation (Hoekstra *et al.* 2006; Elmer & Meyer 2011; Smadja & Butlin 2011; Jones *et al.* 2012; Seehausen *et al.* 2014; Baack *et al.* 2015). A better understanding of the historical context of adaptive divergence is nonetheless needed to appreciate the actual timing and role of gene flow during speciation (Abbott *et al.* 2013; Bierne *et al.* 2013; Feder *et al.* 2013).

The need to understand the divergence history of speciating lineages is especially pronounced in taxonomic groups with highly dynamic species ranges, for example groups situated in regions strongly affected by the climate-induced geographic range shifts of the pleistocene (Hewitt 2000; Tzedakis *et al.* 2013). Here, range shifts and secondary contact are expected, pervasive outcomes of organisms' tendency to track fitness optima in the face of changing environments (Davis & Shaw 2001). Range evolution and demographic processes will be especially dynamic in species with great dispersal capacities and high levels of standing genetic variation such as wind-pollinated forest trees (Petit & Hampe 2006; de Carvalho *et al.* 2010; Stölting *et al.* 2015) or marine species (Tine *et al.* 2014). Nevertheless, many/most organisms inhabiting Earth's temperate and Arctic zones and many tropical/subtropical regions are affected by these phenomena (Hewitt 2000). Not surprisingly therefore, there is currently an increasing interest in studying the genomic footprints of demographic history in the genomes of diverging populations (Gutenkunst *et al.* 2009; Chapman *et al.* 2013; Excoffier *et al.* 2013; Harris & Nielsen 2013; Roux *et al.* 2013; Sousa & Hey 2013; Tine *et al.* 2014).

Genomewide scans are widely recognized as a useful approach for studying genomic patterns of diversity and differentiation in speciating taxa, and for pinpointing the likely mechanisms behind these patterns (Bierne *et al.* 2011; Feder *et al.* 2012; The Heliconius Genome Consortium 2012; Ellegren 2014; Seehausen *et al.* 2014). Most studies thus far have focused on understanding the distributions and sizes of highly divergent genomic 'islands' and 'continents' of population divergence (Nosil *et al.* 2009; Feder *et al.* 2012; Roesti *et al.* 2014). However, the approach may also offer opportunities to search for low-differentiation regions embedded within the large, highly differentiated genome tracts of taxa that are at an advanced stage of divergence but are not yet fully isolated (Wu 2001; Feder *et al.* 2012; Roux *et al.* 2013; Stölting *et al.* 2013; Christe *et al.* 2016). Such low-differentiation 'pores' in the genome may stem from introgression, shared selection pressures or intrinsic features of genomes (Roux *et al.* 2013; Stölting *et al.* 2013; Renaut *et al.* 2014; Roesti *et al.* 2014), thus studying them may facilitate the identification of ecological and nonecological speciation trajectories. Genomewide scans involving functionally annotated genomes also hold great promise for identifying both adaptive amino acid substitutions and inferring the effect of purifying selection on segregating deleterious mutations (Smith & Eyre-Walker 2002; Nielsen 2005; Stoletzki & Eyre-Walker 2011). The former can be helpful for pinpointing functional groups of genes or even specific candidate genes involved in adaptation and speciation, while the latter facilitates testing the potential role of deleterious alleles in the weakening of reproductive barriers through heterotic effects (Bierne *et al.* 2002; Harris & Nielsen 2015), and the genetic load imposed by introgressed haplotypes on the recipient genome (Juric *et al.* 2015).

Complementary to genome scans, much can be learned from model-based inference of the joint demographic history of diverging populations from genomic data (e.g. reviews by Beaumont 2010; Pinho & Hey 2010). One increasingly popular approach of demographic inference involves the computation of joint SFS from genetic polymorphism data for two or more divergent populations, and estimation of demographic parameters such as population sizes, migration rates and time intervals since particular events using composite likelihood (Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). This approach of studying demographic footprints in genomic data can complement genome scans for diversity and differentiation along the chromosomes of diverging populations (above). In fact, adaptation and speciation genomic studies have started to interpret both types of genomic footprints in parallel (Roux *et al.* 2013; Tine *et al.* 2014; Der Sarkissian *et al.* 2015).

The forest tree genus *Populus* (poplars, aspens, cottonwoods) offers many excellent examples of ongoing DWGF involving ecologically important 'keystone' or foundation species, in which the genetic variation accumulating in diverging populations potentially affects entire communities and ecosystems (Whitham *et al.* 2006; Bernhardsson *et al.* 2013; Geraldes *et al.* 2014; Caseys *et al.* 2015a). As *Populus* spp. occur primarily in Earth's climatically unstable temperate zones and exhibit extraordinary dispersal capacities (wind-dispersed pollen and seed), understanding the joint demographic history of diverging populations and species in this genus is particularly challenging. This calls for the

combined analysis of genomic footprints along chromosomes and across SFS. Also, many recently diverged *Populus* species are ecologically divergent (Lexer *et al.* 2005; Dickmann & Kuzovkina 2008; Geraldes *et al.* 2014), which raises conceptual questions regarding the precise roles of ecological divergence in speciation (Barton & De Cara 2009; Bierne *et al.* 2011; Smadja & Butlin 2011).

Populus alba (White poplar) and *Populus tremula* (European aspen) are two genetically and ecologically divergent (flood plain species *vs.* upland pioneer) Eurasian taxa with 'porous' genomes, that is taxa with incomplete reproductive barriers (Lexer *et al.* 2005, 2010; Stölting *et al.* 2013). Recent high-throughput genotyping-by-sequencing studies of hybrid zones of these two species indicated steep genomic clines and fairly strong RI, maintained primarily by selection against recombinants ('intragenomic coadaptation') and an unknown contribution of prezygotic barriers (Lindtke *et al.* 2014; Christe *et al.* 2016). The two species are known for their temporarily separated demographic histories following the last glacial maximum (LGM), with the thermophilic *P. alba* recolonizing from southern and southeastern Europe, and the cold-adapted *P. tremula* from middle European refugia located much closer to the ice shields (Fussi *et al.* 2010; Christe *et al.* 2016). A genomewide analysis of local ancestry and differentiation in hybrid zones based on restriction site-associated DNA (RAD) sequencing revealed a striking and unexpected preponderance of F_1 hybrids relative to recombinants in present-day contact zones, but also pointed to subtle traces of more ancient admixture pulses (Christe *et al.* 2016). RAD-seq provides only limited numbers of markers, and the analytical methods used in that earlier study were not well suited for revealing complex demographic events in the more distant past. Thus, we were motivated to explore the genomic footprint of these species' joint demographic history at a much greater depth, powered by whole-genome sequencing (WGS).

In particular, the objectives of this study were to (i) provide a fine-grained picture of divergence, differentiation and allele sharing between these ecologically divergent, hybridizing forest tree species by whole-genome resequencing, (ii) infer these species' demographic history based on their joint genomewide SFS and (iii) scan their genomes for the relative impact of segregating deleterious alleles and adaptive substitutions and ask which functional groups of genes were most strongly affected by the latter. We discuss our results in the context of the likely mechanisms involved in divergence and the maintenance of RI and episodic introgression between these hybridizing, ecologically divergent species. We suggest potential ways forward for addressing the functional significance of the genes and molecules involved.

Materials and methods

Whole-genome sequencing and polymorphism detection

Sampling of species and populations. *Populus alba* and *Populus tremula* are two ecologically divergent tree species (floodplain vs. upland pioneer) that form large 'mosaic' hybrid zones within a large zone of geographic overlap in Europe and Asia (Lexer *et al.* 2005; Dickmann & Kuzovkina 2008). Populations of the two species were sampled adjacent to two well-known hybrid zone localities situated in the Ticino river valley in northern Italy (45.28°N, 8.98°E) and the Tisza river valley in eastern Hungary (48.32°N, 22.26°E) (Lexer *et al.* 2010). Restriction site-associated DNA (RAD) sequencing data for these two hybrid zone localities are consistent with the presence of different phylogeographic recolonization lineages in *P. alba* and weak population structure in *P. tremula* (Christe *et al.* 2016), in agreement with expectations from earlier phylogeographic data for these species (de Carvalho *et al.* 2010; Fussi *et al.* 2010; Du *et al.* 2015). The present WGS study was based on population samples of $N = 24$ individuals for each species in each locality. None of these trees showed signs of recent genome admixture in previous ancestry analyses based on mapped microsatellites (Lindtke *et al.* 2012). Individuals were sampled at a minimum distance of 50 m from each other to avoid sampling clonal ramets. The population samples of *P. alba* were previously characterized with a small amount of SOLiD4 (Applied Biosystems, Thermo Fisher Scientific) resequencing data (Stölting *et al.* 2015). This study greatly expands this data set using a combination of SOLiD5500 and Illumina HiSeq sequencing technologies for populations of both species in both localities.

Pooled whole-genome sequencing. We extracted total genomic DNA from silica-dried leaf tissue using Qiagen's DNeasy Plant Mini Kit. All DNA extracts were quantified on a Nanodrop 1000 system (Thermo Fisher Scientific). We combined DNA extracts for 24 individuals from each species and sampling locality into DNA pools of equal molarity. All DNA pools were individually barcoded and sequenced on both the SOLiD5500 and Illumina HiSeq systems at the Functional Genomics Center Zurich, using 75- and 100-bp paired-end chemistry, respectively. Data curation and polymorphism detection are reported in detail in Appendix S1 (Supporting information). Briefly, high-quality reads were reference-mapped against the *Populus trichocarpa* v3/build 210 genome sequence, using allelic information

available for *P. alba* and *P. tremula* to generate majority-rule consensus reference sequences. We formatted alignments with PICARDTOOLS 1.125 and realigned around indels using GATK version 3.3.0. We tabulated the depth of coverage for any given variable site using DepthOfCoverage in GATK, filtering reads at a minimum base quality of 20 and a minimum mapping quality of 20. We only retained sites with a minimum coverage of eight reads in each pool, and a minimum coverage of three reads for the minor allele to avoid spurious single nucleotide polymorphism (SNP) calls. We used the fractions of raw sequencing coverage for each allele and site as estimates of population allele frequencies. Keeping in mind the known limitations of allele frequency estimation from pool-seq data (Boitard *et al.* 2012; Schlötterer *et al.* 2014), allele frequency estimates were cross-validated against Sanger sequence data in a previous study (Stölting *et al.* 2015) and against RAD-seq data in the present study (below).

Genomewide scans for genetic differentiation, diversity and allele sharing

Allele frequency differentials (AFDs; Shriver *et al.* 1997; Turner *et al.* 2010) were calculated for each SNP from the read counts used to estimate allele frequencies. Then, AFDs were averaged for nonoverlapping windows of 8 kb along the genome, following the rationale of Stölting *et al.* (2015) and using known LD distances from a large poplar resequencing study as a guidance (Slavov *et al.* 2012). Window-based analyses allowed us to examine AFDs together with other statistics that were only available at the window level. These included the ratio of sites with fixed differences between species over the total number of polymorphic sites in each window ('fixratio'), which represents a simple measure of genetic differentiation and allele sharing (Stölting *et al.* 2013, 2015), and absolute divergence (Dxy), which is not influenced by within-population diversity (Cruickshank & Hahn 2014). Dxy was calculated from the counts of reference and alternative allele calls for species 1 and 2 in each locality as pairwise nucleotide diversity among populations following Ferretti *et al.* (2013). To identify genome regions with exceptionally high levels of allele sharing (=low differentiation), we made use of our 'fixratio' differentiation measure. More specifically, we estimated the size and distribution of low-fixation regions in the genome using R 3.1.1, considering adjoining windows with no fixation between species to be of the same region if they were separated by no more than a single intermittent window with non-zero fixation. We tested for departures of the distribution of these low-fixation regions from an even distribution using Kolmogorov–Smirnov (K-S) tests following Ellegren *et al.* (2012). These tests compare the distance matrices between mid-points of no-fixation windows to expectations from distance matrices of evenly distributed windows along each chromosome.

In addition to differentiation and divergence, we also estimated the following genetic diversity parameters at the window level: pooled heterozygosities (Rubin *et al.* 2010) and the selective sweep test statistic lnRH (Schlötterer & Dieringer 2005) as the ratio of pooled heterozygosities between species in each sampling locality, as in Stölting *et al.* (2015). To capture structural features of the poplar genome, we also estimated the fraction of repetitive DNA (repfrac) in each window. A comparison of repeat-masked and non-repeat-masked *P. trichocarpa* build 210 genomes allowed us to directly quantify the fraction of repetitive DNA in any given 8 kb window. As yet another structural genomic feature of potential interest, we predicted approximate centromere positions in the *P. trichocarpa* v2 build 156 genome assembly based on Slavov *et al.* (2012) and then lifted these coordinates to build 210 of the genome, used for reference-mapping here (above). To achieve this, we identified genes flanking each centromere in build 156 and obtained synonymous gene models in build 210 as indicated on www.popgenie.org. We double-checked these approximate centromere positions based on their expected high content of repetitive DNA. All analyses in this study were restricted to windows with at least 500 of 8000 bases covered. We also removed all windows with nucleotide diversity (pi) >0.1 and/or a fraction of variable sites >0.2 to guard against overinterpretation of potentially biased windowed parameter estimates. We note that the removal of high-diversity windows may have reduced our power to detect interspecific 'migrant tracts' (discussed below).

Demographic inference based on joint site frequency spectra

Filtering for demographic inference from SFS. Pool-seq data are not yet widely used for demographic inference (but see Boitard *et al.* 2012), presumably because of the known difficulty of estimating allele frequencies in the face of heterogeneity in coverage (Schlötterer *et al.* 2014). Here, we approached this issue by careful filtering of SNPs prior to demographic inference, and by validating SFS results from pool-seq data against those obtained from individually tagged RAD-seq data for the same populations, available from a previous study (Christe *et al.* 2016).

We filtered the pool-seq data using the 1st and 3rd quartiles of the depth distribution within each species as minimum and maximum coverage depth thresholds,

respectively. This was done to account for variation in coverage depth between populations and remove loci in the tails of the distribution, likely representing sequencing errors in the case of low coverage or duplicated loci in the case of high coverage. The difference in depth of coverage between *P. alba* and *P. tremula* in Hungary (Results) was of a magnitude that prompted us to refrain from further demographic inference for this locality; thus, all subsequent analyses of demographic inference were focused on the Italian locality only.

We thinned the pool-seq data for the Italian locality to randomly keep one SNP per window of 8000 bp. This was done to meet the requirement of independence among loci, which is a prerequisite for model comparisons with likelihood ratio tests (below). The joint SFS for *P. alba* and *P. tremula* in the Italian locality was estimated from the data with the program ∂a∂i (Gutenkunst *et al.* 2009) using *P. trichocarpa* as an outgroup to distinguish ancestral and derived alleles. To keep the maximum number of SNPs that matched our stringent coverage thresholds (above), the SFS was downsampled to 20 reads (assumed to correspond to 10 diploid individuals) for each population.

Demographic models. As the reality of our data (above) dictated a focus on a single pairwise population comparison, we chose the well-known composite-likelihood, diffusion approximation-based approach to demographic inference implemented in ∂a∂i (Gutenkunst *et al.* 2009). The flexibility of this approach allows incorporating heterogeneity in demographic parameters between groups of loci (Tine *et al.* 2014). A set of simple and plausible scenarios was chosen as basic models. They included strict isolation (SI), isolation with migration (IM), ancient migration (AM) with one and two periods of ancient gene flow (PAM) and secondary contact with one (SC) and two (PSC) periods of contact.

Complexity was added to these models to take into account specific aspects of divergence between these species and intrinsic features of genomes known to affect genomewide patterns of differentiation and allele sharing (Renaut *et al.* 2013; Burri *et al.* 2015) (Table S1, Supporting Information). Additional classes of models including recent expansion (prefix 'ex') were also tested to allow for known demographic events in the history of European forest trees following the LGM (Hewitt 2000; Fussi *et al.* 2010; Tzedakis *et al.* 2013) (Table S1, Supporting Information). We note that the IM, SC and PSC models included migration during expansion, whereas the AM and PAM models did not. For models including gene flow, heterogeneous gene flow across the genome (2M2P, Table S1, Supporting Information) was also incorporated to take into account the genomewide heterogeneity in introgression rates between speciating lineages (Roux *et al.* 2013; Sousa & Hey 2013). We considered that a certain proportion of sites were neutrally evolving and were exchanged at a fixed rate between species, whereas the remaining sites had a zero effective migration rate. Further, all models were also implemented by specifying heterogeneity in effective population size (2N), and in both gene flow and population size (2N2M2P). Our rationale was to explore the impact of variation in genomic features affecting diversity and recombination rates, that is the effects of background selection at linked sites in low recombination regions (Cruickshank & Hahn 2014). The five possible models of the family 2N2M2Pex (adding recent expansion to heterogeneous gene flow and N_e) were also constructed, but were not pursued further because of persistent convergence issues.

In total, 39 different models (Table S1, Supporting Information) were evaluated and fitted with the observed joint SFS using 30 replicate runs per model. Models were ranked according to their log likelihoods. For nested models, likelihood ratio tests were used to establish differences in model support. For non-nested models, the relative likelihood of the Akaike information criterion (AIC) was used instead. Model comparisons and the estimation of parameters for the best supported model are described in detail in Appendix S1 (Supporting information).

Validation of pool-seq results with RAD-seq data. To validate the results on demographic inference obtained from pool-seq WGS data, we used a data set for the same populations of *P. alba* and *P. tremula* from the Italian locality, individually sequenced for RAD-seq (Christe *et al.* 2016). The individually tagged RAD-seq data were subjected to a similar procedure as for pool-seq (above), with small modifications specific to this type of data (Supporting information). The same models explored for pool-seq WGS data were also examined for the RAD-seq data.

Exploring the genomic distribution of candidate 'migrant loci'. To examine the genomic distributions of putative 'migrant loci' identified by ∂a∂i, the best supported model (PAM2Nex) and the strict isolation model with expansion (SIex) were compared using ∂a∂i 's function Plotting.plot_2d_comp_ multinom. Because the two models represented very different demographic scenarios, they differed by a large number of cells that were mainly found in the centre of the joint SFS. Loci with derived allele frequencies corresponding to the greatest difference between the two models were retrieved from a nonthinned data set containing 8 307 756 SNPs with homogenous coverage required for ∂a∂i analyses. We then tested the effects of fixratio

(localization in no-fixation windows) and centromeric region on the proportion of 'migrant loci' in each 8 kb genome window using a generalized linear model with a quasi-binomial error distribution using R software v3.1.

Detection of long-term selection and characterization of affected gene sets

Exploring traces of positive selection at the sequence level (i.e. adaptive substitutions) is of interest in the context of understanding the processes that facilitate and/or accompany species divergence (Smith & Eyre-Walker 2002; Nielsen 2005). Likewise, knowing the proportions of segregating deleterious mutations present in hybridizing species is of interest because of the expected effects of heterospecific haplotypes on the fitness of recipient genomes (Harris & Nielsen 2015; Juric *et al.* 2015) and the weakening of reproductive barriers through heterotic effects (Bierne *et al.* 2002; Harris & Nielsen 2015). Based on our pool-seq data, we calculated alpha, the proportion of substitutions fixed by positive selection (Smith & Eyre-Walker 2002), at the level of individual genes in the poplar genome. We used NPSTAT v0.99 (Ferretti *et al.* 2013) to count and classify substitutions in reference-mapped reads as either synonymous or nonsynonymous fixed or polymorphic changes by comparing reads for *P. alba* and *P. tremula* separately against the *P. trichocarpa* reference sequence for each transcript and sampled population. We subsequently estimated the proportion of substitutions driven by positive selection based on equation 3 in Smith & Eyre-Walker (2002) for each transcript using a custom script in R version 3.1.1. We used *G*-tests as originally intended by McDonald & Kreitman (1991) to establish whether there were significant differences in the distributions of nonsynonymous and synonymous substitutions between fixed and polymorphic sites for each transcript. In subsequent analyses, we considered genes that had significant alpha in at least one of the two localities sampled for each species. We also used the proportions of fixed and polymorphic synonymous and nonsynonymous sites from NPSTAT v0.99 to estimate the direction of selection DoS, which facilitates the detection of adaptive protein evolution (positive DoS) and deleterious segregating variants (negative DoS; Stoletzki & Eyre-Walker 2011). We were particularly interested in comparing DoS between species, and between genome regions with indications of introgression (no-fixation windows; above) *vs.* the remainder of the genome. To check whether sequencing coverage in pool-seq WGS affected the results, we checked for associations between coverage and DoS across our comparisons of interest.

Genes with evidence for positive selection as identified by alpha tests were gene ontology (GO)-annotated using BLAST2GO-PRO v. 2.7.1 under default settings (Conesa *et al.* 2005) based on blastx searches of all *P. trichocarpa* transcripts against a local copy of the nr database as of June 2014. Gene ontology term enrichment analyses were performed using Fisher's exact tests as implemented in the Bioconductor package TOPGO v2.16 using a minimum node size of five as recommended by Alexa & Rahnenfuhrer (2010).

Results

WGS and polymorphism detection

Pooled whole-genome resequencing on the Illumina and SOLiD sequencing platforms yielded a total of 684 132 484 quality-trimmed sequencing reads for a total of 60.26 billion base pairs. Of these, 88.1% mapped against the *Populus trichocarpa* reference genomes. Quality sites were covered at 36.3× and 53.74× in *Populus alba* from Italy and Hungary, while *Populus tremula* pools were covered with 27.9× and 13.0× in Italy and Hungary, respectively. Given this relatively low sequencing coverage of *P. tremula* in Hungary, we focused subsequent demographic inference (below) primarily on the Italian interspecific comparison. Nevertheless, we note the extremely high correlation in genomewide interspecific differentiation levels between sampling localities ($r = 0.977$, $t = 884.92$, d.f. = 36916, $P < 0.001$, Fig. S1, Supporting Information). After carefully safeguarding against the presence of genome windows that were either poorly sampled or unusually diverse (likely reflecting the effects of past genome duplications; Tuskan *et al.* 2006), the final data set thus comprised 36 918 windows of 8KB each, equivalent to 74.9% of the *P. trichocarpa* assembly (Table S2, Supporting Information). These windows covered 138 109 168 bases (on average 3741 ± 1633 SD; median = 3828), including 8 607 742 variable SNP sites.

Genomewide patterns of genetic differentiation, diversity and allele sharing

Genomewide analysis (Table 1) indicated relatively homogeneous levels of interspecific Dxy (Italy, 0.027 ± 0.013 SD; Hungary, 0.027 ± 0.013 SD). Average levels of interspecific differentiation (AFD) were high (Italy, 0.393 ± 0.093 SD; Hungary, 0.399 ± 0.094 SD). High-differentiation windows more than two standard deviations (SD) different from genomewide means were frequent (Italy, 931 windows; Hungary, 988 windows) and widely scattered along chromosomes (Table 1; Figs 1 and S1, Supporting Information). We considered high-differentiation outlier windows that also exhibited reduced diversity (lnRH) as candidate regions for

Table 1 Genomewide counts and averages for numbers (no.) of covered sites, allele frequency differentials (AFDs), fractions of fixed SNPs, absolute divergence Dxy, the selective sweep test statistic lnRH and levels of selection (alpha)

	Italy	Hungary
No. of covered sites	138 109 168	138 109 168
Fraction fixed SNP's	0.057	0.115
Average AFD	0.393	0.399
No. no-fixation windows	2323	1460
No. AFD outlier windows*	931	988
Average AFD in no-fixation windows	0.253	0.233
Average AFD in AFD outlier windows	0.622	0.631
Average Dxy	0.027	0.027
No. lnRH outlier windows *Populus alba*†	1039	987
No. lnRH outlier windows *Populus tremula*†	885	903
No. hard sweep candidates *P. alba*‡	102	100
No. hard sweep candidates *P. tremula*‡	19	44
Average alpha *P. alba*§	0.435	0.434
Average alpha *P. tremula*§	0.481	0.442

*Outlier window counts are indicated for all windows with observed values >2SD different from genomewide averages.
†lnRH windows are considered outliers for extreme values of >2SD difference from genomewide means in either a positive or negative direction.
‡Hard sweep candidates are windows in *P. alba* or *P. tremula* with outlier status for both elevated AFD and reduced diversity as measured by lnRH.
§Alpha values are genomewide averages for all windows with alpha significantly different from zero.

classical, 'hard' selective sweeps. We observed 102 and 100 of these sweep candidate regions in *P. alba* in Italy and Hungary, respectively, and 19 and 44 in *P. tremula* in the two sampling sites, respectively (Table 1). We note that lnRH indicates species-specific sweeps only.

Despite high levels of genomewide differentiation, we frequently observed windows that did not exhibit any fixed allelic differences between species: 2323 such windows (6.3% of well-covered windows) representing 1538 low-fixation regions were detected in Italy, while 1460 windows (3.95%) representing 936 low-fixation regions were observed in Hungary (Table 1). Low-fixation regions were generally small both in Italy (mean 13 210 bp ± 12 684 SD) and Hungary (mean 13 600 bp ± 12 434 SD), respectively (Table S3, Supporting Information). K-S tests indicated significant departures from an even distribution of no-fixation regions in all chromosomes in Italy (mean D-statistic = 0.570 ± 0.206 SD, $P < 0.001$) and Hungary (mean D-statistic = 0.313 ± 0.184 SD, $P < 0.001$) (Fig. S2, Supporting Information).

Footprint of demographic history in joint site frequency spectra

After filtering, 29 808 SNPs were retained for demographic inference in the Italian locality. Observed joint SFS obtained from pool-seq (Fig. 2) and RAD-seq data (Fig. S3, Supporting Information) yielded congruent results, with most of the derived alleles residing in the 'frame' of the joint SFS (private and fixed variants) and few residing in the centre (shared polymorphisms), thus indicating a long history of divergence (Gutenkunst *et al.* 2009). In fact, the joint SFS (Fig. 2) indicates the presence of two different types of genome pairs 'superimposed' on one another, a highly divergent and a less divergent one (R. Gutenkunst, personal communication), as one might expect for highly divergent taxa with 'porous' genomes (Wu & Ting 2004).

Fit of different demographic models to pool-seq data. Globally, support for the seven basic divergence models followed a similar order within each major class of models, each of which was set up to add a new layer of complexity (Methods) (Tables 2 and S1, Supporting Information). Models allowing for gene flow generally performed significantly better than strict isolation models (SI) (Fig. 4; Tables 2 and S1, Supporting Information). Ancient migration models (AM/PAM) received significantly stronger support than models implementing continuous gene flow through time (IM) and secondary contact models (SC and PSC) (Table 2 and S1, Supporting Information). Patterns recovered by individually sequenced RAD-seq data were broadly congruent with those from pool-seq WGS (Fig. S4; Table S4, Supporting Information).

Comparisons between different model classes clearly favoured expansion over constant demography. Likewise, heterogeneity was favoured over homogeneity of gene flow (Fig. 4; Tables 2 and S1, Supporting Information). The strongest statistical support, however, was obtained for a model implementing repeated ancient gene flow with heterogeneity in N_e and expansion, PAM2Nex (Fig. 4; Table 2 and S1, Supporting Information).

Parameter estimates. For the best supported model PAM2Nex (Table 2; Figs 2 and 3; Table S1, Supporting Information), parameter estimates are shown in Table S5 (Supporting Information). Within this model, total divergence time between *P. alba* and *P. tremula* approximated >2.8 million (Mio) years (2.807 ± 0.129 SD; Table S5, Supporting Information). The period without contact was approximately three times longer than the time with ancient gene flow (Table S5, Supporting Information), consistent with high levels of interspecific

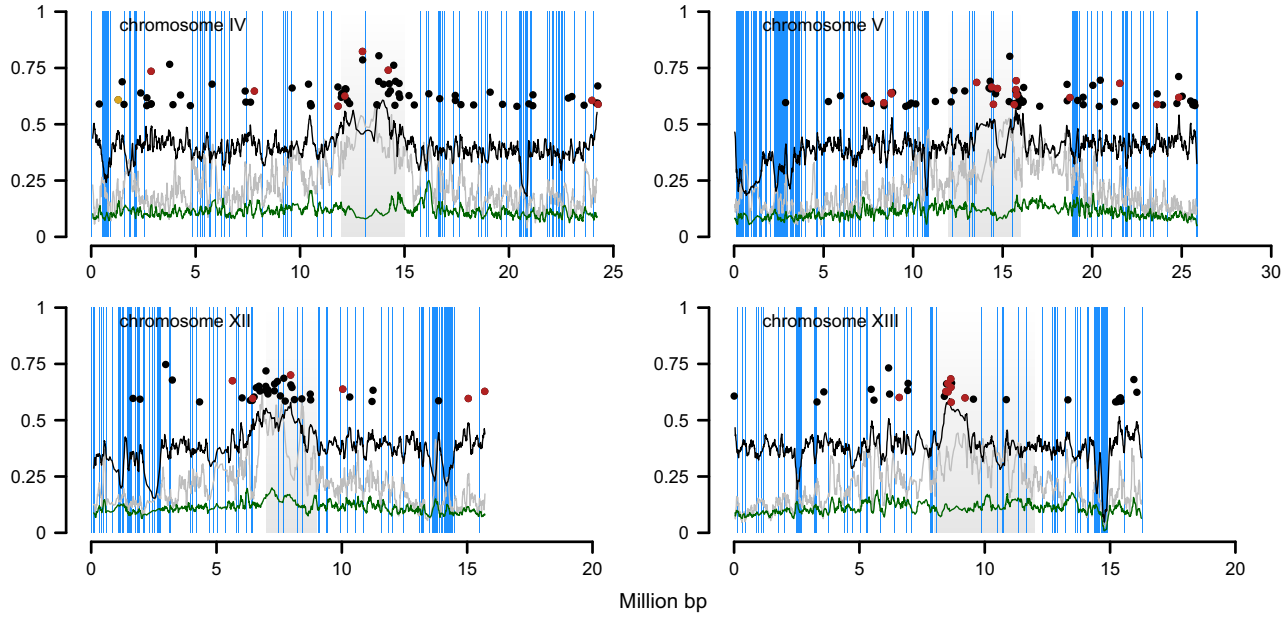

Fig. 1 Chromosome-wide patterns of differentiation (allele frequency differentials: AFDs, black), sequence divergence (Dxy, green) and genomic features (fraction of repetitive DNA, grey) for four *Populus* chromosomes in Italy. Approximate centromere positions are indicated by rectangular gradient boxes (grey), as are genomic windows free of fixation (8 kb length, with blue shades). AFD outlier windows ≥2SD different from genomewide averages are indicated with black dots, and we highlight hard sweep candidate windows with simultaneous outlier status for elevated differentiation and reduced diversity in *Populus alba* (red dots) and *Populus tremula* (golden dots). Dxy was scaled fourfold for better visibility.

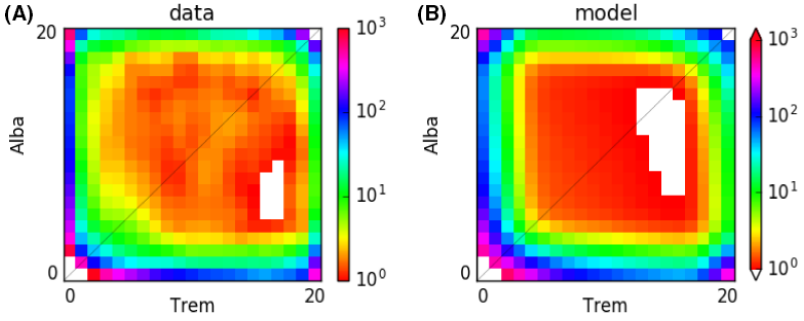

Fig. 2 Demographic history of *Populus alba* and *Populus tremula* for the Italian hybrid zone inferred from pool-seq data. (A) Joint site frequency spectrum (SFS) for populations of *P. alba* and *P. tremula* showing the count of derived alleles for 29 808 oriented single nucleotide polymorphisms in 20 random reads (assumed to represent 10 individuals) for each population. The empty (white) spots in the lower left corner are due to data quality filtering (see Materials and methods). (B) Maximum-likelihood SFS obtained under the PAM2Nex model.

divergence. The increase in population size due to expansion was pronounced for *P. alba* and was unnoticeable for *P. tremula*. Amounts of ancient gene flow were moderate and slightly asymmetric. A sizable portion of the genome (nr = 51%±0.06) was detected as being nonrecombining and potentially affected by background selection, and reductions in Ne were pronounced (Table S5, Supporting Information).

Validation of pool-seq results with RAD-seq. Very similar model fitting results were obtained with RAD-seq data (Fig. S4, Supporting Information), which effectively validated our demographic inference based on pool-seq. Most obviously, ancient and repeated ancient migration models (AM/PAM) performed better than all other models. In general, the rankings of models supported by likelihood ratio tests and AIC differences were very similar between pool-seq and RAD-seq (Tables 2, S1 and S4, Supporting Information). RAD-seq data also supported expansion and heterogeneity of gene flow and N_e across the genome as important components in the demographic history of these species (Tables 2, S1

Table 2 Results of model fitting for nine representative demographic models of divergence. Models are ranked according to their log likelihood (log L). Number of parameters (k) and Akaike information criterion (AIC) are given for each model. For model details, see footnote

Model	k	logL	AIC
SI	4	−2421.2	4850.5
PSC	7	−2383.5	4781.0
IM	6	−2383.5	4779.0
PAM	7	−2360.5	4735.1
PSC2M2P	9	−2241.3	4500.7
PAM2M2P	9	−2229.3	4476.7
SI2Nex	11	−2180.3	4405.3
PSC2M2Pex	12	−2136.9	4378.7
PAM2M2Pex	12	−2112.5	4249.0
PAM2Nex	12	−2086.3	4196.6

SI, strict isolation; IM, isolation with migration; PSC, secondary contact with two periods of contact; PAM ancient migration with two periods of contact. 2M2P, models implementing heterogeneity in gene flow; 2N, heterogeneity in effective population size (N_e); prefix ex, with or without recent expansion; best model: PAM2Nex, repeated ancient gene flow with heterogeneity in N_e and expansion.

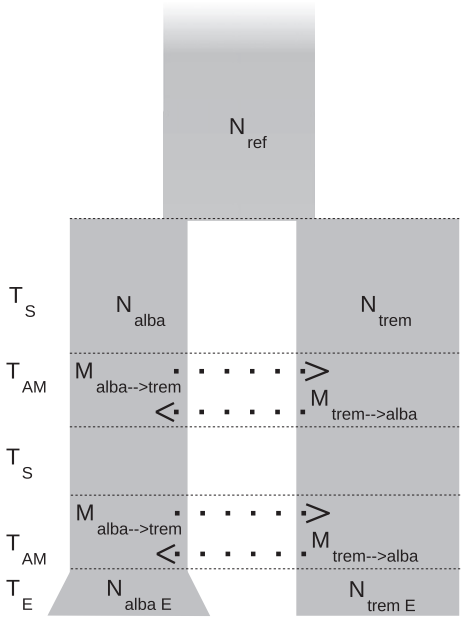

Fig. 3 Ancient migration model with two contact periods, heterogeneity in N_e and expansion (PAM2Nex), including 12 parameters: ancestral population size (N_{ref}), population sizes of *Populus alba* and *Populus tremula* after the split (N_{alba} and N_{trem}) and after expansion (N_{albaE} and N_{tremE}), time of divergence in strict isolation (T_s), time of ancient migration (T_{am}), time since expansion was initiated (T_E), locus migration rate from *P. alba* into *P. tremula* ($M_{alba \rightarrow trem}$) and in the opposite direction ($M_{trem \rightarrow alba}$). Proportion of nonrecombining regions affected by background selection (nr) and extent of population size reduction (bf) were estimated but are not illustrated in the figure.

and S4, Supporting Information). Accordingly, results from demographic modelling of individually tagged RAD-seq broadly agreed with those from pool-seq (Tables 2, S1 and S4; Fig. S4, Supporting Information).

Genomic distribution of candidate 'migrant loci'. Plotting of residuals for our best supported model (PAM2Nex) and a strict isolation model with expansion (SIex) indicated that the greatest between-model differences for shared, derived polymorphisms between *P. alba* and *P. tremula* were for variants with frequencies of 26.3–73.7% (Fig. S5, Supporting Information). A total of 23 447 SNPs along the 19 chromosomes of poplar exhibited derived allele frequencies within these bounds, and 3484 of these putative 'migrant loci' coincided with no-fixation windows identified by our differentiation genome scans (above). Generalized linear models indicated a significant association of 'migrant loci' with such no-fixation windows along the genome (slope = 1.53107 ± 0.03661 SE, $t = 41.82$, $P < 0.001$), but not with their presence or absence in centromere regions (slope = −0.12436 ± 0.07409 SE, $t = -1.678$, $P = 0.093$).

Genome fractions and gene sets affected by selection

We were able to estimate alpha for at least one of the studied populations in 25 870 of 33 733 genes covered by this sequencing effort (76.7%), thus allowing us to test for positive selection during divergence from *P. trichocarpa*. At the species level, we successfully estimated alpha for 24 530 genes in *P. alba* and 6600 of these estimates were indeed significant for positive selection. In *P. tremula*, alpha estimates were obtained for 22 990 genes, 6375 of which were significant for positive selection (Tables 3 and S6, Supporting Information).

Gene ontology annotations were available for a total of 13 348 genes with available alpha values in *P. alba* and for 12 531 genes with alpha values in *P. tremula*, respectively (Tables 3 and S6, Supporting Information). Genes with significant alpha values (=evidence for positive selection) were enriched for many *Biological Processes* GO terms, and 37% of these were overrepresented in both species (Table 4). Perhaps most conspicuously, these included genes involved in pollen recognition, response to hormones and innate immune response. The last group comprised mainly members of the nucleotide binding site (NBS) leucine-rich repeat (LRR) family of disease resistance (R-) genes, previously shown to be involved in postzygotic geneflow barriers in plants (Bomblies & Weigel 2007; Chae *et al.* 2014). While we were

Table 3 Alpha tests for evidence of positive selection in the form of adaptive amino acid substitutions in both species and localities (Ita, Italy; Hun, Hungary). Numbers of genes for which selection tests (alpha) were possible and for which significantly positive selection was observed (*G*-tests; $P < 0.05$). The number of annotated genes is indicated for the same categories, broken down by species and sampling sites

	All genes		Only GO annotated genes	
Population	No. of genes with alpha	No. of significant alpha genes	No. of genes with alpha	No. of significant alpha genes
Populus alba Ita	22 949	4860	12 547	2974
P. alba Hun	23 687	4980	12 907	3057
Populus tremula Ita	22 405	4247	12 230	2635
P. tremula Hun	20 036	4422	11 068	2760
P. alba Total	24 530	6600	13 348	3985
P. tremula Total	22 990	6375	12 531	3918

mainly interested in GO terms of the *Biological Processes* domain, we document results for all other GO domains in Table S7 (Supporting Information).

The direction of selection (DoS; Stoletzki & Eyre-Walker 2011) did not differ between species, but it differed significantly between candidate genome regions for introgression (no-fixation windows) and the remainder of the genome (Fig. 5). Whereas DoS was positive for the remainder of the genome, it was on average negative for genes in no-fixation regions (Fig. 5), thus indicating slightly more deleterious alleles segregating in these regions. These results are unlikely to be due to differences in sequencing coverage between genome fractions, as there were no significant differences in coverage between no-fixation windows and the remainder of the genome (SD overlapped), and correlations between coverage and DoS were close to zero (range: −0.0249 to +0.0015).

Discussion

It is now increasingly appreciated that great progress in adaptation and speciation genomics can be made by integrating genomic and molecular approaches with the goal of understanding how natural selection, drift and demographic processes drive the divergence of populations and species (Stapley *et al.* 2010; Feder *et al.* 2012; Jones *et al.* 2012; The Heliconius Genome Consortium

Table 4 Major gene ontology (GO) terms for biological processes with significant enrichment of genes with evidence for positive selection (adaptive amino acid substitutions) compared with all genes covered in this whole-genome resequencing effort

Significantly enriched GO term in genes with signal of positive selection	*Populus alba*		*Populus tremula*	
	# Genes	# Pos. sel. genes	# Genes	# Pos. sel. genes
Protein phosphorylation	1128	**467*** **	1089	**480*** **
Apoptotic process	202	**101*** **	197	**99*** **
Recognition of pollen	83	**47*** **	82	**51*** **
Innate immune response	53	**31*** **	55	**29*** **
Signal transduction	315	**96** *	291	91
Response to aluminium ion	17	**11** *	17	8
RNA metabolic process	1420	**407** *	1341	387
Response to hormone	25	**14** *	24	**16*** **
Oligopeptide transport	43	**21** *	41	**21** *
Negative regulation of catalytic activity	51	**24** *	52	18
ATP biosynthetic process	44	**16** *	38	11
Cell communication	418	**153** *	390	148
Protein ubiquitination	76	**30***	77	**35** *
Cation transport	288	83	269	**82***
Microtubule-based movement	56	22	51	**22***
Proteolysis	438	138	423	**152***
Intracellular signal transduction	166	38	145	**37***
Cell death	217	106	214	**108***
Ion transport	355	104	333	**107*** **

Genes, total number of annotated genes; # Pos. Sel. Genes, number of genes with evidence of outlier status for positive selection as identified by alpha tests.
Significance levels for Fisher's exact tests are abbreviated ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$, and all these significant test results are highlighted in bold.

2012; Ellegren 2014; Seehausen *et al.* 2014; Tine *et al.* 2014). Here, we have taken a three-pronged approach to explore the process of speciation in two ecologically divergent, hybridizing species of the 'model tree' genus *Populus* from a genomic perspective. Below, we shall discuss our results in a synthetic, integrative manner, including caveats of the approaches taken and perspectives for future work.

Signature of joint demographic history of hybridizing species revealed by genomewide patterns of differentiation and site frequency spectra

Our genomewide scan for differentiation and allele sharing revealed numerous genomic patterns expected for two incompletely isolated species at an advanced stage of speciation (Michel *et al.* 2010; Feder *et al.* 2012; Stölting *et al.* 2013; Christe *et al.* 2016), such as high and widespread genomic differentiation (Michel *et al.* 2010) and a large number of differentiation outliers, some of which also showed the reduction of genetic diversity expected for selective sweeps (Table 1; Fig. 1) (Nielsen 2005; Ellegren 2014). Most conspicuously, however, many genome windows in these hybridizing *Populus* taxa exhibited zero (or close to zero) fixed polymorphisms despite high levels of genomewide differentiation (Figs 1 and S1, Table S2, Supporting Information). These regions, many of which are hypothesized to stem from recurrent gene flow (see below; Stölting *et al.* 2013), were not evenly distributed in the genome with a clear trend for more frequent occurrence towards the ends of chromosomes. As recombination rates are generally higher towards chromosome ends in animals and plants (Pigozzi 2008; Ren *et al.* 2012; Tortereau *et al.* 2012), these results are highly suggestive of a role for structural genomic features in shaping genomewide patterns of allele sharing. Also, genomic differentiation (AFD) was visibly elevated around centromeres (Figs 1 and S1, Supporting Information), consistent with structural features of genomes driving genomewide differentiation, as recently observed in sunflowers (Renaut *et al.* 2013) and flycatcher birds (Burri *et al.* 2015). In the same context, similarities and differences between allele frequency differentiation (measured as AFD in our study) and absolute divergence (Dxy) have recently attracted considerable interest in connection with genomic variation in recombination rates and diversity (Cruickshank & Hahn 2014; Burri *et al.* 2015). Although our present pool-seq genome scan is certainly not ideal for an in-depth analysis of these issues, we note that Dxy in our study was weakly but significantly correlated with AFD (e.g. Italian locality: Pearson's $r = 0.398$, $P < 0.001$) and that many low-fixation regions revealed by our study exhibited a slight reduction in Dxy (Figs 1 and S1, Supporting Information), consistent with an erosion of the differentiation by secondary gene flow.

Results of our genome scan for differentiation and allele sharing were largely concordant with those from demographic modelling of genomewide SFS (Fig. 2). For example, our extensive model comparisons consistently revealed ancient (AM) and repeated ancient geneflow (PAM) scenarios as superior to simple geneflow or strict isolation models of demographic history (Table 2). This is perfectly in line with the small sizes of low-fixation islands seen in these species' genomes, which are suggestive of past gene flow (Bierne *et al.* 2011; Stölting *et al.* 2013; Christe *et al.* 2016). Adding demographic expansion in the recent past greatly increased model support, in agreement with available knowledge of the pleistocene demographic history of temperate forest trees with Eurasian distributions (Hewitt 2000; Tzedakis *et al.* 2013). Heterogeneous geneflow models (e.g. PAM2M2P) received stronger support than models that did not implement heterogeneity along the genome, as expected when genomes are 'porous' and isolated by a semipermeable barrier to gene flow (Barton & Hewitt 1985; Harrison 1993; Wu & Ting 2004). Note that the most recent secondary contact event between these species has been fully documented by analysing local ancestries across the full genomic admixture gradient present in these hybrid zone localities (Christe *et al.* 2016). Models implementing heterogeneity in N_e also received strong model support, and in fact, the most strongly supported model included heterogeneity in N_e (Fig. 4; Table 2). Strong support for these models was expected, as our genome scan already indicated considerable genomewide variation in differentiation, coinciding with structural features of genomes such as repetitive DNA and centromeres.

Demographic parameters within our best supported model (repeated ancient gene flow with heterogeneity in N_e along the genome and expansion; PAM2Nex; Fig. 3) yielded sensible results with regard to time periods and rates of ancient gene flow and provided an estimate for the timing of divergence between these species: *P. alba* and *P. tremula* appear to have diverged *c.* 2.8 Mio years and experienced low amounts of repeated ancient gene flow (Table S5, Supporting Information). Our estimate of divergence time lies within the age range of fossils from species of this section of the genus *Populus* (Eckenwalder 1996). We note that we may have slightly under- or overestimated divergence time by assuming a generation time of 20 years, which is a nontrivial assumption in perennial plants with overlapping generations, long reproductive life phase and the ability to persist clonally (Petit & Hampe 2006; van Loo *et al.* 2008).

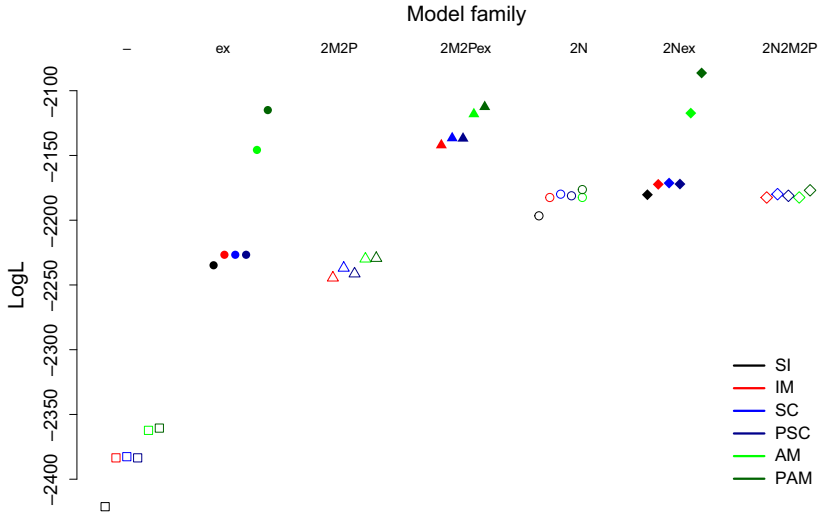

Fig. 4 Graphical comparison of statistical support for different families of demographic models examined with dadi based on pool-seq data. Models grouped within model families are shown along the horizontal axis, and log likelihoods (logL) along the vertical axis. Models within each model family include SI (strict isolation), IM (isolation with migration), SC (secondary contact), PSC (repeated secondary contact), AM (ancient gene flow) and PAM (repeated ancient gene flow). Model families include basic models (-), models with expansion (ex), models implementing heterogeneity in migration (2M2P), models including both of these features (2M2Pex), heterogeneity in effective population size (2N), heterogeneity in effective population size and expansion (2Nex), and heterogeneity in both migration and population size (2N2M2P). For exact model definitions, see scripts available as Appendix S1 (Supporting information).

Genomic footprints of selection

Positive selection is widely expected to contribute to species differentiation during DWGF (Coyne & Orr 2004; Presgrave; Elmer & Meyer 2011; Smadja & Butlin 2011; The Marie Curie SPECIATION Network 2012; Seehausen *et al.* 2014; Burri *et al.* 2015); thus, we used our genome-scale data to test for selective sweeps and adaptive protein evolution. We found between 885 and 1039 selective sweeps (lnRH outliers) in each locality and species (Table 1). In *P. alba*, the species with lower migration rates and stronger population structure (Lexer *et al.* 2005; Christe *et al.* 2016), *c.* 100 of these were also associated with significant AFDs between species in each locality, whereas this number was lower in *P. tremula*, known to be a 'high geneflow species' (de Carvalho *et al.* 2010). The results are suggestive of a role for positive selection in maintaining species differentiation. Nevertheless, sweep signatures based on diversity are ephemeral (Nielsen 2005) and are thus unlikely to cover the long timescales relevant to speciation and divergence in these species.

Our estimates of alpha (the proportion of amino acid substitutions fixed by positive selection; Smith & Eyre-Walker 2002), on the other hand, revealed at least several thousands of genes affected by selection during divergence of each species from *P. trichocarpa*, used as a reference in this study (Table 3). Genes with significant alpha were enriched for numerous GO terms of the *Biological Processes* domain, including genes involved in pollen recognition, response to hormones including auxin, and innate immune response (Table 4). Only 103 of 2405 genes with significant alpha and available GO annotations also coincided with high-differentiation (AFD) outlier windows in our genome scan, which suggests that the selection we detected often predates the split between *P. alba* and *P. tremula*. This makes sense, as our comparison involved *P. trichocarpa* as an outgroup, which diverged from *P. alba* and *P. tremula* for *c.* 5–10 Mio years, judging from the fossil record (Eckenwalder 1996) and molecular data (Ingvarsson 2005).

As expected (Stoletzki & Eyre-Walker 2011), DoS provided a clearer picture of the direction of selection: whereas DoS was generally positive throughout the genome fraction exhibiting some fixation between species (Fig. 5; 95% CIs did not include zero), confirming a substantial proportion of proteins evolved adaptively in these species, it was negative for 'no-fixation' candidate regions for introgression (Fig. 5; difference significant in both species). Negative DoS is expected when slightly deleterious nonsynonymous mutations segregate in populations. This result matches the prediction that introgression (inferred by no-fixation windows) should not only be enhanced in genome regions devoid of barrier loci, but also in regions bearing segregating

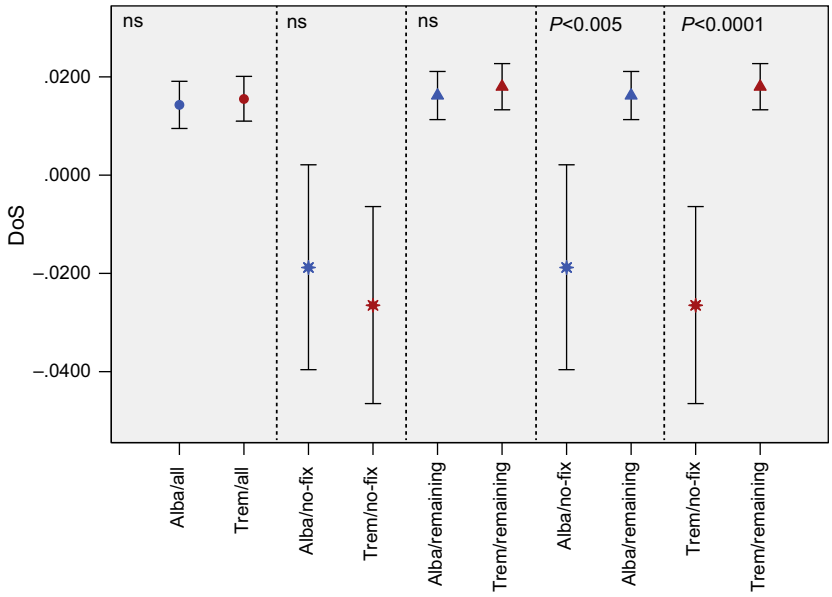

Fig. 5 Direction of selection (DoS) for different comparisons of species and genome fractions in the Italian locality, including means and 95% confidence limits of the means. The comparisons include *Populus alba* (alba, in blue) and *Populus tremula* (trem, in red) and the genome fractions 'all windows' (all, indicated by circles), 'no-fixation windows only' (no-fix, indicated by stars), and 'all remaining windows' (remaining, indicated by triangles). Pairwise comparisons are separated by vertical dashed lines. Results of Welch's *t*-tests are indicated for each comparison. The significant change from positive to negative DoS comparing no-fixation windows (candidate regions for introgression) and the remaining genome windows in each species is clearly visible. The analyses were based on matching data sets of 19 184 genome windows for interspecific comparisons. Of these, 1045 of windows were free of fixed allelic differences between *P. alba* and *P. tremula*, and the remaining 18 139 windows exhibited some degree of fixation between species.

deleterious alleles. Such mutations are expected to produce dominance heterosis in F_1s and in heterozygous genome tracts of later-generation introgressants, thus contributing to the weakening of barriers and widespread introgression in both animals and plants (Ingvarsson & Whitlock 2000; Bierne *et al.* 2002; Harris & Nielsen 2015).

The 'innate immune response' GO term captured by our genome scan for adaptive protein evolution included 30 members of the well-known NBS-LRR family of plant disease resistance (R-) genes (Table S7, Supporting Information). These genes are known to have broad and important functional roles in the immune response of plants (McHale *et al.* 2006) including trees (Tuskan *et al.* 2006; Tobias & Guest 2014). Interestingly, members of the same subfamily of NBS-LRR genes have previously been demonstrated to be involved in strong postzygotic geneflow barriers in plants, caused by sublethal or lethal autoimmune responses due to deleterious epistasis in hybrids between divergent populations (Bomblies & Weigel 2007; Chae *et al.* 2014). In fact, these 'hybrid necrosis' genes are among the best characterized speciation genes in plants (Rieseberg & Blackman 2010). We note that hybrid seedlings of *P. alba* and *P. tremula* often are necrotic, and genotyping of inviable first-year seedlings from the Italian hybrid zone revealed that up to 100% of them were genetically intermediate hybrids (A. Jordan, L. Bresadola, C. Lexer, unpublished data). This is in stark contrast to the broad range of different genotypes commonly found among successfully germinated survivors from the same locality (Lindtke *et al.* 2014).

Of 30 NBS-LRR genes with significant positive selection detected by our study (Table S6, Supporting Information), 26 had positive alpha in both *P. alba* and *P. tremula*, and one was located in a high-differentiation (AFD) outlier window detected by our genome scan. This suggests that genes with potential involvement in RI do not necessarily need to exhibit elevated interspecific differentiation. In the case of plant R-genes – and immune system genes more generally – negative frequency dependence leading to balancing selection can maintain alleles in gene pools for a range of different timescales (Bakker *et al.* 2006; Charlesworth 2006; Bomblies & Weigel 2007), likely including trans-species polymorphisms. We note that most of the NBS-LRRs under long-term positive selection in our study exhibited significantly increased nucleotide diversity compared with genomewide expectations (Table S8, Supporting Information) and that balancing selection has previously been detected for markers tagging NBS-LRR genes in these species (Caseys *et al.* 2015b).

Speciation with recurrent gene flow: an intrinsic consequence of adaptive divergence?

Studies of DWGF often assume that adaptation to different environmental conditions drives or facilitates the evolution of RI, and for many organisms studied by evolutionary biologists, this assumption is indeed supported by convincing data (Coyne & Orr 2004; Hoekstra *et al.* 2006; Elmer & Meyer 2011; Smadja & Butlin 2011; Baack *et al.* 2015). On the other hand, it has long been known that intrinsic (=endogenous) barriers to gene flow are more likely to result in strong RI (Barton & Bengtsson 1986), and theory predicts the 'coupling' of environmental (=exogenous) and intrinsic barriers under a wide range of ecological and spatial settings (Barton & De Cara 2009; Bierne *et al.* 2011). This raises the question as to what extent the steep spatial or genomic clines often seen between ecologically divergent, hybridizing species (Barton & Hewitt 1985; Bierne *et al.* 2011) including *P. alba* and *P. tremula* (Lexer *et al.* 2010; Lindtke *et al.* 2012; Christe *et al.* 2016) have arisen/are currently maintained by environmental vs. intrinsic mechanisms.

In the case of the ecologically divergent *P. alba* and *P. tremula*, patterns of local ancestry along hybrid genomes and differences in juvenile survivorship in a common garden trial indicate strong intrinsic barriers due to deleterious epistasis in recombinant hybrids (Lindtke *et al.* 2014; Christe *et al.* 2016), consistent with DM incompatibilities and/or intrinsic coadaptation of genes in genetic or biochemical pathways (Caseys *et al.* 2015a; Lindtke & Buerkle 2015). An important role for intrinsic barriers is also suggested by extremely steep genomic clines and the presence of epistatic interactions in natural hybrids (Lexer *et al.* 2010). Our present study supports a scenario in which adaptive divergence affects the evolution and maintenance of RI in hybridizing species primarily via intrinsic and demographic processes, rather than ecologically divergent selection counteracting gene flow.

Genome scanning and demographic modelling both agree that these species experienced low levels of gene flow repeatedly across extended time periods and that this process affected their genomes in a heterogeneous way, thus resulting in many small low-fixation 'pores' preferentially located in genome regions likely experiencing non-negligible levels of recombination (Fig. 1). These consistently recovered ancient/repeated gene-flow scenarios agree well with biogeographic and genomic data supporting the existence of spatially separated ice age refugia, with southern European refugia for the thermophilic *P. alba* and Central European ones for the cold-tolerant *P. tremula* (Fussi *et al.* 2010; Christe *et al.* 2016). In effect, their divergent ecological preferences appear to 'pull' these species to separate refugial areas in glacial periods as they track their different optima (Davis & Shaw 2001), thus allowing for extended periods of divergence (Hewitt 2000; Tzedakis *et al.* 2013), followed by episodes of secondary contact in interglacials involving strong selection against recombinants due to deleterious epistasis (Lindtke *et al.* 2014; Christe *et al.* 2016). In Europe, genetic contact occurs primarily in localities where hybrid zones of these and other species get trapped by environmental clines (Bierne *et al.* 2011), for example in the upstream portions of Eurasian river valleys in the case of these two forest trees (Lexer *et al.* 2010; Lindtke *et al.* 2012).

Selection apparently affected both divergence and episodic introgression during these species' evolutionary history, and it likely shaped their genomic landscape of divergence in nontrivial ways. For example, our results support the prediction that regions bearing slightly deleterious variants had an elevated chance to cross the species barrier by triggering heterosis (Bierne *et al.* 2002) possibly by early pseudo-overdominance, while migrant tracts are sufficiently long to mask several recessive detrimentals (Harris & Nielsen 2015). The hypothesis that segregating load can enhance introgression has thus far been largely overlooked, but this hypothesis can now be tested with genomic data. We do currently not know whether some of the introgressed alleles in these poplar species reflect adaptive introgression. Rigorous tests of this hypothesis would require genomewide scans of the recipient species for positively selected variants at different distances from contact zones. With regard to the maintenance of RI in secondary contact, our many positively selected NBS-LRR 'innate immune response' genes (Table S6, Supporting Information) may be a case in point, as members of this subfamily of plant R-genes are known to act as potent 'speciation genes' by triggering deleterious epistatic autoimmune phenotypes when recombined in hybrids (Bomblies & Weigel 2007; Rieseberg & Blackman 2010; Chae *et al.* 2014). Thus, this group of genes appears to represent an example of ecologically important loci (host–pathogen interactions) that nevertheless exert their effects on RI and plant speciation via essentially intrinsic, nonecological mechanisms. Gathering direct evidence for an involvement of NBS-LRR's (and other groups of genes) in RI in poplars would benefit from testing the phenotypic and fitness effects of different alleles and haplotypes in common garden trials, and identifying the most relevant mutations by functional assays (Plett *et al.* 2014; Suarez-Gonzalez *et al.* 2016). This could be complemented by proteomic and metabolomic approaches (e.g. Caseys *et al.* 2015a) to understand the effects of specific functional gene sets

and pathways on RI. In poplars and many other taxa, this combination of approaches may ultimately reveal the precise mechanisms by which ecology exerts its impact on the origin and maintenance of biological diversity.

Acknowledgements

We thank Stefano Castiglione, Stefano Gomarasca, Denes Bartha and István Asztalos for help with field work, Alexa Oppliger, Thelma Barbará, and Aurore Jordan for help in the laboratory, the Functional Genomics Centre Zurich (FGCZ) and the Lausanne and Berne genomics facilities for DNA sequencing, Pierre Alexandre Gagnaire, Yoann Anciaux, Sylvain Glemin, Ryan Gutenkunst, Berthold Heinze, Alex Widmer, and Alex Buerkle for helpful discussions, Luca Ferretti for explaining specifics of the NPSTAT software, Vital-IT and the University of Bern computing centre for computational support, and three anonymous reviewers for insightful comments. This work was supported by SNF grants PDFMP3_134660, the 31003A_149306 of the Swiss National Science Foundation (SNSF) to CL and the Agence Nationale de la Recherche (HYSEA project, ANR-12-BSV7- 0011) to CF and NB.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Alexa A, Rahnenfuhrer J (2010) *topGO: topGO: Enrichment analysis for Gene Ontology*. R package version 2.22.0.

Arnold ML (2006) *Evolution Through Genetic Exchange*. Oxford University Press, Oxford, UK.

Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D (2015) The origins of reproductive isolation in plants. *New Phytologist*, **207**, 968–984.

Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. *The Plant cell*, **18**, 1803–1818.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57**, 357–376.

Barton NH, De Cara MAR (2009) The evolution of strong reproductive isolation. *Evolution*, **63**, 1171–1190.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Beaumont MA (2010) Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Bernhardsson C, Robinson KM, Abreu IN *et al.* (2013) Geographic structure in metabolome and herbivore community co-occurs with genetic structure in plant defence genes. *Ecology Letters*, **16**, 791–798.

Bierne N, Lenormand T, Bonhomme F, David P (2002) Deleterious mutations in a hybrid zone: can mutational load decrease the barrier to gene flow? *Genetics Research*, **80**, 197–204.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it...? why are F_{ST} outliers sometimes so frequent? *Molecular Ecology*, **22**, 2061–2064.

Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting selective sweeps from pooled next-generation sequencing samples. *Molecular Biology and Evolution*, **29**, 2177–2186.

Bomblies K, Weigel D (2007) Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics*, **8**, 382–393.

Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, **25**, 1656–1665.

de Carvalho D, Ingvarsson PK, Joseph J *et al.* (2010) Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Molecular Ecology*, **19**, 1638–1650.

Caseys C, Stritt C, Glauser G, Blanchard T, Lexer C (2015a) Effects of hybridization and evolutionary constraints on secondary metabolites: the genetic architecture of phenylpropanoids in european Populus species. *PLoS ONE*, **10**, e0128200.

Caseys C, Stölting KN, Barbará T, González-Martínez SC, Lexer C (2015b) Patterns of genetic diversity and differentiation in resistance gene clusters of two hybridizing European Populus species. *Tree Genetics & Genomes*, **11**, 81.

Chae E, Bomblies K, Kim S-T *et al.* (2014) Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell*, **159**, 1341–1351.

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

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

Christe C, Stölting KN, Bresadola L *et al.* (2016) Selection against recombinant hybrids maintains reproductive isolation in hybridizing Populus species despite F1 fertility and recurrent gene flow. *Molecular Ecology*, **25**, 2482–2498.

Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, Massachusetts.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. *Science*, **292**, 673–679.

Der Sarkissian C, Ermini L, Schubert M *et al.* (2015) Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, **00**, 2577–2583.

Dickmann D, Kuzovkina YA (2008) Poplars and Willows in the World. FAO. Poplars and Willows in the World: Meeting the needs of society and the environment. International Poplar Commission 9-2, FAO, Rome, Italy.

Du S, Wang Z, Ingvarsson PK *et al.* (2015) Multilocus analysis of nucleotide variation and speciation in three closely related *Populus* (Salicaceae) species. *Molecular Ecology*, **24**, 4994–5005.

Eckenwalder JE (1996) Systematics and evolution of *Populus*. In: *Biology of Populus and its Implications for Management and Conservation, Part I* (eds Stettler RF *et al.*), pp. 7–32. NRC Research Press, Ottawa.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, **29**, 51–63.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

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

Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 461–474.

Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P (2013) Geographic mode of speciation and genomic divergence. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 73–97.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.

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

Geraldes A, Farzaneh N, Grassa CJ *et al.* (2014) Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, **68**, 3260–3280.

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

Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, **9**, e1003521.

Harris K, Nielsen R (2015) The genetic cost of neanderthal introgression. *Genetics*, **203**, 881–891.

Harrison RG (ed.) (1993) *Hybrid Zones and the Evolutionary Process*. Oxford University Press, New York, New York.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics*, **169**, 945–953.

Ingvarsson PK, Whitlock MC (2000) Heterosis increases the effective migration rate. *Proceedings of the Royal Society*, **267**, 1321–1326.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Juric I, Aeschbacher S, Coop G (2015) The strength of selection against Neanderthal introgression. *bioRxiv*. doi: 10.1101/030148

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

Lexer C, Joseph JA, van Loo M *et al.* (2010) Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. *Genetics*, **186**, 699–712.

Lindtke D, Buerkle CA (2015) The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution*, **69**, 1987–2004.

Lindtke D, Buerkle CA, Barbará T *et al.* (2012) Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in *Populus*. *Molecular Ecology*, **21**, 5042–5058.

Lindtke D, Gompert Z, Lexer C, Buerkle CA (2014) Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. *Molecular Ecology*, **23**, 4316–4330.

van Loo M, Joseph JA, Heinze B, Fay MF, Lexer C (2008) Clonality and spatial genetic structure in *Populus* x canescens and its sympatric backcross parent *P. alba* in a Central European hybrid zone. *New Phytologist*, **177**, 506–516.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.

McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biology*, **7**, 1–11.

Michel AP, Sim S, Powell THQ *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of Americas*, **107**, 9724–9729.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 187–214.

Pigozzi MI (2008) Relationship between physical and genetic distances along the zebra finch Z chromosome. *Chromosome Research*, **16**, 839–849.

Pinho C, Hey J (2010) Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 215–230.

Plett JM, Williams M, LeClair G, Regan S, Beardmore T (2014) Heterologous over-expression of ACC SYNTHASE8 (ACS8) in *Populus tremula* x *P. alba* clone 717-1B4 results in elevated levels of ethylene and induces stem dwarfism and reduced leaf size through separate genetic pathways. *Frontiers in Plant Science*, **5**, 514.

Ren Y, Zhao H, Kou Q *et al.* (2012) A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS ONE*, **7**, e29453.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Renaut S, Owens GL, Rieseberg LH (2014) Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Molecular Ecology*, **23**, 311–324.

Rieseberg LH, Blackman BK (2010) Speciation genes in plants. *Annals of Botany*, **106**, 439–455.

Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D (2014) The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, **23**, 3944–3956.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution*, **30**, 1574–1587.

Rubin C-J, Zody MC, Eriksson J *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, **464**, 587–591.

Schlötterer C, Dieringer D (2005) A novel test statistic for the identification of local selective sweeps based on microsatellite gene diversity. In: *Selective Sweep* (ed. D Nurminsky), pp. 55–64. Landes Bioscience, Georgetown, Texas.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, **60**, 957–964.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Slavov GT, DiFazio SP, Martin J *et al.* (2012) Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, **196**, 713–725.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. *Nature*, **415**, 1022–1024.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Molecular Biology and Evolution*, **28**, 63–70.

Stölting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.

Stölting KN, Paris M, Heinze B *et al.* (2015) Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of *Populus alba* (Salicaceae), a widespread forest tree. *New Phytologist*, **207**, 723–734.

Suarez-Gonzalez A, Hefer C, 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.

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

The Marie Curie SPECIATION Network (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27–39.

Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.

Tobias PA, Guest DI (2014) Tree immunity: growing old without antibodies. *Trends in Plant Science*, **19**, 367–370.

Tortereau F, Servin B, Frantz L *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, **13**, 586.

Turner TL, Bourne EC, Von Wettberg EJ *et al.* (2010) Population resequencing reveals local adaptation o *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Tuskan GA, Difazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Tzedakis PC, Emerson BC, Hewitt GM (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology & Evolution*, **28**, 696–704.

Whitham TG, Bailey JK, Schweitzer JA *et al.* (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics*, **7**, 510–523.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu C-I, Ting C-T (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114–122.

C.C., K.N.S., N.B. and C.L. conceived the study; C.C., K.N.S., and C.L. gathered the data; C.C., K.N.S., M.P., and C.F. analyzed the data; C.C., K.N.S. and C.L. wrote the manuscript with input and revisions from all co-authors.

Data accessibility

Extensive data sets and scripts are available on DRYAD (provisional doi:10.5061/dryad.3bc76). Raw sequence reads have been submitted to NCBI SRA (Accession ID SRP076693).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1. Chromosome-wide patterns of differentiation (allele frequency differentials: AFDs, black), sequence divergence (Dxy, green), and genomic features (fraction of repetitive DNA, grey) for all 19 *Populus* chromosomes in Italy and Hungary.

Fig. S2. Observed and expected pairwise distances between no-fixation regions for interspecific comparisons of *Populus alba* and *P. tremula* in two sampling localities (Italy, Hungary).

Fig. S3. Demographic modelling of RAD-seq data for *P. alba* and *P. tremula* for the Italian hybrid zone locality.

Fig. S4. Graphical comparison of statistical support for different families of demographic models examined with dadi based on RAD-seq data.

Fig. S5. Comparison between the strict isolation model with expansion (SIex) and best model (PAM2Nex) based on pool-seq data.

Table S1. Results of model fitting for the 39 alternative models of divergence for pool-seq data.

Table S2. Filtered windowed results for 36 918 well covered windows of 8 kb length along the Populus trichocarpa genome.

Table S3. Size and distribution of low-fixation regions in Italy and Hungary.

Table S4. Results of model fitting for the 39 alternative models of divergence for RAD-seq data.

Table S5. Parameter estimates (time periods scaled in years) for the best model PAM2Nex, the ancient migration model with two periods of contact, heterogeneity of Ne and recent expansion with means and standard deviations (stdev) derived from 100 bootstrapped datasets.

Table S6. Description and gene ontology annotations of the 14 018 genes analyses for alpha.

Table S7. Major gene ontology (GO) terms with significant enrichment among genes with evidence of positive selection compared to all genes covered in this whole-genome resequencing effort.

Table S8. Allele frequency differentials (AFD), nucleotide diversity (π), and Tajima's D for NBS-LRR genes with significant alpha (positive selection).