Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species

Dorothea Lindtke[1,2], Zachariah Gompert[3], Christian Lexer[2], and C. Alex Buerkle[1]

[1] Department of Botany, University of Wyoming, Laramie, WY 82071, USA

[2] Unit of Ecology and Evolution, Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland

[3] Department of Biology and Ecology Center, Utah State University, Logan, UT 84322, USA

Corresponding author: Dorothea Lindtke
1000 E. University Ave.
Department of Botany, 3165
University of Wyoming
Laramie, WY 82071, USA
dlindtke@uwyo.edu
Fax: 307-766-2851

Keywords: paternity, parentage, admixture, reproductive isolation, Bayesian inference, next-generation sequencing

Running title: Hybridization frequency in *Populus*

Abstract

In the context of potential interspecific gene flow, the integrity of species will be maintained by reproductive barriers that reduce genetic exchange, including traits associated with prezygotic isolation or poor performance of hybrids. Hybrid zones can be used to study the importance of different reproductive barriers, particularly when both parental species and hybrids occur in close spatial proximity. We investigated the importance of barriers to gene flow that act early versus late in the life cycle of European *Populus* by quantifying the prevalence of homospecific and hybrid matings within a mosaic hybrid zone. We obtained genotypic data for 11,976 loci from progeny and their maternal parents and constructed a Bayesian model to estimate individual admixture proportions and hybrid classes for sampled trees, and for the unsampled pollen parent. Matings that included one or two hybrid parents were common, resulting in admixture proportions of progeny that spanned the whole range of potential ancestries between the two parental species. This result contrasts strongly with the distribution of admixture proportions in adult trees, where intermediate hybrids and each of the parental species are separated into three discrete ancestry clusters. The existence of the full range of hybrids in seedlings is consistent with weak reproductive isolation early in the life cycle of *Populus*. Instead, a considerable amount of selection must take place between the seedling stage and maturity to remove many hybrid seedlings. Our results highlight that high hybridization rates and appreciable hybrid fitness do not necessarily conflict with the maintenance of species integrity.

Introduction

Species are thought to arise through the accumulation of traits that contribute to reproductive isolation (RI) and evolutionary independence, with intermediate stages of speciation achieving some evolutionary independence even in the presence of some amount of hybridization and gene flow (Wu, 2001; Mallet, 2005). One of the central questions in evolutionary biology thus concerns the understanding of the origin and maintenance of reproductive barriers that reduce genetic exchange to a level that creates or maintains independent groups of organisms in primary or secondary contact (Coyne & Orr, 1998; Turelli *et al.*, 2001; Ortiz-Barrientos *et al.*, 2002). Barriers that reduce gene flow between populations or species are typically categorized according to timing (pre- or postzygotic) or mechanism (intrinsic or extrinsic). Early-acting reproductive barriers are often thought to be more important to total RI than late-acting barriers (Kirkpatrick & Ravigne, 2002; Ramsey *et al.*, 2003; Rieseberg & Willis, 2007; Lowry *et al.*, 2008), although the contrary might equally be true (Coyne & Orr, 1998, 2004; Bomblies & Weigel, 2007).

Incomplete species barriers resulting in (occasional) hybridization and gene flow between diverging lineages are common in nature, often without threatening species integrity (Wu, 2001; Mallet, 2005; Mallet *et al.*, 2007; Butlin *et al.*, 2008; Abbott *et al.*, 2013). One of the major explanations for the maintenance of species differences in the absence of complete geographic or prezygotic isolation is that hybrids show poor performance due to extrinsic or intrinsic factors and are thus quickly eliminated by selection (e.g., Wu, 2001; Schluter, 2001). However, theoretical and empirical data indicate that some hybrids can be as fit or even fitter than their parental species (Arnold & Hodges, 1995; Barton, 2001). This can weaken interspecies RI through the opportunity for recombination between divergent traits or incompatible loci in hybrids. For example, loci involved in assortative mating and differential adaptation can recombine or non-detrimental allele combinations of genetic incompatibilities can be recovered (Kirkpatrick & Ravigne, 2002; Ortiz-Barrientos *et al.*, 2002; Bank *et al.*, 2012). This might lead to genetic swamping and loss of evolutionary independence (Seehausen, 2004; Mallet, 2005; Seehausen *et al.*, 2008). In contrast, hybridization can also be a creative force in evolution through adaptive introgression or hybrid speciation (Rieseberg *et al.*, 2003; Seehausen, 2004; Jiggins *et al.*, 2008; Abbott *et al.*, 2013). The fate of hybridizing species can therefore depend strongly on the fitness of interspecific genotypes, as well as the genetic architecture, strength, and type of isolation barriers between diverging lineages.

Studies of hybrid zones are expected to contribute to our understanding on the establishment and maintenance of RI and species differences in the face of gene flow (Barton & Hewitt, 1989; Buerkle & Lexer, 2008; Payseur, 2010; Gompert *et al.*, 2012b). Particularly 'mosaic' hybrid zones can provide an excellent opportunity to study RI, as the effect of ecogeographic isolation on species barriers is diminished by the close spatial proximity of parental species and their hybrids (Harrison & Rand, 1989; Vines *et al.*, 2003). Although divergent habitat preferences or habitat-associated fitness could explain the frequently observed patchy distribution of different hybrid classes, several mosaic hybrid zones lack clear evidence for habitat-associated effects (Nosil *et al.*, 2005). The existence of fit hybrids in such a nearly sympatric setting will continuously challenge the maintenance of reproductive barriers between species that are threatened by the homogenizing effects of gene flow and recombination (e.g., Buerkle *et al.*, 2000; Ortiz-Barrientos *et al.*, 2002).

Mosaic hybrid zones between *Populus alba* (white poplar) and *P. tremula* (European aspen) exhibit such a challenging setting. The parental source populations are ecologically and morphologically well differentiated, with differences in habitat type (lowland flood-plains for *P. alba* and upland for *P. tremula*), latitudinal distributions, and in leaf morphology associated with ecology (Lexer *et al.*, 2005, 2009; Dickmann & Kuzovkina, 2008). Despite these differences, species distributions overlap considerably along European river systems, where *P. alba* and *P. tremula* can form large mosaic hybrid zones (Lexer *et al.*, 2005, 2010). Diploid hybrids, sometimes referred to as *P.* ×*canescens* (gray poplar), are frequently found as pioneers in disturbed habitats (e.g. river flood-plains; natural or anthropogenic disturbance), where hybrids and parental trees often grow within tens of meters (van Loo *et al.*, 2008; Lindtke *et al.*, 2012). Comprehensive genetic analyses of these *Populus* hybrid zones have indicated that, even in such sympatric settings, parental species and their hybrids are separated into three distinguishable ancestry groups (Lexer *et al.*, 2010; Lindtke *et al.*, 2012). Hybrids are considered to be highly recombinant, and are genetically and phenotypically mainly intermediate between their parental species (Lexer *et al.*, 2009, 2010). Few mature trees show backcross-like genotypes (those that are likely backcrosses more commonly have *P. alba* as the non-hybrid parent), and F_1 hybrid genotypes are either very rare or absent (Lexer *et al.*, 2009, 2010; Lindtke *et al.*, 2012). Previous work indicates that hybrids consist only of a subset of possible genotypic combinations, pointing to an important role of genome interactions in RI (Lindtke *et al.*, 2012).

In this study, we quantify the abundance of homospecific and hybrid matings in a *Populus* hybrid zone, using an approach related to parentage analysis. Although parentage analysis is a powerful tool for evolutionary biology and hybrid zone research, it requires exhaustive sampling of potential parents, a task that would be prohibitive in a wind-pollinated, dioeceous tree like *Populus*. Instead, we use DNA sequence data for thousands of loci to estimate genetic ancestries for hundreds of sampled trees, including 17 maternal trees and their open pollinated progenies. The genetic data from mothers and progeny allow us to infer the genetic ancestries of the unknown and unsampled paternal trees (pollen donors) within a hierarchical Bayesian model.

By studying mating patterns in a mosaic hybrid zone between *P. alba* and *P. tremula*, we aim at a better understanding of the maintenance of species despite hybridization and potential interspecific gene flow. Our main goal is to infer the timing of RI and selection responsible for the observed separation of adult trees and hybrids into three discrete ancestry clusters. This distribution of ancestry in adult trees could arise from selection that acts early in the life cycle, before or during seed formation, or later, between seedling establishment and maturity. We argue that random or non-random paternity with respect to the genetic ancestry of the mother will help to reveal the timing of RI and selection. Non-random, assortative paternity in seedlings would provide evidence for early-acting RI and might arise if mothers mate assortatively with fathers with similar ancestry, or through lethality of heterospecific embryos. By contrast, random mating patterns would indicate ineffective early-acting RI and that the remarkable genetic discontinuities present in these hybrid zones likely arise from substantial viability selection at later stages of the life cycle.

Methods

Sampling

The study population is part of a large, natural, mosaic hybrid zone between *P. alba* and *P. tremula* within the protected area of the Parco Lombardo della Valle del Ticino in northern Italy. Here trees of each of the parental species often grow within a few meters from hybrids, in a natural flood-plain habitat. The genetic structure of adult trees within this hybrid zone has been characterized previously (Lindtke *et al.*, 2012). Poplars are dioecious, wind-pollinated trees, allowing sampling of large maternal, open pollinated families. Both flowering phenology and ecogeography are unlikely to contribute substantially to RI between species, as *P. alba* and *P. tremula* have overlapping flowering times (Lauber *et al.*, 2012), and wind-dispersed pollen is expected to travel on average a few hundreds of meters within the genus (Bialozyt, 2012). Although the principal hybrid zone stretches over 20 km along the river flood-plain (N $45^\circ\,18'\,23.5''$ E $8^\circ\,55'\,11.4''$ to N $45^\circ\,10'\,36.8''$ E $9^\circ\,08'\,43.2''$), we restricted our sampling of all but two reproductive maternal trees to a mixed stand within 1.3 × 0.7 km (N $45^\circ\,17'\,18.4''$ to N $45^\circ\,18'\,0.8''$, E $8^\circ\,55'\,43.8''$ to E $8^\circ\,56'\,15.8''$; the core study stand) to ensure that all trees had a similar set of potential mates available (Fig. 1).

We collected seeds by cutting catkins from 15 maternal trees in spring 2011 at the time when fruits burst and seeds were mature, using scissors on a telescopic pole. Seeds were separated from wool and subsequently germinated in the greenhouse in covered trays on a 5–10 mm sand layer on top of autoclaved common garden soil. To monitor germination success and survival, seedlings were counted seven and 14 days after sowing. For each family, leaf tissue, or the entire seedling, was collected and stored in silica gel prior to DNA extraction. Because most putative *P. alba* trees produced very few seeds in 2011, progeny samples of two *P. alba* trees that were collected and germinated during a pilot study in 2010 were also included (17 maternal trees total). We collected leaf tissue from all mother trees and included population reference samples from within and outside the hybrid zone (40 adult trees each of *P. alba*, *P. tremula* and intermediate hybrids; a total of 120 reference individuals). Reference samples were collected in 2008 and have been previously characterized by 77 microsatellite markers (Lindtke *et al.*, 2012).

Genetic data

DNA extraction, library preparation and Illumina sequencing

We obtained genetic data from all 17 mother trees, 484 progeny samples (8 to 34 samples per family, Table 1), and 120 population reference samples. Genomic DNA was extracted from approximately 7 mg of silica gel-desiccated plant tissue using Qiagen's DNeasy 96 Plant Kit and DNeasy Plant Mini Kit (Qiagen Inc.). We generated restriction fragment libraries for each individual (with each maternal sample replicated four times) using a protocol described in detail in Parchman *et al.* (2012). Briefly, approximately 250 ng of genomic DNA was digested with endonucleases *EcoRI* and *MseI* (NEB Inc.). Fragments were ligated to double stranded adaptor sequences including the Illumina sequencing adaptor and, for the *EcoRI* adaptor, 8–10 base pairs (bp) of unique barcode for each individual. Adaptor-ligated fragments were PCR amplified with one reaction per individual using Illumina PCR primers. We combined barcoded PCR products of maternal and population reference samples and all progeny into two libraries, purified them by ethanol precipitation, and size selected them on a 2.5% agarose gel. Fragments 250–400 bp length were excised and purified using Qiagen's QiaQuick gel extraction kit (Qiagen Inc.). Single-end sequencing was performed on an Illumina HiSeq 2000 platform at the Lausanne Genomic Technologies Facility using one lane per library.

Sequence curation and assembly

DNA sequences from both libraries were passed through Illumina quality filters to remove low quality reads. Additionally, sequence reads were removed that had substantial similarity to the PhiX genome, which is used as an internal positive control for the instrument, as well as oligonucleotides used in the library preparation. A total of 254,074,113 reads were retained. All reads were then processed with a custom perl script to excise, and potentially correct, the internal barcodes and associate sequences with tree sample identifiers. After removal of barcodes and invariant restriction site sequences, reads were 85–87 bp in length. DNA sequences were assembled to the *P. tremula* draft genome assembly v0001 (UPSC draft genome release, http://popgenie.org; http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/Pota/), using `bwa` ver. 0.7.5 algorithms `index`, `aln` and `samse` (Li & Durbin, 2009). To align sequences to the reference genome, we first seeded a 20 bp part of the sequence by tolerating 2 mismatches, and then aligned the total sequence allowing not more than 4 mismatches. Sequences were trimmed to 35 bp if the quality score was below 10. Only sequences with a unique best hit to the reference were aligned. Single nucleotide variants were called with `samtools mpileup` and `bcftools view` algorithms vers. 0.1.19 from the aligned sequences. This generated a `vcf` output file containing likelihood estimates for bi-allelic genotypes at variant nucleotide positions. We called single nucleotide polymorphisms (SNPs) at sites that were variant with an at least 99% probability, where at least 95% of our samples were covered (had at least one read), and by using a full prior for `bcftools view` with the scaled mutation substitution rate set to 0.001. Variant calling was done for either the full data set including all individuals, or a reduced set excluding progeny samples to omit rare variants potentially present only in one family. We obtained 32,912 and 64,649 SNPs for the full and reduced data sets (the same stringent criteria applied to a smaller number of individuals allowed more SNPs to be identified). For both sets, we excluded variants with minor allele frequencies ≤ 0.05, which were less likely to be informative about ancestry between the highly genetically divergent parental species (*P. alba* and *P. tremula*). We retained 11,976 SNPs that were present in both sets for the final analyses. To test for the potential dependence among linked SNPs within the same sequence read (85–87 bp in length), we randomly selected only one variant from sites that were less than 88 bp apart from each other, and retained 5,226 variants used for additional analyses (Supporting information). Average sequence depth was > 8 per SNP and individual for both data sets (Table S1 and S2).

Estimation of genetic ancestry

To learn more about the timing of reproductive isolation between *P. alba* and *P. tremula*, our main goal was to estimate the ancestries of parents that mated naturally in the hybrid zone. In particular, we wanted to (i) compare the species ancestry of the father to the ancestry of the mother (i.e. is mating random or non-random with respect to the ancestry of the parents), and (ii) compare the distribution of the genome-wide admixture proportions in the seedlings class to that of the established adult trees. By knowing the genetic ancestry of the mother and her progeny, it is possible to infer the genetic ancestry of father's gamete without the need to sample putative fathers in the hybrid zone directly. We therefore built a model that estimates the genetic ancestry of mothers, their progeny, population reference samples, and the genetic ancestry of the gametes of the unknown fathers. One or more fathers are possible within an open pollinated family. As our focus was to estimate the prevalence of homospecific versus hybrid matings, not the number of contributing fathers, we estimate genetic ancestry of the pollen donor independently for each progeny.

We used a Bayesian clustering method similar to that implemented in the software `structure` (Pritchard *et al.*, 2000; Falush *et al.*, 2003). The main difference is that we incorporated genotype uncertainty arising from DNA sequence data with limited coverage (per locus and individual), and that we estimated ancestry at both allele copies at a locus jointly, enabling better distinction between different hybrid classes. A closely related, but independent, model and its application in an empirical study will be described elsewhere (Gompert *et al.* in review; Buerkle *et al.* in prep.). In this study, we modified the model to include the use of family data to estimate the genetic ancestry of unsampled fathers. We briefly describe the model here and refer to the Supporting information for entire details. Figure 2 gives a graphical representation of the model.

As in the 'admixture model' implemented in the software `structure` (Pritchard *et al.*, 2000), our primary goal was to use a hierarchical Bayesian model to estimate the proportion q_k of an individual's genome that has ancestry in population k out of K source populations. For both models, the probability of observing the genotype $\mathbf{g}$ is conditional on the unknown population of origin $\mathbf{z}$ of the alleles that form the genotype, and the unknown allele frequencies $\mathbf{p}$ in the source populations, $P(\mathbf{g}|\mathbf{z}, \mathbf{p})$. Relative to the `structure` model, our modifications include the use of genotype likelihoods associated with next generation DNA sequence data, considering both allele copies in a diploid genotype simultaneously (as noted above) and different parameterization at a few levels in the hierarchical model (including incorporation of family data).

Because contemporary DNA sequencers result in stochastic and finite sampling of DNA sequences from a locus and individual, diploid genotypes can only be inferred with some uncertainty. Rather than requiring high sequence coverage and using arbitrary thresholds to "call genotypes" from genotype likelihoods, it is possible to utilize the genotype likelihoods directly in models and thereby utilize a greater fraction of the sequence data by explicitly taking genotype uncertainty into account (e.g., Gompert *et al.*, 2010; Li, 2011; Buerkle & Gompert, 2013). In addition, methods using genotype likelihoods rather than potentially erroneous genotype calls can give more accurate ancestry estimates (Gompert & Buerkle, 2013; Skotte *et al.*, 2013). The genotype likelihoods, $L(\mathbf{g}|\mathbf{x})$, are pre-calculated from the sequence data $\mathbf{x}$ using `bcftools`, where the number of reads and allelic state, and the read specific error rate $\boldsymbol{\epsilon}$ (computed from the base quality scores), are taken into account (Li, 2011; Gompert & Buerkle, 2013). The pre-calculated genotype likelihoods were used directly in the model: $L(\mathbf{g}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{g})$. Genotypes were updated based on genotype likelihoods, estimated population allele frequencies and locus-specific ancestry, enabling the handling of residual genotype uncertainty resulting from missing data or sequencing errors in the genotype likelihood data (Equation 7, Supporting information). We restricted our model to bi-allelic loci and diploid individuals, allowing four different genotypic states $g_{ij} \in \{00, 01, 10, 11\}$ for each individual j and locus i (0 denotes the reference allele, and 1 the alternative allele), and assume independence among loci. We estimated the allele frequencies p_{ik} in the K populations under an F-model, in which p_{ik} are related to an allele frequency in the common ancestor of the populations (π_i) through a variance term F_k (analogous to Wright's F_{ST}; as in Falush *et al.*, 2003; Gompert *et al.*, 2012a; Buerkle & Gompert, 2013; Parchman *et al.*, 2013). The ancestral allele frequencies $\boldsymbol{\pi}$ were modeled to depend on parameter α that reflects the genetic diversity in the common ancestor.

For each individual, we estimated locus-specific ancestry $\mathbf{z}$ conditional on the genome-wide ancestry class matrix $\mathbf{Q}$, $P(\mathbf{z}|\mathbf{Q})$. $\mathbf{Q}$ gives the proportion of the genome that has intra-source population (on the diagonal) or inter-source population ancestry (off the diagonal; Fig. 3). We use the terms 'intra-source' and 'inter-source ancestry' to clearly distinguish between source population ancestry and genotypic state (i.e. homo- or heterozygosity), as these terms are not equivalent (see Gompert & Buerkle, 2013). The parameters in $\mathbf{Q}$ are related to the admixture proportion $\mathbf{q}$, but have the advantage that ancestry can be modeled as a diploid combination of ancestry in the genotype rather than considering ancestry for each allele copy separately. For example, if $K = 2$, $\mathbf{z}$ can have four states: both allele copies can originate from source population 1, both allele copies can come from source population 2, or the two allele copies can be derived from different source populations (these are two states with equal probability for our unphased data). Accordingly, $\mathbf{Q}$ can also have four different states, which describe the proportion of the individual's genome that has intra-source (Q_{11} and Q_{22}) or inter-source ancestry ($Q_{12}+Q_{21}$; we will refer to inter-source ancestry as Q_{12} for simplicity as for our unphased data, $\mathbf{Q}$ is symmetrical above and below the main diagonal; Fig. 3). From $\mathbf{Q}$, the marginal genome-wide admixture proportion $\mathbf{q}$ can easily be calculated. Estimating ancestry for both allele copies jointly has two important advantages. First, it better distinguishes between different hybrid classes, particularly between F_1 hybrids, recombinant hybrids, and backcrosses (Fig. 3; similar to the model of Anderson & Thompson, 2002). For example, with $K = 2$, the expected genome-wide admixture proportion q_1 is 0.5 for both F_1 and F_2 hybrids; however, their expected inter-source ancestry differs. For an F_1 hybrid, the whole genome will be heterospecific, thus inter-source ancestry will be 1. For an F_2 hybrid, inter-source ancestry will have an expected value of 0.5, with some variance around it depending on the genetic map size (affected by the number of chromosomes, genetic map size of chromosomes, and recombination rate). By using $\mathbf{Q}$ as a model parameter, we can learn directly about the proportion of an individual's genome that has inter-source or intra-source ancestry, and thus better estimate its hybrid class. Second, when estimating genetic ancestry of progeny, it is highly beneficial to compute it for the ancestry combination in the genotype at a locus, because the two allele copies are not independent of one another. Instead, if one has been inherited from the mother, the other must have been inherited from the father.

We used different priors on $\mathbf{Q}$ for adult trees (mothers, $\mathbf{Q}_m$; fathers, $\mathbf{Q}_f$; population reference samples, $\mathbf{Q}_r$) or progeny samples ($\mathbf{Q}_p$). For adults, we calculated $\mathbf{Q}$ only conditional on the parameter $\boldsymbol{\gamma}$ that describes admixture in the whole population, $P(\mathbf{Q}_{m,f,r}|\gamma)$. For progeny, we have additional information and the genetic ancestry of a progeny will depend on the genetic ancestry of its parents. We therefore modeled the admixture of the progeny conditional on a function of the admixture proportions of the parents, $\boldsymbol{\nu} = f(\mathbf{Q}_m, \mathbf{Q}_f)$, and variance β, $P(\mathbf{Q}_p|\boldsymbol{\nu}, \beta)$. The parameter β models the deviation from the expected ancestry $\boldsymbol{\nu}$ that results from non-independence among loci (i.e. limited meiotic recombination of ancestry blocks in early generation hybrids, where the variance in gamete ancestry is affected by genetic map size). In the absence of direct observations of data for fathers, the parameter $\mathbf{Q}_f$ is estimated by its effect on $\mathbf{Q}_p$, independently for each progeny.

The full model is given by:

$$\begin{gathered}P(\mathbf{g}, \mathbf{z}, \mathbf{p}, \boldsymbol{\pi}, \boldsymbol{F}, \alpha, \mathbf{Q}, \boldsymbol{\gamma}, \beta|\mathbf{x}) \propto \\ P(\mathbf{x}|\mathbf{g})P(\mathbf{g}|\mathbf{z}, \mathbf{p})P(\mathbf{p}|\boldsymbol{\pi}, \boldsymbol{F})P(\boldsymbol{\pi}|\alpha)P(\alpha)P(\boldsymbol{F}) \\ P(\mathbf{z}|\mathbf{Q})P(\mathbf{Q}_{r,m}|\boldsymbol{\gamma})P(\mathbf{Q}_p|\boldsymbol{\nu}, \beta)P(\mathbf{Q}_f|\boldsymbol{\gamma})P(\boldsymbol{\gamma})P(\beta). \qquad (1)\end{gathered}$$

We provide full details on the implementation of the MCMC algorithm and settings used for analysis in the Supporting information and software.

Simulations

We used simulated data sets to evaluate how well our model could recover the genetic ancestry of fathers, given the genetic data for reference samples, mothers, and progeny. Overall the simulations were meant to emulate the type of empirical data we have in this study. We simulated 15,000 loci distributed evenly across 20 chromosomes, with population allele frequencies $\mathbf{p}$ drawn independently for two populations from a beta distribution with parameters shape$_1 = \boldsymbol{\pi}(1/F - 1)$ and shape$_2 = (1 - \boldsymbol{\pi})(1/F - 1)$. The $(1/F - 1)$ term acts as a precision parameter for the distribution, and F is a variance relative to the common ancestral allele frequency $\boldsymbol{\pi}$ and is analogous to Wright's F_{ST} (as in e.g., Gompert *et al.*, 2012a; Buerkle & Gompert, 2013). The ancestral allele frequency $\boldsymbol{\pi}$ was drawn from a symmetrical beta distribution with shape α. We chose different α depending on F to obtain U-shaped population allele frequency distributions similar to those of the empirical data. We generated genotypic data for 5,000 randomly selected loci for species and hybrid individuals (F$_1$, F$_2$, F$_3$ and first generation backcrosses), including 17 mothers, 30 fathers, and 510 progeny (also including later generation crosses). Full details are in the Supporting information. We explored nine different simulation settings that are likely to influence the information content of the data sets, namely the combinations of $F = \{0.1, 0.5, 0.8\}$, and mean sequence coverage of $\{1, 3, 8\}$ drawn from a negative binomial distribution with size 5. Details on the settings are provided in Table S3. Simulations were done in R (R Development Core Team, 2012), and subsequently analyzed in the same way as the empirical data.

Results

Seed sampling and germination rate

Based on classification by genetic ancestry, the sampled mother trees consisted of four pure *P. tremula* ($q \leq 0.1$), three pure *P. alba* ($q \geq 0.9$; including two trees sampled in 2010: F010 and F011), and 10 trees with admixed ancestry ($0.1 < q < 0.9$; Table 1). All trees sampled in 2011 showed a high synchrony in fruit burst allowing sampling of all seeds within a single week. There was a tendency of fruit burst to start with *P. tremula*, followed by the hybrids, and *P. alba* fruits bursting at the end of the collection period (Fig. S1). Seed output was substantially different between species, with *P. tremula* producing huge quantities of seeds, in contrast to *P. alba* and hybrid mothers that often set seedless fruits (observation not quantified). Of the trees sampled in 2011, 91–163 seeds per family were sown to obtain tissue samples and to determine germination and survival rates (Table 1). Germination and survival rates after seven and 14 days were high (89% for both dates; sd <10%), and showed no difference between *P. tremula* and admixed mother trees (t-test, p-value = 0.30 and 0.71 after seven and 14 days, respectively; Table 1, Fig. 4; data for only one family was available for *P. alba*). Unfortunately, many of the seedlings died shortly after two weeks, probably because of an infestation with root maggots.

Estimation of genetic ancestry

Simulations

To assess the performance of our model for estimating genetic ancestry for population and family samples, as well as ancestry of fathers, we generated simulated data sets with different amounts of ancestry and genotype information (Table S3). Overall, we found a high correlation between true and estimated admixture proportion q_1 and inter-source ancestry Q_{12} (Fig. S2, S3, S4). The accuracy in ancestry estimates was considerably higher for data sets with higher population divergence and thus information content in terms of ancestry ($F \geq 0.5$). For small F, ancestry estimates were more centered toward intermediate values (i.e. few individuals reached the extremes of the possible distribution; compare Fig. 5 and S5). This characteristic (which is consistent with the usage of uninformative priors) constitutes an opportunity to check for the precision of ancestry estimates in our empirical data, as the expected ancestries for our population reference samples are known. Except for low F, estimates of the genetic ancestry of mothers and progeny were more accurate than those of population reference samples, likely due to the mutually increased information in ancestry for family samples. By contrast, multiple levels of uncertainty in ancestry within family samples and unsampled fathers may have led to less precise estimates in some cases where F was low. Simulations of different levels of genotype uncertainty (i.e. differences in sequence depth) had minor effects on ancestry estimates for the whole genome of individuals (Fig. S2, S3, S4; and compare Fig. 5 and S6).

Genetic ancestry of population reference samples, mothers and progeny

For all maternal trees, their progeny, and population reference samples, we estimated the genome-wide inter-source ancestry (Q_{12}) and admixture proportion (q) for $K = 2$ (Data S1). Our estimates of the genetic differentiation between the parental populations and the hypothetical common ancestor were high ($F_{P.\,alba} = 0.70$ and $F_{P.\,tremula} = 0.76$). In addition, *P. alba* and *P. tremula* reference samples matched their expected ancestries (q approaches 0 or 1, Q_{12} approaches 0; Fig. 6). We therefore assume reliable estimates of ancestry for our other samples too. As the difference between the full and the reduced set of SNPs on ancestry estimates was minor (11,976 or 5,226 SNPs; Fig. S7; Data S1), we only present results based on the full data set. We excluded one of the progeny samples (F011_26) due to difficulty in obtaining reliable parameter estimates from the model.

Based on their genome-wide admixture proportions, maternal trees were either classified as pure species or intermediate hybrids, with two exceptions (Table 1). One *P. alba* tree appears to be slightly introgressed (F039; $q = 0.94$, 95% equal tail credible interval (CI) 0.937–0.947), and one admixed mother tree shows a backcross-like genotype toward *P. tremula* (I_345; $q = 0.22$, 95% CI 0.210–0.226). Three pairs of maternal trees are likely to be ramets of the same clone based on their very high genetic similarity (F008-F009, F036-I_396, and F021-F031; Table S4). Progeny samples spanned the whole range of possible admixture proportions from 0 to 1 (Fig. 6a, d). Hybrid population reference samples that were preselected based on data from 77 microsatellite markers to include only intermediate individuals (mean $q = 0.51$, range 0.39–0.62 based on `structure` analysis; Lindtke *et al.*, 2012) showed similar admixture proportions based on our sequence data and in our model (mean $q = 0.50$, range 0.38–0.60). Inter-source ancestries reached maximum possible values given q, indicating that most adult intermediate hybrids are likely F_1 hybrids rather than recombinants. By contrast, genome-wide inter-source ancestry was lower in most of the progeny compared to the adult samples (Fig. 6a), consistent with the presence of recombinant hybrids in the seedlings. The differences in the distributions of inter-source ancestry and genome-wide admixture proportions between hybrid adults and hybrid progeny were highly significant (Q_{12}, Kolmogorov-Smirnov test for individuals with $Q_{12} > 0.1$, adults vs. progeny, D = 0.89, p-value < 2.2e-16; q, Kolmogorov-Smirnov test for individuals with $0.1 < q < 0.9$, adults vs. progeny, D = 0.59, p-value = 1.423e-13). These significant differences between adults and progeny still hold when comparing the distributions of all adult and progeny samples, or by considering progeny of pure species mothers only (reflecting the extreme hypothetical case that hybrid mothers produce negligible seed quantities compared to species mothers; Table S5, Fig. S8). As ancestry estimates could potentially be affected by using different priors for adults and progeny (for the latter, the prior was computed as a function of the ancestry of the parents), we additionally ran our model without providing family information. We obtained very similar results, indicating that different prior constructions had little effect on our findings (see Supporting information for further details).

Genetic ancestry of fathers

We estimated genome-wide admixture proportions of the unknown fathers based on genetic data of progeny and their mothers (Data S1). As for the progeny samples, admixture proportions of fathers spanned the whole range of possible values from 0 to 1, although paternity from pure species and intermediate hybrids was clearly predominant (Fig. 6e; for CIs, see Fig. S9). Most pollen (49.3%) originated from *P. alba* ($q \geq 0.9$), 29.2% from admixed trees ($0.1 < q < 0.9$), and 21.5% from *P. tremula* ($q \leq 0.1$). Note that these estimates involve some uncertainty for the admixed fathers, particularly in cases in which one or both of the parents were likely early-generation hybrids (F_1 or backcrosses). In that case, the stochasticity of recombination and segregation of the 19 *Populus* chromosomes can lead to overly confident estimates of ancestry, as linkage within jointly inherited chromosome blocks would violate the assumed non-independence among loci.

Frequency of homospecific and hybrid matings

There was considerable variation in admixture proportions of fathers among and within families (Table 1, Fig. 7). The admixture proportions of fathers were largely dependent on the ancestry of the corresponding mother. When pooled according to the mother's ancestry, admixture proportions of fathers showed significantly different distributions (*P. alba* vs. *P.*

tremula, D = 0.74; *P. alba* vs. intermediate, D = 0.43; *P. tremula* vs. intermediate, D = 0.65; Kolmogorov-Smirnov test, all p-values < 0.00001; with mothers classified as *P. tremula* if $q \leq 0.1$, as *P. alba* if $q \geq 0.9$, and as intermediate if $0.25 < q < 0.75$). We also detected variation in admixture proportions of fathers among families of the same hybrid category of mothers (i.e. not all families shared the same mating pattern). Particularly admixed mothers showed high variation in admixture proportions of inferred pollen parents both among and within families (Table 1, Fig. 7).

Mothers of each species mated more often with conspecific males (66.3%) than with hybrid males (22.9%) or heterospecifics (10.7%; Table 1). The only maternal tree that deviates from this pattern (I_373) was sampled from outside the core study stand (Fig. 1, 7). Hybrid mothers showed a high propensity to backcross toward *P. alba* (57.6%) or to mate with hybrid males (33.8%), whereas comparatively few seeds were fertilized by *P. tremula* pollen (8.6%; Table 1, Fig. 7). The backcross-like individual I_345 showed much more similarity in the distribution of fathers' admixture proportions to *P. tremula* than to the other hybrids (Table 1, Fig. 7). Further, four hybrid trees sampled in close proximity to *P. alba* also shared a similar mating pattern with *P. alba* (F008, F009, F020 and F033; Fig. 1, 7). The mating patterns of the hybrids resulted in a high proportion of *P. alba* backcrosses in the seedlings, which strongly contrasts to the distribution in adult trees where backcrosses are mostly absent (Fig. 6a, b, d, S8; Lexer *et al.*, 2010; Lindtke *et al.*, 2012). Mothers that are likely to be ramets of the same clone showed very similar mating patterns (F008-F009; F036-I_396; F021-F031; Table 1, Fig. 7).

Discussion

In this study, we have investigated mating patterns in a natural mosaic hybrid zone of European *Populus* to assess the importance of early-acting and late-acting reproductive barriers in species isolation. Although the importance of prezygotic barriers has been emphasized by recent theoretical and empirical studies (e.g., Ramsey *et al.*, 2003; Rieseberg & Willis, 2007; Lowry *et al.*, 2008), our results show that early-acting mechanisms are insufficient for the maintenance of RI in *Populus*. Rather, the prevalence of a full range of hybrids in progeny and the discrete distribution of ancestry in adult trees suggest an important role for traits that act later in the life cycle to contribute to postzygotic isolation and to maintain species integrity. In addition, we found viable hybrid progeny spanning the whole range of possible admixture proportions and fertile hybrids that contribute to seedling production. This suggests that some hybrids are possibly as fit as their parental species, which can lead to the breakdown of species identity through recombination and genetic exchange (Kirkpatrick & Ravigne, 2002; Ortiz-Barrientos *et al.*, 2002; Bank *et al.*, 2012). Despite this potential for interspecific gene flow, the pronounced genetic clustering of the parental species and their hybrids in adult trees indicates that species barriers are nevertheless strong (see Lexer *et al.*, 2010; Lindtke *et al.*, 2012). We propose that RI is potentially maintained by repeated episodes of selection between germination and sexual maturation, in which the majority of recombinant hybrids have low viability, effectively reducing genetic exchange between parental genomes.

Mating patterns and hybridization frequency

By estimating genetic ancestries of maternal trees, progeny, and pollen parents within 17 open pollinated families, we determined hybridization frequencies and mating patterns within a mosaic hybrid zone. We detected a tendency for homospecific matings despite the close spatial proximity of species and hybrids (Table 1, Fig. 1, 7), consistent with early-acting barriers to reproduction between *P. alba* and *P. tremula*. However, even though 66.3% of matings were homospecific, hybridization rates were high, resulting in the formation of a substantial number of F_1 hybrid (10.7%) and backcross progeny (22.9%; Table 1). In addition, we found evidence for fertility among admixed trees, which contributed as male and female parents in reproduction (Fig. 6c, e). Furthermore, germination and survival rates in hybrid families were high, and did not differ from the species samples (Table 1, Fig. 4). These findings indicate that while early-acting barriers are clearly important contributors to RI, they are also insufficient to explain the discrete ancestry clusters (i.e. species and intermediate hybrids) found among adult trees. Although we observed a very low seed output for hybrid and *P. alba* mothers compared to *P. tremula* mothers in 2011 (not quantified), mating pattens of *P. tremula* females alone resulted in high hybridization rates (Table 1).

The rarity of conspecific pollen is unlikely to be a major source for the observed hybridization rates, as pollen disperses on average over several hundreds of meters in this wind-pollinated tree genus (Bialozyt, 2012), and pollen from pure species and hybrids was clearly available in the hybrid zone (Fig. 6e). Nevertheless, our results are consistent with some spatial effect on hybridization frequency (Fig. 1, 7), which can result from the higher relative abundance of pollen from nearby males and thus a higher probability of pollination success (Lepais *et al.*, 2009; Niggemann *et al.*, 2012; Lagache *et al.*, 2013). The tendency for hybrid females to backcross more often toward *P. alba* than toward *P. tremula* (Table 1, Fig. 7) can be explained by an overall higher abundance of *P. alba* trees in flood-plain habitats (Lexer *et al.*, 2005), differences in flowering time, pollen competition, or intrinsic postzygotic factors (Lowry *et al.*, 2008).

Deviation in genetic ancestry between seedlings and mature trees

We found marked differences in the ancestry of seedlings and adult trees, for both the distributions of admixture proportions and inter-source ancestry (Fig. 6, Table S5, Fig. S8). The three ancestry categories in adults in the *Populus* hybrid zone (two species and intermediate hybrids) suggests that RI requires selection that acts later in the life cycle, after germination. The species and intermediate hybrid genotypes are likely to have a selective advantage over other hybrid classes during later life stages. In addition, our results indicate that survival is highest for those intermediate genotypes that also show maximal possible inter-source ancestries (i.e. F_1 hybrids; Fig. 6a). Although we cannot completely exclude the possibility that the rarity of recombinant hybrids in mature trees relative to their abundance in progeny results from a recent formation of the hybrid zone, this scenario is unlikely. First, the presence of some backcross-like and putatively introgressed mature individuals (I_345, F039; also see Lindtke *et al.*, 2012) suggests that an opportunity for a second generation of hybrids exists. Likewise, the flood plain area has been protected for > 30 years and includes some old trees of both species and hybrids. Second, the Italian peninsula is considered to have served as a refugium during the last glacial maximum for *P. alba* and *P. tremula*, rendering a long history of hybridization likely (Svenning *et al.*, 2008; Fussi *et al.*, 2010). Despite this opportunity for hybridization, species remain genetically well differentiated, indicating strong RI. Third, previous work has shown that very similar discrete ancestry clusters in adult trees exist in "replicate" *Populus* hybrid zones in Austria and Hungary (Lindtke *et al.*, 2012), which are unlikely to have all been formed recently.

The finding that adult intermediate hybrids are most likely to be F_1 hybrids contrasts to previous conclusions gathered on the same population reference samples but that were based on microsatellite markers (see Lexer *et al.*, 2010; Lindtke *et al.*, 2012). According to these previous studies, hybrids were inferred to be highly recombinant (Lexer *et al.*, 2010), with an unusual high level of inter-source ancestry that was nevertheless lower than that expected for F_1 hybrids (Lindtke *et al.*, 2012). These contrasting findings are potentially due to overconfidence in genotype associated to microsatellite markers that are nevertheless prone to allele dropout (Taberlet *et al.*, 1996), but likely also reflect the difference in the models used to estimate ancestry (Fig. 3). In particular, for our previous studies, we applied a model that estimates ancestry for each of the two allele copies independently, where the prior probability of ancestry for each allele copy equals the genome-wide admixture proportion q. With little information about locus-specific ancestry from the data (e.g., missing data, multiple segregating alleles, low and uncertain allele frequency differential between the parental species), locus-specific ancestry estimates will be strongly influenced by their prior probability, resulting in an expected inter-source ancestry of $2q(1-q)$. In that case, the function of expected locus-specific inter-source ancestry given q has its maximum at 0.5, and hence does not include the value of 1 expected for F_1 hybrids. By contrast, in our current model, we estimated ancestry for both allele copies jointly using the $K \times K$ matrices $\mathbf{Q}$ and $\mathbf{z}$ and included an explicit parameter for genome-wide inter-source ancestry (Q_{12}), which was allowed to range from 0 to 1. This approach is expected to lead to more precise and accurate results for inter-source ancestry, as loci with little information about ancestry will reflect prior information on inter-source ancestry from other loci, rather than a function of q (where even a low proportion of loci with little ancestry information may bias genome-wide inter-source ancestry toward $2q(1-q)$).

A current limitation of our model is the assumption of independence among loci. Nevertheless, tightly linked loci had a negligible effect on our ancestry estimates, as the difference in results between the full and reduced set of SNPs was minor (Fig. S7; Data S1). Although we cannot exclude over-representation of particular parts of the genome in our sample, our results should only be affected if these regions coincide with directional introgression in recombinant hybrids. Incorporating linkage and haplotype information in future models is likely to further improve inference of hybrid class and recombination history (e.g. see Wegmann *et al.*, 2011; Gompert & Buerkle, 2013). Regardless of whether adult hybrid individuals are truly F_1 or potentially more advanced-generation hybrids with exceptionally high inter-source ancestry, the deviations in admixture proportion and inter-source ancestry between seedlings and adults highlight the likely importance of post-germination reproductive barriers between the investigated species.

Evolutionary implications and further directions

Our study on *Populus* offers an example system in which species boundaries are maintained without substantial premating isolation in a secondary contact zone between highly divergent species. This indicates that the often generalized assumption regarding the importance of prezygotic barriers (e.g., Kirkpatrick & Ravigne, 2002; Ramsey *et al.*, 2003; Rieseberg &

Willis, 2007; Lowry *et al.*, 2008) potentially results from the limited diversity of species that has been studied so far, and from the difficulties in detecting postzygotic isolation mechanisms once prezygotic barriers are complete (Coyne & Orr, 2004). Our study provides further support for the importance of late-acting barriers even in highly divergent species as *Populus*. Although some of the hybrid individuals are fertile, only very few of their progeny reach maturity. This reduces the opportunity to pass recombinant chromosomes to the next generation, and thus limits the potential for interspecific gene flow. Interestingly, a number of other studies have similarly documented high hybridization rates (Field *et al.*, 2011; Lepais & Gerber, 2011; Moran *et al.*, 2012), high fitness of at least a subset of hybrid progeny (e.g., Cruzan & Arnold, 1994; Lepais & Gerber, 2011), or differences in the distribution of admixture proportion among life stages within hybrid zones (e.g., Cruzan & Arnold, 1994; Cornman *et al.*, 2004; Curtu *et al.*, 2009; Lepais & Gerber, 2011). This suggests that future studies on the evolution of RI that continue to investigate organisms with various life histories might reveal additional examples for a strong role of late-acting barriers for the maintenance of species.

One potential explanation for the reduced life-time fitness of recombinant hybrids is coadaptation within each of the parental species' genomes, where recombination can result in unbalanced genotypes (outbreeding depression; Lynch, 1991; Edmands & Timmerman, 2003). Accumulative minor disadvantages between seedling formation and maturity (resulting for example from repeatedly misexpressed genes in recombinant hybrids) might explain the rarity of adult recombinant hybrids despite high germination rates and fitness at the seedlings stage. Selective disadvantages of recombinant hybrids may be amplified by the life-history of *Populus* (i.e., several years to reach maturity, the large quantity of progeny but limited carrying capacity for adult trees). A long pre-reproductive period of selection and its potential importance for speciation has been demonstrated recently in a marine coral, characterized by a life-history very similar to long-lived trees (Prada & Hellberg, 2013). Further steps toward a better understanding of reproductive barriers in *Populus* would involve the identification of the set of genetic combinations or environmental factors that are responsible for the low fitness of a subset of recombinant hybrids during the transition from the seedling stage to maturity. Modeling the null expectation for hybridization frequencies given the spatial distribution and demography of trees, quantifying the number of seeds produced in each hybrid class (preferably over several years), and measuring different components to RI in experiments, can help to better determine the strength of different reproductive barriers in future studies.

Acknowledgments

We thank Stefano Castiglione, Alberto Spada, Berthold Heinze, Tom Parchman, Eligio Bossolini, Alan Brelsford, Pär Ingvarsson, Liz Mandeville and Monia Haselhorst for their help and advice. We thank Associate Editor Nick Barton and three anonymous reviewers who helped to improve an earlier version of the manuscript. This work was facilitated by SNSF mobility grant PBFRP3_145869 to DL and an SNSF project grant to CL.

References

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

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Arnold M L, Hodges SA (1995) Are natural hybrids fit or unfit relative to their parents? *Trends in Ecology and Evolution*, **10**, 67–71.

Bank C, Buerger R, Hermisson J (2012) The limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-island model. *Genetics*, **191**, 845–U345.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Bialozyt R (2012) Gene flow in poplar - experiments, analysis and modeling to prevent transgene outcrossing. *Iforest-Biogeosciences and Forestry*, **5**, 147–152, 1st Biosafety Workshop of COST Action FP0905, Hamburg, Germany, Sep 09, 2010.

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

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Buerkle CA, Lexer C (2008) Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, **23**, 686–694.

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

Butlin RK, Galindo J, Grahame JW (2008) Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2997–3007.

Cornman R, Burke J, Wesselingh R, Arnold M (2004) Contrasting genetic structure of adults and progeny in a Louisiana iris hybrid population. *Evolution*, **58**, 2669–2681.

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

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

Cruzan M, Arnold M (1994) Assortative mating and natural selection in an *Iris* hybrid zone. *Evolution*, **48**, 1946–1958.

Curtu AL, Gailing O, Finkeldey R (2009) Patterns of contemporary hybridization inferred from paternity analysis in a four-oak-species forest. *BMC Evolutionary Biology*, **9**.

Dickmann DI, Kuzovkina J (2008) *FAO. Poplars and Willows in the World: Meeting the needs of society and the environment*, chap. Poplars and willows in the world. International Poplar Commission 9-2, FAO, Rome, Italy.

Edmands S, Timmerman C (2003) Modeling factors affecting the severity of outbreeding depression. *Conservation Biology*, **17**, 883–892.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Field DL, Ayre DJ, Whelan RJ, Young AG (2011) The importance of pre-mating barriers and the local demographic context for contemporary mating patterns in hybrid zones of *Eucalyptus aggregata* and *Eucalyptus rubida. Molecular Ecology*, **20**, 2367–2379.

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.

Gompert Z, Buerkle CA (2013) Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology*, **22**, 5278–5294.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson R, 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.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012a) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Gompert Z, Parchman TL, Buerkle CA (2012b) Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 439–450.

Harrison RG, Rand DM (1989) Mosaic hybrid zones and the nature of species boundaries. In: *Speciation and its consequences* (eds. Otte D, Endler J), pp. 110–133, Sinauer Associates.

Jiggins CD, Salazar C, Linares M, Mavarez J (2008) Hybrid trait speciation and *Heliconius* butterflies. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3047–3054.

Kirkpatrick M, Ravigne V (2002) Speciation by natural and sexual selection: Models and experiments. *American Naturalist*, **159**, S22–S35.

Lagache L, Klein EK, Guichoux E, Petit RJ (2013) Fine-scale environmental control of hybridization in oaks. *Molecular Ecology*, **22**, 423–436.

Lauber K, Wagner G, Gygax A (2012) *Flora Helvetica*. Haupt Verlag, Bern.

Lepais O, Gerber S (2011) Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution*, **65**, 156–170.

Lepais O, Petit R, Guichoux E, *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology*, **18**, 2228–2242.

Lexer C, Fay MF, Joseph JA, Nica MS, Heinze B (2005) Barrier to gene flow between two ecological 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 J, van Loo M, *et al.* (2009) The use of digital image-based morphometrics to study the phenotypic mosaic in taxa with porous genomes. *Taxon*, **58**, 349–364.

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.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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.

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

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3009–3021.

Lynch M (1991) The genetic interpretation of inbreeding depression an doutbreeding depression. *Evolution*, **45**, 622–629.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, **20**, 229–237.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in Heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28.

Moran EV, Willis J, Clark JS (2012) Genetic evidence for hybridization in red oaks (*Quercus* sect. *Lobatae, Fagaceae*). *American Journal of Botany*, **99**, 92–100.

Niggemann M, Wiegand T, Robledo-Arnuncio JJ, Bialozyt R (2012) Marked point pattern analysis on genetic paternity data for uncertainty assessment of pollen dispersal kernels. *Journal of Ecology*, **100**, 264–276.

Nosil P, Vines TH, Funk DJ (2005) Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution*, **59**, 705–719.

Ortiz-Barrientos D, Reiland J, Hey J, Noor M (2002) Recombination and the divergence of hybridizing species. *Genetica*, **116**, 167–178.

Parchman TL, Gompert Z, Braun MJ, *et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**, 3304–3317.

Parchman TL, Gompert Z, Mudge J, Schilkey F, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Payseur BA (2010) Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources*, **10**, 806–820.

Prada C, Hellberg ME (2013) Long prereproductive selection and divergence by depth in a Caribbean candelabrum coral. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 3961–3966.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Wegmann D, Kessner DE, Veeramah KR, *et al.* (2011) Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, **43**, 847–853.

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

Data Accessibility

Genotype likelihoods (in vcf file format) for the full and reduced data sets, formatted input files and source code for the genetic ancestry model and simulation code have been deposited at Dryad (doi:10.5061/dryad.kh7sc). Sequence data have been deposited at NCBI SRA (SRP040523).

Author Contributions

DL and CL designed the study and collected samples. DL generated the DNA sequence data and simulations. ZG, CAB and DL worked on data processing and new analytical tools. DL analyzed the data. DL and CAB wrote the paper. All authors revised the paper.

Tables and Figures

Table 1: Summary of homospecific and hybrid matings for 17 open pollinated families in this study, based on 11,976 SNPs. Fam, family ID; $q(m)$, admixture proportion of mother (from 0, *P. tremula*, to 1, *P. alba*); $q(pro)$, mean (sd) admixture proportion of progeny within family (note that these values need to be interpreted with caution for multimodal and bounded distributions); N pro, number of progeny used for paternity inference; trem, hyb or alba give number of pollen parents categorized into three hybrid classes, with $q \leq 0.1$ assigned as *P. tremula*; $0.1 < q < 0.9$ assigned as hybrid, and $q \geq 0.9$ assigned as *P. alba*; homo (homospecific), F_1-like, BCtrem, BCalba or F_n (hybrid × hybrid) show proportion of crossing types in the progeny given the hybrid class assignments of their parents; N seeds, number of seeds sown; GR, germination and survival rate (not available for families F010 and F011). The last three rows give summary on pollen source, and mating patterns of species and hybrid mothers.

				Pollen parents			Crossing types							
Fam	$q(m)$	$q(pro)$	N pro	trem	hyb	alba	homo	F_1-like	BCtrem	BCalba	F_n	N seeds	GR week1	GR week2
F021	0.00	0.04 (0.11)	30	25	4	1	0.83	0.03	0.13	-	-	155	0.94	0.91
F030	0.00	0.06 (0.15)	30	25	2	3	0.83	0.10	0.07	-	-	155	0.86	0.94
F031	0.00	0.06 (0.16)	30	25	2	3	0.83	0.10	0.07	-	-	163	0.91	0.91
I_373	0.00	0.32 (0.17)	30	3	14	13	0.10	0.43	0.47	-	-	163	0.84	0.87
I_345	0.22	0.18 (0.12)	30	19	10	1	-	-	0.63	0.03	0.33	155	0.94	0.95
F008	0.51	0.75 (0.04)	30	0	0	30	-	-	0.00	1.00	0.00	155	0.94	0.88
F009	0.50	0.72 (0.11)	30	1	1	28	-	-	0.03	0.93	0.03	155	0.92	0.97
F020	0.50	0.68 (0.14)	8	0	1	7	-	-	0.00	0.88	0.12	91	0.92	0.92
F022	0.50	0.63 (0.12)	30	0	15	15	-	-	0.00	0.50	0.50	155	0.92	0.90
F026	0.50	0.57 (0.13)	30	0	20	10	-	-	0.00	0.33	0.67	155	0.86	0.86
F032	0.50	0.66 (0.14)	30	1	8	21	-	-	0.03	0.70	0.27	101	0.95	0.94
F033	0.50	0.71 (0.13)	30	2	0	28	-	-	0.07	0.93	0.00	155	0.81	0.81
F036	0.50	0.58 (0.16)	30	0	18	12	-	-	0.00	0.40	0.60	155	0.97	0.95
I_396	0.50	0.52 (0.15)	30	1	21	8	-	-	0.03	0.27	0.70	155	0.97	0.95
F039	0.94	0.89 (0.10)	34	0	14	20	0.59	0.00	-	0.41	-	155	0.65	0.57
F010	1.00	0.97 (0.07)	19	0	3	16	0.84	0.00	-	0.16	-	NA	NA	NA
F011	1.00	0.91 (0.15)	32	2	8	22	0.69	0.06	-	0.25	-	NA	NA	NA
Overall				21.5%	29.2%	49.3%								
Species							66.3%	10.7%	10.7%	12.2%				
Hybrid									8.6%	57.6%	33.8%			

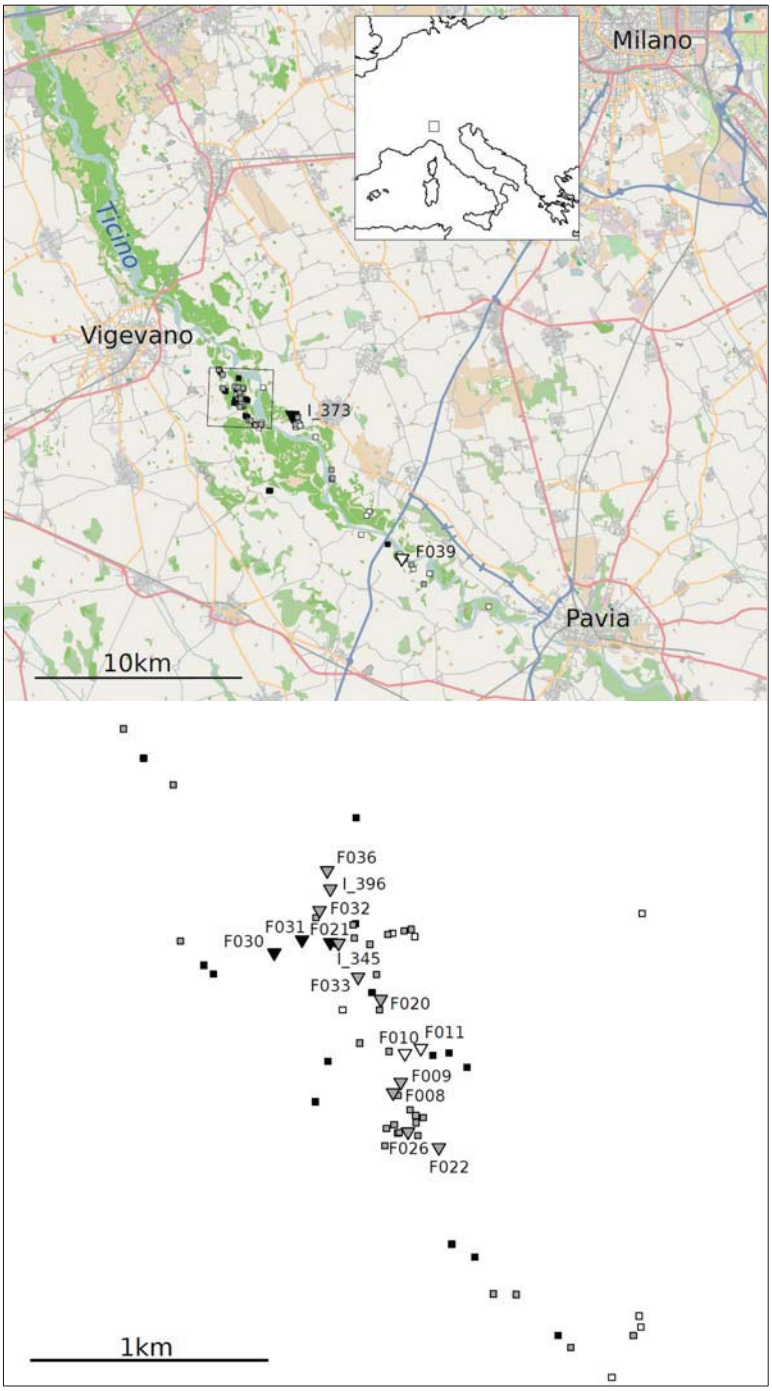

Figure 1: Map of sampling localities. Top, locality of the *Populus* hybrid zone in northern Italy; bottom, enlarged map of the core study stand. Triangles indicate maternal trees, squares population reference samples. 20 *P. tremula* reference samples collected approximately 50 km north of the hybrid zone are not shown. Black, *P. tremula*; gray, admixed individuals, white, *P. alba*.

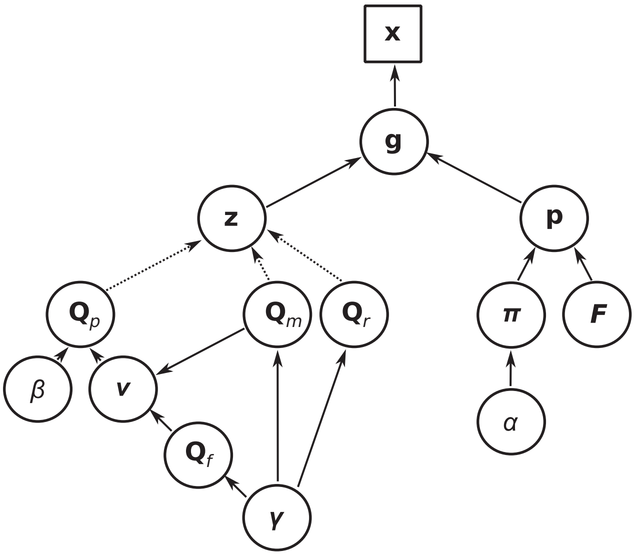

Figure 2: Model graph. Dotted lines denote alternative paths to model $\mathbf{Q}$ dependent on the family class of the sample (r, population reference; m, mother; p, progeny; f, father). See Methods and Supporting text for further details.

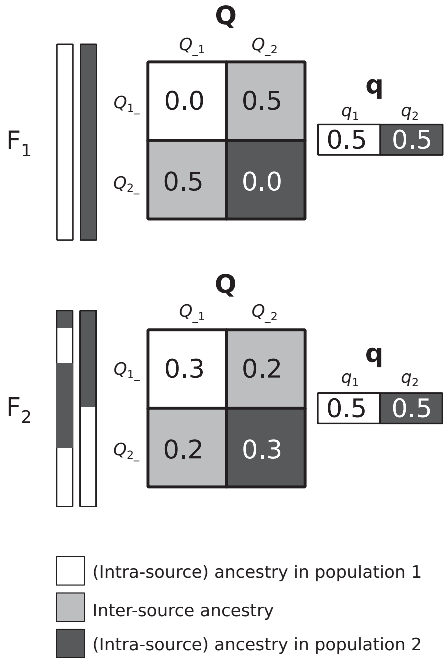

Figure 3: Genome-wide ancestry estimated by admixture class matrix **Q** or admixture proportion **q** for F$_1$ and F$_2$ (or other recombinant) hybrid genomes. **Q** gives diploid ancestry combinations (intra-source ancestry: proportion of the diploid complement in the genome with ancestry in the same source population, Q_{11} or Q_{22}; inter-source ancestry: proportion of the diploid complement in the genome with ancestry in different source populations, $Q_{12} + Q_{21}$). The admixture proportion **q** gives the overall proportion of the genome with ancestry in population 1 or 2 (q_1 and q_2), independent on their diploid combination. Genome-wide ancestry between F$_1$ and F$_2$ hybrids differs for **Q**, but is identical for **q**. See text for further details.

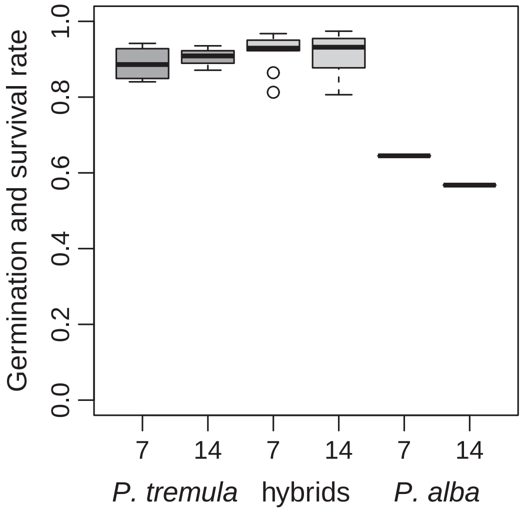

Figure 4: Germination and survival rates after seven and 14 days within open pollinated families. Data were pooled according to the species assignment of the mother.

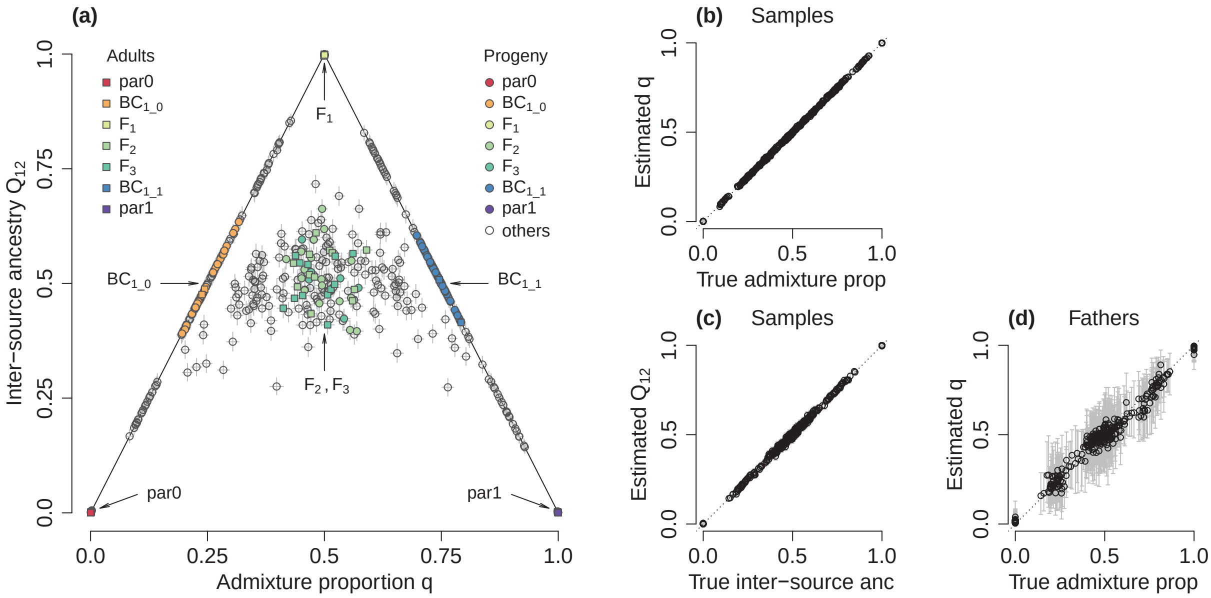

Figure 5: Ancestry estimates for simulated data. The performance of the model and software is shown for one of the simulations, with $F = 0.8$ and coverage $= 8$. (a) inter-source ancestry (Q_{12}) as a function of admixture proportion (q) for all samples excluding fathers; lines indicate maximum possible Q_{12} given q; species samples (par0 and par1), hybrids (F_1, F_2, F_3), and first generation backcrosses toward par0 or par1 (BC_{1_0} or BC_{1_1}) are labeled; (b) comparison of true vs. estimated admixture proportion for all samples excluding fathers; (c) comparison of true vs. estimated inter-source ancestry for all samples excluding fathers; (d) comparison of true vs. estimated admixture proportion for the gametes from fathers. Gray lines show 95% equal tail credible intervals.

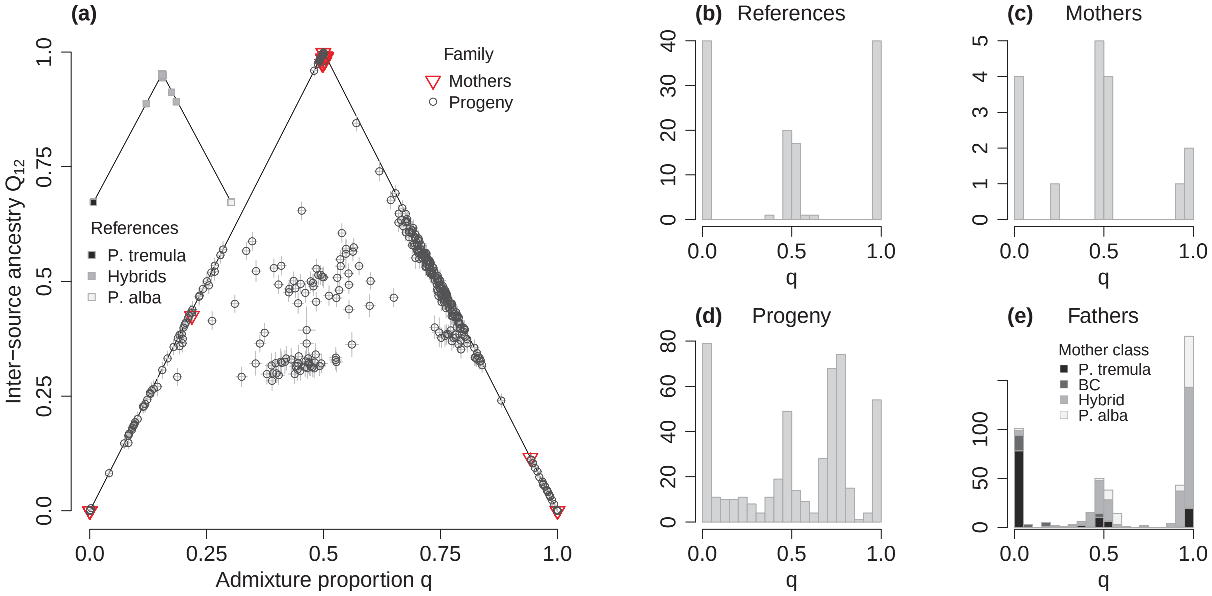

Figure 6: Ancestry estimates for empirical data, based on 11,976 SNPs. (a) Inter-source ancestry (Q_{12}) is plotted as a function of admixture proportion (q) for all samples excluding fathers; lines indicate maximum possible Q_{12} given q; thin lines show 95% equal tail credible intervals; circles, progeny; triangles, mothers; population reference samples (squares) are plotted separately as an insert at the left; (b)–(e) Histograms of admixture proportions (q) for (b) population reference samples, (c) mothers, (d) progeny, and (e) gametes of fathers; for the latter, shades of gray indicate the hybrid class of the corresponding mother of the mating pair.

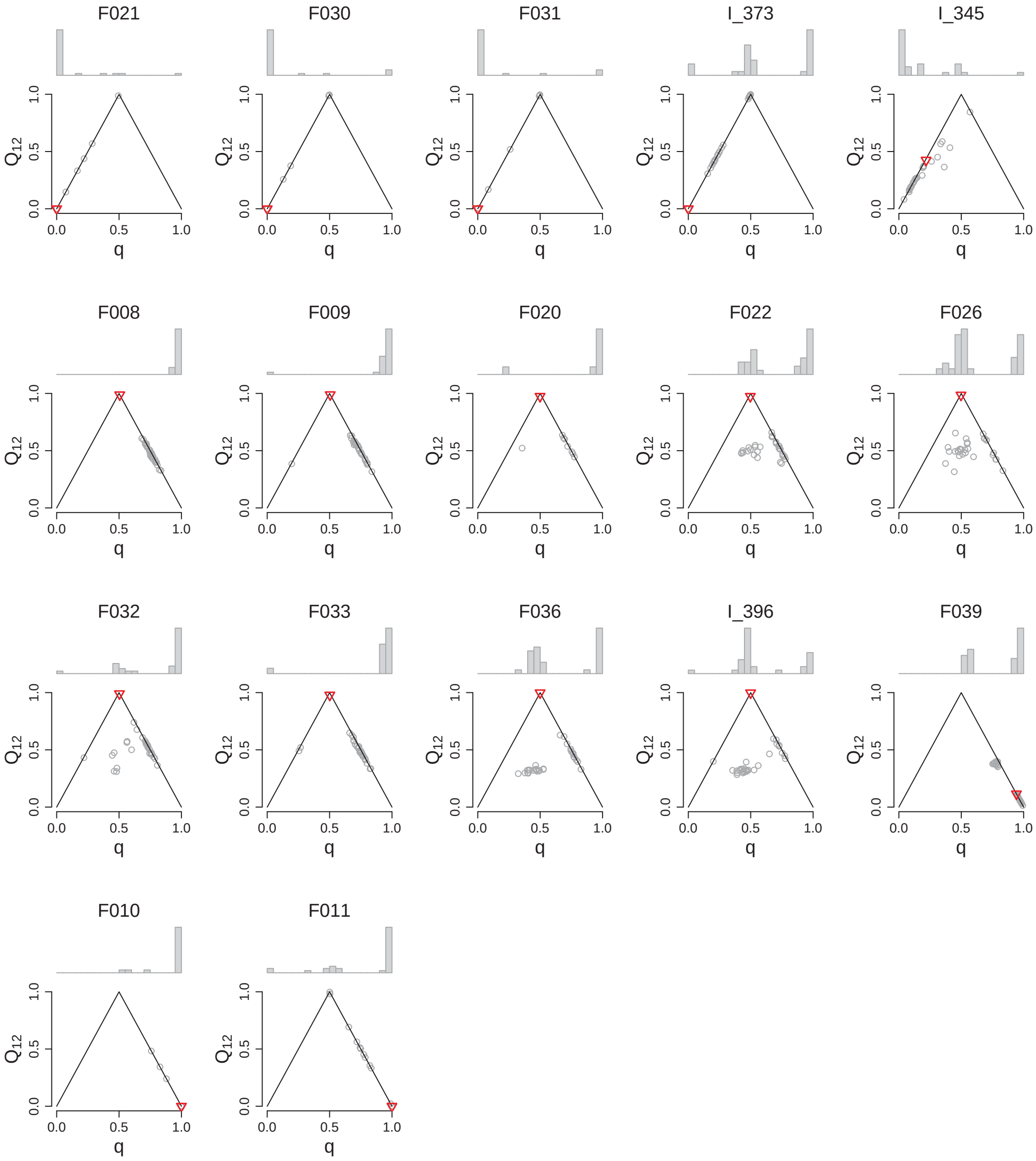

Figure 7: Mating patterns and ancestry estimates per family for empirical data, based on 11,976 SNPs. Histograms show admixture proportions for fathers within an open pollinated family (family ID at top of each plot); below each histogram, inter-source ancestry (Q_{12}) is plotted as a function of admixture proportion (q) for the mother (triangle) and her progeny (circles) of that family.