Published in "Molecular Ecology 22(3): 842–855 , 2013"
which should be cited to refer to this work.

Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species

KAI N. STÖLTING,* RICK NIPPER,† DOROTHEA LINDTKE,* CELINE CASEYS,* STEPHAN WAEBER,* STEFANO CASTIGLIONE‡ and CHRISTIAN LEXER*

*Department of Biology, Unit of Ecology & Evolution, University of Fribourg, Chemin du Musée 10, CH-1700 Fribourg, Switzerland, †Floragenex, 2828 SW Corbett Ave, Suite 145, Portland, OR, 97201, USA, ‡Department of Chemistry, University of Salerno, 84084 Fisciano, Italy

Abstract

Recent advances in population genomics have triggered great interest in the genomic landscape of divergence in taxa with 'porous' species boundaries. One important obstable of previous studies of this topic was the low genomic coverage achieved. This issue can now be overcome by the use of 'next generation' or short-read DNA-sequencing approaches capable of assaying many thousands of single nucleotide polymorphisms (SNPs) in divergent species. We have scanned the 'porous' genomes of *Populus alba* and *Populus tremula*, two ecologically divergent hybridizing forest trees, using >38 000 SNPs assayed by restriction site associated DNA (RAD) sequencing. Windowed analyses indicate great variation in genetic divergence (e.g. the proportion of fixed SNPs) between species, and these results are unlikely to be strongly biased by genomic features of the *Populus trichocarpa* reference genome used for SNP calling. Divergence estimates were significantly autocorrelated ($P < 0.01$; Moran's I up to 0.6) along 11 of 19 chromosomes. Many of these autocorrelations involved low divergence blocks, thus suggesting that allele sharing was caused by recurrent gene flow rather than shared ancestral polymorphism. A conspicuous low divergence block of three megabases was detected on chromosome XIX, recently put forward as an incipient sex chromosome in *Populus*, and was largely congruent with introgression of mapped microsatellites in two natural hybrid zones ($N > 400$). Our results help explain the origin of the 'genomic mosaic' seen in these taxa with 'porous' genomes and suggest rampant introgression or extensive among-species conservation of an incipient plant sex chromosome. RAD sequencing holds great promise for detecting patterns of divergence and gene flow in highly divergent hybridizing species.

Keywords: admixture, hybridization, population genomics, RAD sequencing, sex chromosomes, speciation genetics

Introduction

Understanding the genomic signatures associated with the origin and breakdown of species boundaries is of great current interest in molecular ecology and evolutionary biology (Wolf *et al.* 2010). Recent years have seen a wealth of population genomic studies on taxa with 'porous' genomes, that is, pairs or groups of diverging populations in which 'whole-genome isolation' has not yet been achieved (Dopman *et al.* 2005; Turner *et al.* 2005; Savolainen *et al.* 2006; Minder & Widmer 2008; Nosil *et al.* 2009; Michel *et al.* 2010; Andrew *et al.* 2012). This work has already started to change the way we think about the 'genomic landscape of divergence' in taxa with 'porous' species boundaries and the suite of population genetic forces shaping it.

Correspondence: Christian Lexer, Fax: +41 26 300 9698;
E-mail: christian.lexer@unifr.ch

Metaphors such as genomic 'islands' and 'continents' of speciation have been coined to articulate the complex sequence of events leading to speciation in the face of gene flow (i.e. in sympatry or parapatry; Turner *et al.* 2005; Via & West 2008). This process is sometimes thought to start from just a few 'speciation genes' subject to strong divergent selection (the genomic 'islands' of divergence), which increase in size due to either adaptive hitchhiking or divergence hitchhiking as newly arisen, adjacent mutations reduce gene flow further (Maynard Smith & Haigh 1974; Wu & Ting 2004; Via & West 2008). Nevertheless, the relative importance and timing of adaptive, divergence and 'genome hitchhiking' (i.e. few linked or many loci under divergent selection during speciation) is contentious, which reinforces the need to study taxa with 'porous genomes' at both early and advanced stages of speciation (Feder *et al.* 2012; Strasburg *et al.* 2012).

To date, most genomic scans are based on small numbers of molecular genetic markers typed in divergent population pairs or hybrid zones (typically dozens or hundreds of loci; Dopman *et al.* 2005; Lexer *et al.* 2010; Michel *et al.* 2010; review by Strasburg *et al.* 2012). Much larger genomic coverage would be desirable to provide a reliable picture of the genomic landscape of divergence (Feder & Nosil 2010), and the necessary technologies to achieve this goal have now become available.

So-called Next Generation Sequencing (NGS) approaches (from here onwards, short-read sequencing) have started to transform the way molecular ecologists approach key issues related to the origin and maintenance of biological diversity. While whole-genome scans of populations based on short-read sequencing have been carried out in few species (Rubin *et al.* 2010; Turner *et al.* 2010; Amaral *et al.* 2011), simpler and more affordable approaches involve partial genomic scans based on sequence templates of reduced complexity, such as Restriction site associated DNA (RAD) sequencing (Baird *et al.* 2008) or similar methods of genotyping-by-resequencing (Parchman *et al.* 2012). This approach has already provided intriguing insights into the genomic basis of parallel evolution (Hohenlohe *et al.* 2010) and into previously unresolved genetic structure in phylogeographic surveys (Emerson *et al.* 2010). Here, we apply RAD sequencing (from here onwards: RAD-Seq) to the 'porous' species boundary of *Populus alba* and *Populus tremula*, two hybridizing, ecologically divergent forest trees related to the completely sequenced 'model tree', *Populus trichocarpa* (Tuskan *et al.* 2006). Two issues are at the heart of our attention, namely the roles of shared ancestral polymorphism vs. interspecific gene flow as causes of allele sharing between divergent genomes and the role of sex chromosomes in reproductive isolation (RI).

With respect to the former issue, closely related species are expected to share alleles for some time due to incomplete lineage sorting or because of genomic admixture and gene flow across 'porous' species boundaries (Wu & Ting 2004), but distinguishing between these two scenarios is not a trivial task (Muir & Schlötterer 2005; Lexer *et al.* 2006; Palma-Silva *et al.* 2011). Knowing the genomic distribution of shared alleles would help, as gene flow is expected to result in entire blocks of material from one species embedded within another species' genomic background, with the size of shared blocks depending on the precise history of admixture (Buerkle & Lexer 2008; Winkler *et al.* 2010). High-density short-read sequencing–based genome scans should allow the detection of such linkage disequilibrium (LD) blocks shared among related taxa, for example, by examining genomic distributions and autocorrelations of polymorphisms that are variable vs. those that are fixed between species.

With regard to sex chromosomes, recent evidence indicates that 'speciation genes' tend to accumulate in these genomic regions due to suppressed recombination (Saether *et al.* 2007; Qvarnstrom & Bailey 2009), which should manifest itself in increased divergence in genomic scans. *Populus* species are dioecious (separate female and male trees), and sex appears to be controlled by an incipient sex chromosome (Yin *et al.* 2008) without full differentiation into heteromorphy (Macaya-Sanz *et al.* 2011) and with some variation in the size and precise location of the sex determination region among species (Pakull *et al.* 2009; Paolucci *et al.* 2010). Genetic mapping in a controlled backcross (BC_1) of *P. alba* and *P. tremula* revealed that donor alleles in the sex determination region introgressed into their recipient genome significantly better than expected under Mendel's laws (Macaya-Sanz *et al.* 2011). Hence, examining patterns of divergence and gene flow along the incipient *Populus* sex chromosome is of interest to speciation genetics, and short-read sequencing approaches such as RAD-seq provide an opportunity to do so.

Here, we make use of >38 000 reference-mapped, high-quality SNPs and a panel of 32 mapped microsatellites to address the following questions regarding genomic patterns of divergence and gene flow in *P. alba* and *P. tremula*:

1 What is the potential of RAD-Seq to screen divergent genomes for regions of great divergence, indicative of loci involved in RI, and regions of unusually low divergence?

2 Is allele sharing at low divergence loci more likely to be due to shared ancestral polymorphism or recurrent gene flow?

3 What is the extent of allele sharing on chromosome XIX, the incipient sex chromosome of *Populus*, and how can we explain it?

We underpin the results of our RAD-Seq-based genome scan by examining allele sharing and introgression of mapped microsatellites in a large number of individuals from two natural hybrid zones.

Methods

Sampling design

Populus alba and *Populus tremula* are widespread tree species with tractable genomes (~450–500 megabases; 2n = 38; Tuskan *et al.* 2006; *Populus trichocarpa* genome assembly version 2 on www.phytozome.org) and widespread synteny as revealed by comparative genetic mapping (Cervera *et al.* 2001; Macaya-Sanz *et al.* 2011). Sequence divergence between *P. trichocarpa* (the sequenced species) and *P. tremula* is low (between 1% and 5%), according to a recent re-sequencing study based on 77 gene regions (Ingvarsson 2008), which makes it possible to use the *P. trichocarpa* genome for reference mapping of Illumina sequence reads produced by RAD-Seq. The two ecologically divergent (flood plain vs. upland pioneer) target species of this study exhibit incomplete reproductive barriers, resulting in extensive advanced generation hybrid zones (Lexer *et al.* 2005, 2010). To carry out a RAD-Seq-based genomic scan for divergence and gene flow, we sampled seven pure genotypes of each species adjacent to a zone of sym- and parapatry in southern Switzerland and adjacent northern Italy (also known as Ticino river valley hybrid zone; below). The absence of recent admixture in the sampled trees was ascertained by microsatellites in Lexer *et al.* (2010) and Lindtke *et al.* (2012). As RAD-Seq generates codominant markers and no missing data were allowed in the final data set (below), this design allowed us to sample a total of 14 haploid genomes per species. This is sufficient to interrogate divergent genomes for the proportions of SNPs likely to be fixed between species (with the usual caveat that rare alleles segregating within one or both species will go undetected). For details on the samples subjected to RAD-Seq, see Table 1.

The sex of the sampled trees was unknown at the time point of RAD-Seq but was identified phenotypically at a later point. Among ten trees that could unequivocally be sexed, nine turned out to be male (Table 1). Note that males regularly outnumber female trees in natural populations of *Populus* spp. (Stettler *et al.* 1996). The only female sampled in our RAD genome scan showed Illumina read coverage statistics that were well within the range of males (below; Table S1, Supporting information). Thus, our sample is suitable for studying interspecific differentiation of the incipient poplar sex chromosome (below), but not for addressing sexual differentiation within species.

To validate the RAD-Seq results with a much larger panel of individuals, we sampled two large natural hybrid zones of *P. alba* and *P. tremula* for up to 32 mapped microsatellite loci (Table S2, Supporting information). Sampling for this step comprised 219 trees from the Ticino river valley hybrid zone in Italy and 186 trees from the Tisza valley hybrid zone in Hungary (localities described in Lexer *et al.* 2010; Table S3, Supporting information). This sampling included all 14 samples used in RAD-Seq. Our main goal was to compare genomic regions with strongly aberrant SNP patterns such as chromosome XIX, the incipient sex chromosome of poplar (Yin *et al.* 2008), to genomic regions that matched the genomewide average more closely, for example, chromosome VI. This chromosome was chosen for comparisons for consistency with previous studies which also used it as a reference (de Carvalho *et al.* 2010; Macaya-Sanz *et al.* 2011). The two hybrid zones exhibited a broad range of admixture proportions *Q* (Lexer *et al.* 2010), which permits the estimation of locus-specific ancestries (LSA's) compared to the genomic average (Gompert & Buerkle 2009).

RAD sequencing

Total genomic DNA was extracted from silica-dried leaf material using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's protocols. Subsequently, 2.5 μg of spectrophotometrically quantified DNA were submitted to FLORAGENEX, Oregon, who generated and sequenced RAD tags following the methods outlined by Baird *et al.* (2008), Hohenlohe *et al.* (2010) and Emerson *et al.* (2010). In brief, sequencing adaptors and individual barcodes were ligated to PstI-digested total genomic DNA, and the resulting fragments were sequenced from the restriction sites. Individually barcoded RAD samples were jointly sequenced on the Illumina GAIIx platform with single-end 80-bp chemistry. Reads were separated by individual, and sequencing barcodes were removed after the sequencing run, resulting in RAD tags of 70 bp. Summary statistics for the sequencing run can be found in Table S1 (Supporting information). RAD-Seq samples only a fraction of the genome, which reduces the complexity of sequence templates and thus maximizes coverage of individual loci for any given number of sequence reads obtained (Baird *et al.* 2008; Hohenlohe *et al.* 2010). In our case, *Populus trichocarpa* genome assembly version 2 contains

Table 1 Details for *Populus* samples submitted to RAD sequencing. For each sample (ID), species origin, Bayesian ancestry coefficients from STRUCTURE (Q), sex and geographic coordinates are indicated

ID	Species*	Q*	Sex†	Latitude	Longitude
I_309	*Populus alba*	0.989	Unknown	45°14′17.3″N	9°00′57.0″E
I_319	*P. alba*	0.987	Unknown	45°13′08.2″N	9°02′09.4″E
I_326	*P. alba*	0.987	Male	45°12′47.2″N	9°02′45.4″E
I_363	*P. alba*	0.986	Male	45°16′48.7″N	8°58′10.5″E
I_318	*P. alba*	0.984	Male	45°13′05.1″N	9°02′24.4″E
I_324	*P. alba*	0.980	Male	45°12′12.2″N	9°03′06.4″E
I_273	*P. alba*	0.988	Male	45°16′54.6″N	8°58′15.5″E
I_431	*P. tremula*	0.010	unknown	45°45′31.7″N	8°35′28.9″E
I_381	*P. tremula*	0.011	Male	45°45′39.7″N	8°35′24.6″E
I_435	*P. tremula*	0.012	Male	45°45′34.7″N	8°35′26.6″E
I_422	*P. tremula*	0.012	Unknown	45°17′06.4″N	8°56′16.1″E
I_426	*P. tremula*	0.011	Male	45°17′01.6″N	8°56′22.8″E
I_455	*P. tremula*	0.012	Male	45°49′04.8″N	8°28′18.9″E
I_360	*P. tremula*	0.011	Female	45°46′59.7″N	8°38′22.7″E

*Species origin was inferred from individual ancestry coefficients (Q) estimated by the STRUCTURE software based on 18 mapped microsatellite loci (Lexer *et al.* 2010), with pure *P. alba* defined as $Q \geq 0.98$ and pure *P. tremula* as $Q \leq 0.02$.
†Four trees could not be sexed, because they did not flower in that season.

at least 71 995 PstI-CTGCAG restriction sites (www.phytozome.org). Assuming that 2×70 bp could be sequenced from each restriction site, a portion of 10.79×10^6 bp or 2.5% of the *Populus* genome were screened for SNPs in this study. This is likely to be an overestimate of genomic coverage, as the use of a methylation-sensitive restriction enzyme such as PstI increases the fraction of coding DNA among the sequenced fragments, which is of particular interest in complex plant genomes such as *Populus* (Tuskan *et al.* 2006).

Microsatellite genotyping

A total of 32 mapped loci (Table S2, Supporting information) were picked from a larger panel of microsatellites (Lexer *et al.* 2010). These markers were chosen because they reside on well-covered chromosomes previously characterized by genetic mapping in *P. alba* and *P. tremula* (Macaya-Sanz *et al.* 2011; Table S2, Supporting information). All 32 loci were used to estimate genomic admixture gradients during statistical analysis (below), and markers located on chromosomes XIX (six loci) and VI (13 loci) were analysed for locus-specific introgression, to facilitate comparisons with previous studies that compared the same two chromosomes (de Carvalho *et al.* 2010; Macaya-Sanz *et al.* 2011). As sampling comprised >400 individuals from two replicate hybrid zones (above) and included those samples previously submitted to RAD-Seq, the microsatellite data provided a thorough test of the robustness of the RAD genotype data. Table S2 (Supporting information) lists details for all microsatellite loci typed in the two hybrid zones. Note that the chromosome XIX microsatellites were picked from scaffold 117 of *P. trichocarpa* genome assembly version 1 (Yin *et al.* 2008) and from *P. trichocarpa* linkage maps (Yin *et al.* 2004). Genetic mapping of these loci in *P. alba* and *P. tremula* indicates that all of them indeed map to the proximal part of chromosome XIX (Macaya-Sanz *et al.* 2011).

Microsatellites were assayed following Lexer *et al.* (2005). Genotype data for the Ticino and Tisza hybrid zones were available from a different study (Lindtke *et al.* 2012), except for markers Yin1, Yin2, Con58.1, ConC49.1 and Con7.1 on chromosome XIX. These loci were recently developed by Macaya-Sanz *et al.* (2011), but genotype data for natural hybrid zones are presented for the first time here (Tables S2 and S3, Supporting information).

Data analysis

RAD tag processing, reference mapping and SNP calling. The genomes of *Populus* spp. are known to be highly syntenic (Cervera *et al.* 2001). We used the available *Populus trichocarpa* genome (http://www.phytozome.org/cgi-bin/gbrowse/poplar/) for RAD reference mapping and to identify SNP candidates. SNP calling was based on output from the BOWTIE (version 0.11.3; Langmead *et al.* 2009) and SAMTOOLS (0.1.12a; Li *et al.* 2009) algorithms and custom scripts to identify SNP candidates. Reference mapping with BOWTIE took sequence quality information into account, allowed for up to three mismatches (4.28%) between each read and

the reference sequence and ignored reads which mapped against more than a single position in the genome. The stringent mismatch cut-off implies that our analysis may have been constrained to more conserved regions, which was deemed acceptable because our focus was not on the detection of highly divergent outlier loci (see below; Discussion). SAMTOOLS tabulated SNP results for all individuals (using the 'mpileup' module), and we retained information on the number of reads covering each SNP. The initial SNP table comprised more than 350 000 putative SNPs, of which we only accepted those positions which were covered by 6–250 reads in each individual (Table S4, Supporting information). This allowed us to detect alternative alleles if present while at the same time reducing the number of paralogous SNPs from repetitive (duplicated) DNA segments. We additionally restricted our analysis to those SNPs which were unambiguously mapped on the first 19 scaffolds (i.e. chromosomes) of the *Populus trichocarpa* reference genome. Finally, we removed all SNP candidates with three or more alleles and also excluded SNPs which were monomorphic in our two target species but different from the reference genome. We did not implement a minor allele frequency threshold, as our sampling design ($N = 28$ haploid genomes sampled without missing data) already implies a detection threshold of 3.6%. Results of the reference mapping and SNP calling, including information on the position of each SNP, its allelic state, the number of reads covering the SNP and information on RAD tags sequenced in each individual, are shown in Table S4 (Supporting information).

RAD scan for genomic divergence. As surrogates for genomic divergence, we calculated two parameters using sliding window analysis (window size = 1×10^6 bps; step size = 200 kb): the proportion of SNPs that are fixed between species among all variable SNPs in the data set and F_{ST} within Weir & Cockerham's (1984) analysis of variance framework for each locus in ARLEQUIN 3.5 (Excoffier & Lischer 2010). The former (proportion of fixed SNPs) is simpler and more easily interpretable, and comparison to F_{ST} was deemed of interest because of the widespread use of this parameter in population genomics. In taxa with 'porous' genomes, regions with greatly elevated divergence are often thought to point to potential 'islands' or 'continents' of speciation (Nosil *et al.* 2009), and regions with greatly reduced divergence indicate recurrent gene flow or shared ancestral polymorphism (Lexer *et al.* 2006; Palma-Silva *et al.* 2011). Localized genomic divergence is also known to be influenced by a variety of molecular and genomic features (Lercher *et al.* 2002; Guo & Jamison 2005), and quantifying the relative roles of these is of interest before interpreting genomic patterns of divergence in terms of the origin or maintenance of reproductive barriers (e.g. Roesti *et al.* 2012). We therefore measured GC content, the presence of repetitive elements, the fraction of ambiguous (non-ACGT) base pairs and SNP densities in windows of 1×10^5 and 1 million base pairs across the *P. trichocarpa* reference genome, using the BIOCONDUCTOR packages BIOSTRINGS (http://www.bioconductor.org/packages/2.2/bioc/html/Biostrings.html) and GENER (http://www.bioconductor.org/packages/2.3/bioc/html/GeneR.html) for R. Additionally, repetitive elements with motifs <2 kb were identified in *P. trichocarpa* genome assembly version 2 using the TANDEM REPEATS FINDER software (Benson 1999). The relationships between all these features and windowed estimates of genomic divergence were tested using Pearson's correlations.

Genomic autocorrelation analysis of SNP differentiation. Sliding window autocorrelation analyses as implemented in the function correlog, part of the NCF package for the R environment (http://cran.r-project.org/web/packages/ncf/index.html) identified the maximum strength and reach of autocorrelations among RAD polymorphisms along each chromosome. We analysed the autocorrelation between the proportions of fixed SNPs (above) along all 19 *P. trichocarpa* chromosomes (Table S5, Supporting information). Measures of autocorrelation (Moran's I) were calculated for blocks of 4 megabases (4×10^6 bp), sliding by 250 kilobases along chromosomes. Correlog analyses used a distance class size ('increment' argument) of 200 kb, and we permuted the data 1000 times under the null hypothesis to assess significance of Moran's I (http://cran.r-project.org/web/packages/ncf/ncf.pdf). For each 4 megabase window, we calculated an autocorrelogram and extracted the largest Moran's I value from any autocorrelation size class smaller than or equal to the maximum spatial distance for which significant autocorrelation was found. A single Moran's I estimate was thus plotted against the midpoint of each 4 megabase window to identify chromosomal regions of significant autocorrelation. This allowed visualization of spatial autocorrelations along the chromosomes, where genomic blocks appear as rows of neighbouring significant autocorrelations. Note that significant spatial autocorrelations along chromosomes may be caused by many different factors (see Discussion), including cyclic changes in the proportions of fixed SNPs, and that the highest autocorrelations might be observed on the edges of particular regions, not necessarily in their centre. In case of genetically divergent, hybridizing taxa, extended regions of spatial autocorrelation in genomic regions with low divergence are plausibly explained by recent gene flow (Scotti-Saintagne *et al.* 2004). We use our

autocorrelation analysis here in an exploratory fashion on nonindependent, overlapping windows along each chromosome and therefore do not correct for multiple testing.

Ancestry and introgression in hybrid zones. To study ancestry and introgression of mapped microsatellites in hybrid zones in a simple and graphical way, LSA's for 32 loci were computed using the R package INTROGRESS, which calculates maximum-likelihood admixture proportions and genomic clines (Gompert & Buerkle 2009; Gompert & Buerkle 2010). The ancestry of microsatellite alleles was estimated at each locus as the number of alleles originating from the focal species (i.e. 0, 1 or 2 alleles), after binning microsatellite alleles into two allelic classes per locus as in Lexer *et al.* (2007). Bayesian admixture analysis with the linkage model of the STRUCTURE (Falush *et al.* 2003) software was used as an alternative approach to estimate LSAs and clines in hybrid zones, following Lexer *et al.* (2010). As the results and conclusions from both analyses were similar, only the results of the simpler INTROGRESS analyses will be presented here.

Results

SNP calling

RAD sequencing of *Populus alba* and *Populus tremula* ($N = 14$ haploid genomes per species; Tables 1 and S1, Supporting information) yielded on average 93 000 separate RAD tags per individual with high-quality Illumina GAIIx reads at a median sequencing depth of 11-fold (Table S1, Supporting information). Reference mapping against the *Populus trichocarpa* genome identified a preliminary set of 350 301 putative SNP sites. After applying stringent quality checks (see Methods), 38 525 biallelic high-quality SNPs on the first 19 scaffolds of the *Populus* reference genome were chosen for further analysis. These high-quality SNPs were represented in each individual (i.e. no missing data) by an average sequencing depth of 19 Illumina sequencing reads (median = 18; SD = 11). This is sufficient to detect both allelic states if present, within the limits of our population samples to detect rare alleles in heterozygous state.

Given a known genome size of ~378.5×10^6 bp for the first 19 *P. trichocarpa* scaffolds (i.e. chromosomes), SNPs were screened on average every 9.7 kb (SD: 34.5 kb; range, 2–1 470 000 bp; median: 32 bp) across the genome. The low median SNP distance reflects the frequent presence of more than one SNP per RAD tag. Dense genomic coverage is also illustrated by the rare (only 29) number of uncovered genomic intervals larger than 500 kb in length. This SNP density is equivalent to an average number of 88.8 SNPs per million base pairs (±40.3 SD, range, 0–186, median: 93; Fig. 1; Tables S4 and S5, Supporting information).

Genomic patterns of divergence

Considerable genomic divergence was observed between the ecologically divergent target species, *P. alba* and *P. tremula*. This was reflected by the observation that >15% of all high-quality SNPs were fixed between species. Windowed analyses revealed that the proportion of fixed SNPs was significantly correlated with F_{ST} (Fig. 1; Table S6, Supporting information; $r^2 = 0.734$; $P < 0.001$). The two species were differentiated by an average multilocus F_{ST} of 0.634 (95% confidence interval, 0.631–0.638).

Visual inspection of windowed analyses (1×10^6 bp window size) revealed that the proportion of fixed SNPs in any given window was highly variable, as were the average fixation index (F_{ST}) and SNP density (Fig. 1). Detailed windowed analyses (1×10^5 bp window size; Table S5, Supporting information) identified an average proportion of fixed SNPs of 14.8%, with a standard deviation of ±17.9% and a median proportion of fixed SNPs of 11.1% per window. We did not identify any fixed SNP within the first three megabases of chromosome XIX (0–3.5×10^6 bp), the putative incipient sex chromosome of poplar; Fig. 1; Tables S4 and S5, Supporting information). This observation also held after the removal of individuals of female or unknown sex (Table 1; results not shown).

Associations with features of the Populus reference genome

Associations between SNP diversity and genomic features of *Populus* allowed us to identify factors that may potentially influence SNP detection and analysis. For example, the total number of SNPs per window was negatively correlated with the fraction of repetitive DNA elements ($r^2 = -0.336$; $P < 0.001$). Also, the total numbers of both variable and fixed SNPs were correlated with GC content ($r^2 = 0.296$ and 0.217, respectively; $P < 0.001$; Table S6 and Fig. S1, Supporting information). In general, correlations between windowed divergence estimates and genomic features of *Populus* were weak (Table S6 and Fig. S1, Supporting information) and were thus not pursued further. High levels of divergence (great proportions of fixed SNPs) were associated with low SNP densities, whereas low divergence loci exhibited a large range of SNP densities (Fig. 2), two observations of relevance to our discussion of the RAD-Seq data (below).

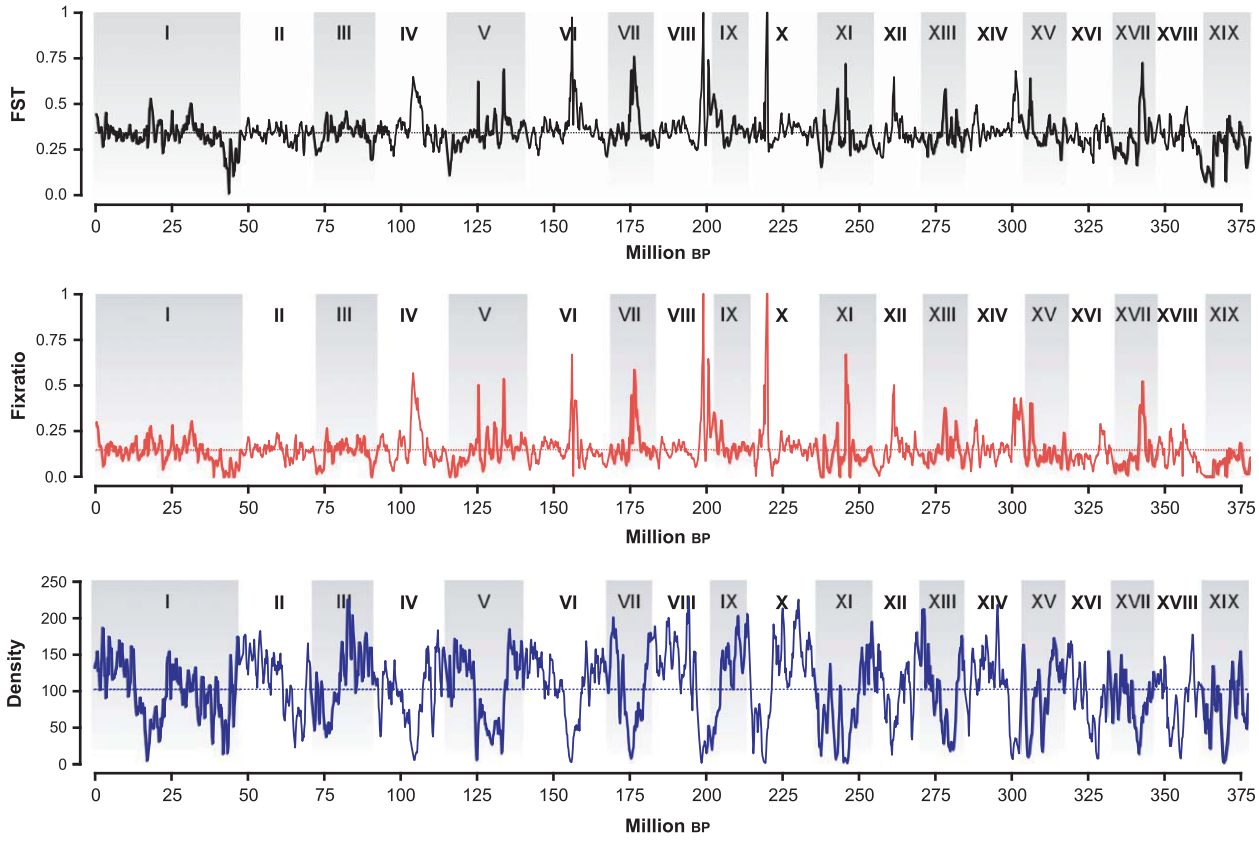

Fig. 1 Patterns of interspecific genomic divergence between *Populus alba* and *P. tremula*. Sliding window analysis (window size: 1×10^6 bp, step size 200 kb) for average fixation indices (F_{ST}, top), proportions of fixed SNPs among all variable SNPs (fixratio, middle) and SNP density (bottom) in each window. Results of all windowed analyses are plotted against window midpoints in millions of base pairs (bp). Grey and white vertical blocks highlight different chromosomes identified by roman numerals.

Genomic autocorrelations

Genomic autocorrelation of the proportion of fixed SNPs was used as a simple way of detecting chromosome blocks of higher or lower than average interspecific divergence in the *Populus* genome. Windowed analyses identified autocorrelated regions particularly within the first three megabases of chromosome XIX (0–3.55×10^6 bp; Fig. 3; Table S5, Supporting information), the incipient sex chromosome of *Populus*, but large autocorrelated blocks were also visible in several other chromosomal regions with unusually low (e.g. chromosomes I, III and XVIII) and unusually high interspecific divergence (e.g. chromosomes V, XIII and XV) (Figs 3 and S2, Supporting information). The extended stretches of autocorrelations in the beginning of chromosome XIX were defined by low divergence regions entirely free of fixed SNPs (Figs 1 and 3; Table S5, Supporting information).

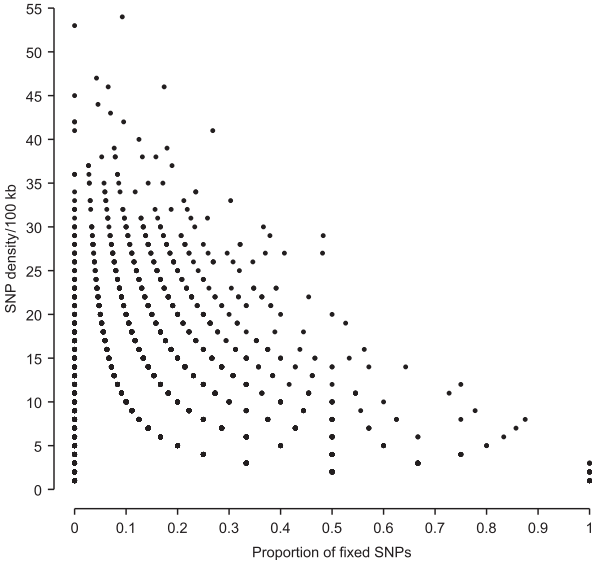

Fig. 2 Relationship between SNP densities (*y*-axis; number of SNPs per 100 kilobase window) and divergence (*x*-axis; proportion of fixed SNPs) in *Populus alba* and *P. tremula*. The association between genomic windows with great divergence and low SNP density is clearly visible.

Ancestry and introgression of mapped microsatellites

The large number of individuals from two independent hybrid zones genotyped for microsatellites allowed us

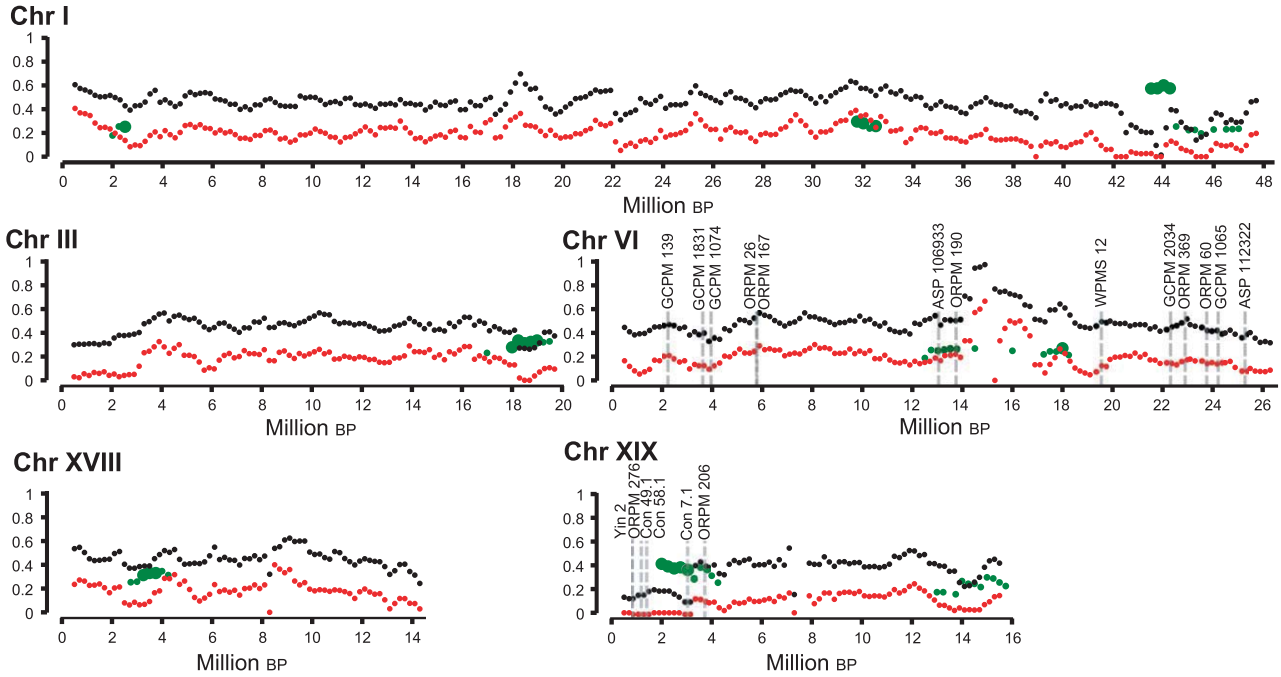

Fig. 3 Autocorrelation (Moran's *I*) analysis for five exemplary *Populus* chromosomes (drawn to scale) based on the proportion of fixed SNPs. Sliding window analyses (width: 4×10^6 bp, sliding by 250 kb) identify highly significant (large green dots; $P < 0.01$) and significant (small green dots, $0.01 < P < 0.05$) Moran's *I* values. Average F_{ST} (black) and proportions of fixed SNPs (red) are indicated for sliding windows of 1×10^6 bp width, sliding by 200 kb. Chromosomes I, III, XVIII and XIX are shown to illustrate low divergence blocks of autocorrelated markers, and chromosome VI is an inconspicuous reference chromosome used in the text. The genomic positions of all analysed microsatellite loci are indicated on chromosomes VI and XIX. Note the strong and significant autocorrelations in the beginning of chromosome XIX, the putative incipient sex chromosome of *Populus*.

to compare ancestries and introgression on chromosomes XIX and VI (Fig. 4). The INTROGRESS results indicate increased introgression of microsatellite alleles across both hybrid zones on chromosome XIX, as visible from the great frequency of homozygotes of one species in the genomic background of the other (Fig. 4). Thus, introgression analysis of mapped microsatellites corroborates genomic patterns of divergence found with RAD-Seq. Results of Bayesian STRUCTURE analysis (not shown) were largely congruent with the INTROGRESS results, thus suggesting a lack of RI in this region.

Discussion

The 'genomic landscape' of population divergence and speciation in taxa with 'porous genomes' is a hotly debated topic in molecular ecology and evolutionary biology (Nosil *et al.* 2009; Feder & Nosil 2010; Michel *et al.* 2010; Pritchard *et al.* 2010; Wolf *et al.* 2010; Feder *et al.* 2012; Roesti *et al.* 2012). Few studies have taken advantage of 'next generation sequencing' based genome scans for addressing these topics, and those that have done so have focused on the genomic signatures of adaptive population divergence or domestication (e.g. Hohenlohe *et al.* 2010; Rubin *et al.* 2010; Turner *et al.* 2010; Amaral *et al.* 2011), rather than the genomics of reproductive barriers between hybridizing species with divergent genomes (but see Nadeau *et al.* 2012; Gompert *et al.* 2012). Here, we have used a recently developed short-read sequencing approach, RAD-Seq (Baird *et al.* 2008; Hohenlohe *et al.* 2010), to infer the genomic signature of gene flow across a 'porous' species boundary between two highly divergent, hybridizing species. Our study complements a growing number of genomic scans of less divergent (usually infraspecific) taxa with 'porous' genomes (e.g. Dopman *et al.* 2005; Turner *et al.* 2005; Roesti *et al.* 2012). More specifically, this study contributes to our understanding of the evolutionary process in taxa that exhibit both, widespread genomic divergence and episodes of sympatry and genetic exchange (Michel *et al.* 2010; Feder *et al.* 2012).

Potential of RAD tags for genome scans of highly divergent species

Little experience is currently available on the usefulness of RAD-Seq for evolutionary genomics work involving highly divergent species. By making use of the *Populus trichocarpa* genome sequence for reference mapping and quality control, we have shown that

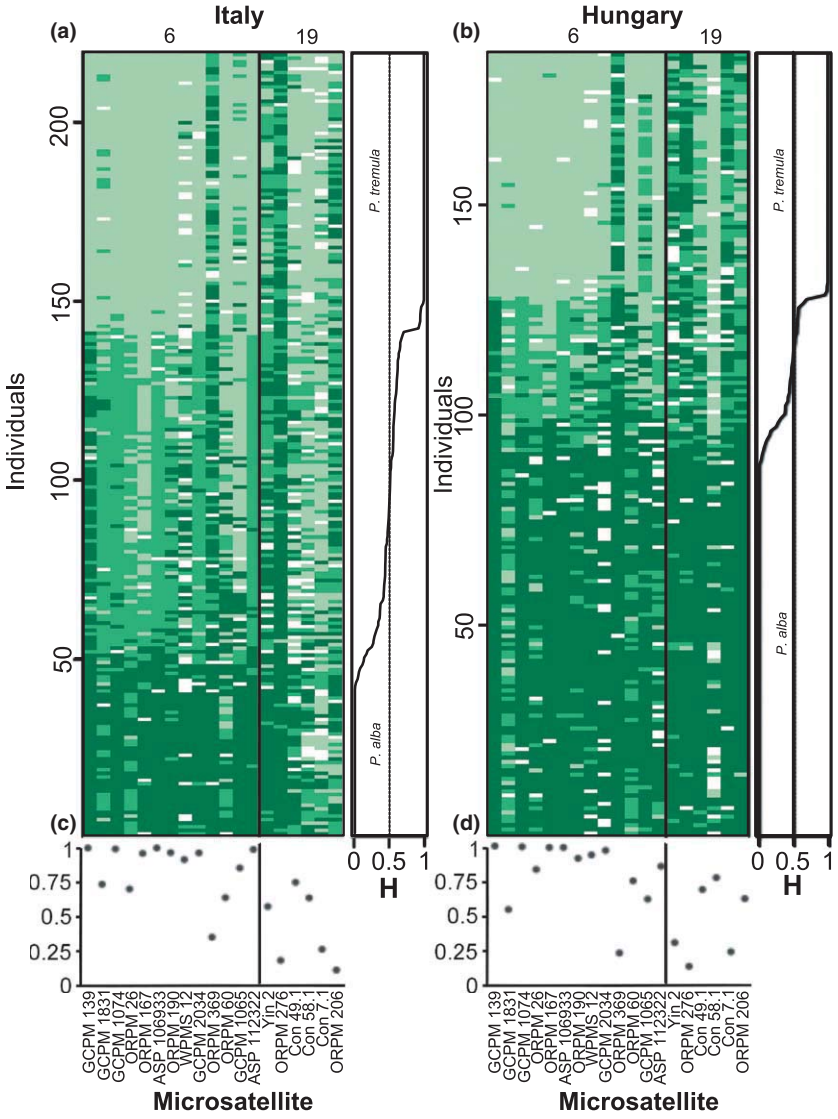

Fig. 4 Locus-specific ancestries (LSAs) for mapped microsatellites on chromosomes VI and XIX. Ancestries for the Italian hybrid zone (a) (N = 219 indviduals) and the Hungarian hybrid zone (b) (N = 186 individuals) are plotted separately. Individuals (y-axis) are sorted by their maximum-likelihood hybrid index (H), and microsatellite loci (x-axis) are ordered by their genomic position along each chromosome. Shades of green reflect allelic ancestry for each locus and individual: dark green, homozygous for *Populus alba*; light green, homozygous for *P. tremula*; medium green, heterozygous for alleles from both species; white, missing data. The allele frequency differentials for each microsatellite marker in the Italian (c) and Hungarian (d) hybrid zone are indicated at the bottom of the graph. See Table S2 (Supporting information) and Methods for details of microsatellite markers and Table S3 (Supporting information) for genotypic data.

RAD tags can facilitate genomic scans between species that have diverged millions of years (Stettler *et al.* 1996) and that hybridize and recombine despite considerable levels of genomic divergence (average multilocus F_{ST} for SNP data in the present study = 0.634; F_{ST} and standardized differentiation (G'_{ST}) for microsatellites up to 0.37 and 0.63, respectively; Lexer *et al.* 2010). The fact that >38 000 high-quality SNPs were recovered in these species after stringent quality checks (Methods) should encourage other students of highly divergent, hybridizing species with complex genomes. Note that *Populus alba* and *Populus tremula* possess diploidized genomes, but that *Populus* has undergone at least two rounds of whole-genome duplication in deep evolutionary history (Tuskan *et al.* 2006). Similar situations are commonly observed in evolutionary radiations of plants, and thus, our results are informative on the utility of RAD-Seq in plant evolutionary genomics.

The availability of a genome sequence for a poplar (Tuskan *et al.* 2006; *P. trichocarpa* genome assembly version 2) allowed us to assess the degree to which genomic features of reference genomes may influence RAD-based genome scans. For example, we uncovered weak negative correlations between SNP density and the fraction of repetitive elements across the *Populus* genome and weak positive correlations between SNP density and GC content (Table S6 and Fig. S1, Supporting information). There also was a conspicuous association between the genomic distributions of divergence (proportion of fixed SNPs, F_{ST}) and SNP densities (Fig. 1). Whereas genomic regions of high divergence were clearly associated with low SNP density, regions of unusually low divergence were not (Fig. 2). The former observation

may indicate that both high divergence and low SNP density are consequences of the same biological process, that is, divergent selection reducing genetic diversity within species or more rapid evolution of some genome regions, effectively elevating divergence and reducing the number of polymorphisms detected by a RAD-Seq scan with stringent quality criteria (see Methods).

While the high divergence peaks uncovered by our genome scan (Fig. 1) may involve genes or genetic elements of importance during advanced stages of speciation, we wish to focus on genomic regions of low divergence in the present study. The genomes of *P. alba* and *P. tremula* are highly divergent overall (Lexer *et al.* 2010; average multilocus F_{ST} in the present study = 0.634); thus, we might expect large 'continents' rather than 'islands' of speciation (Nosil *et al.* 2009; Michel *et al.* 2010; Wolf *et al.* 2010; but see Barton & Bengtsson 1986). In effect, most of the genome will be protected from gene flow except for occasional 'pores' (Wu & Ting 2004; Lexer & Widmer 2008). Thus, loci or regions with greatly reduced differentiation (compared to the genomic average) should point to features of interest to the evolutionary genomics of taxa with 'porous' genomes, such as recent interspecific gene exchange or shared ancestral polymorphism (Muir & Schlotterer 2005; Lexer *et al.* 2006; discussed below). Note that the stringent mismatch cut-off used for sequence alignment (3 mismatches = 4.28% per 70 bp read) implies an underestimation of overall genomic divergence in this study, that is, our detection of low divergence regions against the genomic background is conservative.

Causes of allele sharing between ecologically divergent, hybridizing species

Shared alleles between hybridizing species can be traced back primarily to three mechanisms: homoplasy, shared ancestral polymorphism predating the origin of the species (retained because of uniform selection pressures or insufficient time since divergence) and recurrent gene flow (Martinsen *et al.* 2001; Lexer *et al.* 2006; Palma-Silva *et al.* 2011). At this level of evolutionary divergence, homoplasy is unlikely to be the chief cause of allele sharing for DNA sequence polymorphisms (Zhang & Hewitt 2003). Shared ancestral polymorphism is a more likely explanation, given the long generation times and large effective population sizes of these wind-pollinated, dioecious trees (N_e = 118 000 in *P. tremula*; Ingvarsson 2008). Hybridization and interspecific gene flow, on the other hand, are known to be frequent in localities where these species cooccur (Lexer *et al.* 2005, 2010).

The relative roles of shared ancestral variation vs. recurrent gene flow can sometimes be inferred from biogeographical patterns of diversity (Petit *et al.* 2002; Muir & Schlotterer 2005; Palma-Silva *et al.* 2011). This is the case because the genetic signature of hybridization is likely to exhibit a strong spatial component, whereas that of shared ancestral variation does not. Nevertheless, this method is often limited by the inability to infer all key aspects of species' biogeographical history. A more powerful way to address this issue is to test for genomic block structure among loci with high levels of allele sharing (Scotti-Saintagne *et al.* 2004; Lexer *et al.* 2006). This is based on the expectation that recent gene flow between divergent populations will create blocks of LD along the chromosomes (Chapman & Thompson 2002; Scotti-Saintagne *et al.* 2004), the size of blocks depending primarily on temporal patterns of admixture (Buerkle & Lexer 2008; Winkler *et al.* 2010).

We found 11 chromosomes with windows exhibiting highly significant genomic autocorrelation of divergence (proportion of fixed SNPs), and several of these were due to unusually low divergence, that is, allele sharing. Chromosome XIX, the incipient sex chromosome of *Populus*, represents a prime example (Fig. 3; discussed below). Other examples can easily be found towards the distal ends of chromosomes I and III, or between megabases three and four of chromosome XVIII (Figs 3 and S2, Supporting information).

The observed block structure indicates that recurrent gene flow has contributed to genomic patterns of allele sharing seen between these ecologically divergent forest trees. These findings complement our earlier results based on analyses of genomic admixture in hybrid zones (Lexer *et al.* 2007, 2010). Those studies were able to trace gene flow just within the last few generations, based on the known temporal dynamics of genomic admixture coefficients and hybrid indices (Falush *et al.* 2003; Gompert & Buerkle 2009). In contrast, measures of divergence between species are informative regarding gene flow over hundreds of generations or more (Whitlock 1992). In effect, assuming an average generation time of 30–40 years, it appears that these species' genomes have been affected by hybridization and introgression since thousands if not tens of thousands of years. Our divergence map (Fig. 1) and autocorrelation graphs (Figs 3 and S2, Supporting information) speak for a broad distribution of the sizes of introgressed genomic blocks, ranging from kilobases to megabases, which indicates a complex history of admixture. Future work based on more extensive sampling will help to clarify this issue.

Rampant introgression on an incipient plant sex chromosome?

Sex chromosomes are often seen as 'hotspots' of speciation because of 'Haldane's rule' (Coyne & Orr 2004) and

because suppressed recombination and increased LD in the sex determination region will favour the accumulation of isolation genes (Saether *et al.* 2007; review by Qvarnström & Bailey 2009). Our population genomic data for chromosome XIX, recently suggested as an incipient sex chromosome of *Populus* (Yin *et al.* 2008; Pakull *et al.* 2009; Paolucci *et al.* 2010), suggest the opposite: extensive allele sharing over three megabases surrounding the poplar sex determination region, presumably caused by extensive gene flow across a 'porous' species barrier (Figs 1 and 3). This conclusion is supported by >38 000 SNPs assayed in few individuals and a panel of mapped microsatellites typed in >400 individuals (this study), in addition to previous genetic mapping work carried out in our laboratory (Macaya-Sanz *et al.* 2011).

Early genomic work on sex determination in *P. trichocarpa* suggested a ZW sex system with a female-specific W region on chromosome XIX (Yin *et al.* 2008). Recent genetic mapping in *P. alba* and *P. tremula* has indeed confirmed extensive LD across >560 kilobases of this region (Macaya-Sanz *et al.* 2011); however, codominant microsatellites there behaved like autosomal loci: both homo- and heterozygotes were observed in equilibrium frequencies in known females and males (Macaya-Sanz *et al.* 2011). This suggests that the *Populus* sex chromosome is very young indeed, with pairs or groups of sexually antagonistic loci accumulating in the sex determination region but without full differentiation into heteromorphy (Charlesworth *et al.* 2005), or that factors other than time have constrained its evolution. Either way, the apparent lack of structural differentiation between female and male counterparts facilitates the interpretation of interspecific divergence in trees sampled randomly with regard to sex (see Methods).

Interspecific RAD scanning of chromosome XIX yielded the perhaps most puzzling result of the present study: the complete absence of fixed SNPs in the first three megabases of this chromosome, that is, exactly in the poplar sex determination region (Yin *et al.* 2008; Pakull *et al.* 2009; Paolucci *et al.* 2010), resulting in significant genomic autocorrelation (Fig. 3). This is in line with previous observations of greatly reduced interspecific microsatellite divergence in this region (Macaya-Sanz *et al.* 2011). The results also match with significant segregation distortion of this region in experimental interspecific progeny of *P. alba* and *P. tremula*, consistently favouring donor alleles in the recipient species' background (Macaya-Sanz *et al.* 2011). Microsatellite data for two natural hybrid zones, presented for the first time here, provide even further support for the unusual behaviour of this genomic region.

Locus-specific ancestries estimated by INTROGRESS in both hybrid zones revealed elevated introgression on the proximal end of chromosome XIX (Fig. 4), compared to most loci on our reference chromosome VI (see Methods for rationale of this comparison). Elevated introgression is suggested by several aspects of the INTROGRESS analysis. First, there is a greatly increased number of *P. alba* homozygotes (dark green) in the genomic background of *P. tremula* (light green) (Fig. 4, e.g. microsatellite markers ORPM 276 and Yin 2). This pattern is complicated somewhat by the considerable variation in ancestries present among loci and hybrid zones, which is expected from theory (e.g. Barton & Bengtsson 1986; Buerkle & Lexer 2008; Gompert & Buerkle 2009). Second, a lack of RI of chromosome XIX becomes readily apparent from the absence of clear genomic discontinuities between genetically intermediate, heterozygous individuals (medium green) and either parental species throughout chromosome XIX (Fig. 4). This observation is of relevance because previous studies have found partial RI between genetically intermediate hybrids (medium green in Fig. 4) and their parents (Lexer *et al.* 2010; Lindtke *et al.* 2012). So it appears that the partial reproductive barriers present among these species and hybrids are overcome in the sex determination region. This is consistent with the absence of fixed SNPs in our RAD-based genome scan (Fig. 1) and with generally reduced microsatellite allele frequency differentials in this region (Fig. 4). Third, the results of the present study are congruent with previous results on reduced interspecific microsatellite divergence in wild populations and increased introgression of this region in synthetic interspecific backcross progeny (Macaya-Sanz *et al.* 2011).

An alternative explanation for reduced divergence of the incipient poplar sex chromosome may be uniform selection conserving DNA sequence in this region. Indeed, evolving sex chromosomes are known to be affected by selective sweeps and background selection (review by Bachtrog 2006), which may in principle affect related species in similar ways. Nevertheless, selection pressures such as this would be expected to result in reduced diversity in the sex determination region (the first three megabases of chromosome XIX), which was picked up neither by our RAD scan (Fig. 1) nor by microsatellites (this study, Table S2, Supporting information; Lexer *et al.* 2010; Macaya-Sanz *et al.* 2011).

Our results on gene flow on the incipient poplar sex chromosome are at odds with conventional knowledge on the role of sex chromosomes as 'hotspots' of RI and speciation (Qvarnstrom & Bailey 2009). Future studies should now aim at identifying the molecular and evolutionary mechanisms responsible for increased allele sharing. As this genomic region appears to contain unusually large numbers of nuclear binding site - leucine rich repeat (NBS-LRR) type resistance (R) genes (Yin *et al.* 2008), introgression favoured by balancing selection is a plausible hypothesis (Macaya-Sanz *et al.* 2011).

An alternative explanation could be that this genomic region plays an important role in local adaptation, as already shown for *P. tremula* (de Carvalho *et al.* 2010) and that it thus experiences greatly reduced gene flow within species. This would render this genomic region particularly susceptible to introgression between species due to drift-based processes (Petit & Excoffier 2009). High-density genomic scans for the signature of local adaptation within species would help to clarify these issues, and this work is currently underway.

Conclusion

RAD sequencing of small population samples ($N = 14$ haploid genomes per species in this case) facilitates genomic scans for fixed vs. variable SNPs in hybridizing species. This was demonstrated here by a study on the genomic signature of divergence and gene flow in two ecologically divergent forest trees of the 'model tree' genus *Populus*. Important features of our RAD-Seq results were corroborated by a panel of 32 mapped microsatellites typed in >400 individuals from natural hybrid zones and by previous genetic data on natural populations and experimental progeny. Genomic patterns of allele sharing indicate that the highly divergent *Populus alba* and *Populus tremula* have been affected by recurrent gene flow for timescales in the order of hundreds of generations, if not longer. Our results add to a growing body of evidence that gene flow represents an important force in the evolutionary process, even in diverging taxa that are already highly differentiated throughout much of their genome (Michel *et al.* 2010; Feder *et al.* 2012). The incipient sex chromosome of *Populus* appears to experience rampant introgression rather than the accumulation of isolation genes, which cautions against premature generalizations on the role of sex chromosomes in speciation.

Acknowledgements

We thank Stefano Gomarasca, István Asztalos, Sofia Mangili and others for help during field work, Thelma Barbará and Tressa Atwood for help in the laboratory, Alex Widmer and Nolan Kane for helpful discussions and Alex Buerkle, Maja Greminger and five anonymous referees for reading the manuscript. CL's research on the evolutionary genomics of species barriers in *Populus* was supported by grant no. NE/E016731/1 of the U.K. Natural Environment Research Council (NERC) and grant no. 31003A_127059 of the Swiss National Science Foundation (SNF).

References

Amaral AJ, Ferretti L, Megens HJ *et al.* (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE*, **6**, e14782, doi:10.1371/journal.pone.0014782.

Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH (2012) Recent non-hybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, This Issue.

Bachtrog D (2006) A dynamic view of sex chromosome evolution. *Current Opinion in Genetics and Development*, **16**, 578–585.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376, doi:10.1371/journal.pone.0003376.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **56**, 357–376.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573–580.

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

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.

Cervera MT, Storme V, Ivens B *et al.* (2001) Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. *Genetics*, **158**, 787–809.

Chapman NH, Thompson EA (2002) The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, **162**, 449–458.

Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity*, **95**, 118–128.

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

Dopman EB, Perez L, Bogdanowicz SM, Harrison RG (2005) Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proceedings of the National Academy of Sciences of the USA*, **102**, 14706–14711.

Emerson KJ, Merz CR, Cathen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the USA*, **107**, 16196–16200.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

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.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.

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

Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, **18**, 1207–1224.

Gompert Z, Buerkle CA (2010) Introgress: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, **10**, 378–384.

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

Guo YJ, Jamison DC (2005) The distribution of SNPs in human gene regulatory regions. *BMC Genomics*, **6**, 140, doi:10.1186/1471-2164-6-140.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862, doi:10.1371/journal.pgen.1000862.

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

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lercher MJ, Smith NGC, Eyre-Walker A, Hurst LD (2002) The evolution of isochores: evidence from SNP frequency distributions. *Genetics*, **162**, 1805–1810.

Lexer C, Widmer A (2008) The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3023–3036.

Lexer C, Fay MF, Joseph JA, Nica MS, 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, Kremer A, Petit RJ (2006) Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Molecular Ecology*, **15**, 2007–2012.

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF (2007) Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity*, **98**, 74–84.

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, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindtke D, Buerkle CA, Barbara T *et al.* (2012) Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in *Populus*. *Molecular Ecology*, doi:10.1111/j.1365-294X.2012.05744.x.

Macaya-Sanz D, Suter L, Joseph J *et al.* (2011) Genetic analysis of post-mating reproductive barriers in hybridizing European *Populus* species. *Heredity*, **107**, 478–486.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.

Maynard Smith J, Haigh J(1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

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

Minder AM, Widmer A (2008) A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. *Molecular Ecology*, **17**, 1552–1563.

Muir G, Schlotterer C (2005) Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). *Molecular Ecology*, **14**, 549–561.

Nadeau NJ, Martin SH, Kozak KM *et al.* (2012) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology*, This issue.

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

Pakull B, Groppe K, Meyer M, Markussen T, Fladung M (2009) Genetic linkage mapping in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). *Tree Genetics & Genomes*, **5**, 505–515.

Palma-Silva C, Wendt T, Pinheiro F *et al.* (2011) Sympatric bromeliad species (*Pitcairnia* spp.) facilitate tests of mechanisms involved in species cohesion and reproductive isolation in Neotropical inselbergs. *Molecular Ecology*, **20**, 3185–3201.

Paolucci I, Gaudet M, Jorge V *et al.* (2010) Genetic linkage maps of *Populus alba* L. and comparative mapping analysis of sex determination across *Populus* species. *Tree Genetics & Genomes*, **6**, 863–875.

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

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

Petit RJ, Csaikl UM, Bordacs S *et al.* (2002) Chloroplast DNA variation in European white oaks – phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management*, **156**, 5–26.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

Qvarnstrom A, Bailey RI (2009) Speciation through evolution of sex-linked genes. *Heredity*, **102**, 4–15.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

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

Saether SA, Saetre GP, Borge T *et al.* (2007) Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science*, **318**, 95–97.

Savolainen V, Anstett MC, Lexer C *et al.* (2006) Sympatric speciation in palms on an oceanic island. *Nature*, **441**, 210–213.

Scotti-Saintagne C, Mariette S, Porth I *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615–1626.

Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM (1996) Biology of *Populus* and its implications for management and conservation, NRC Research Press, Ottawa, Ontario.

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 364–373.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285, DOI: 10.1371/journal.pbio.0030285.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *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.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

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

Whitlock MC (1992) Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution*, **46**, 608–615.

Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, **11**, 65–89.

Wolf JBW, Lindell J, Backstrom N (2010) Speciation genetics: current status and evolving approaches. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 1717–1733.

Wu CI, Ting CT (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114–122.

Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA (2004) Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theoretical and Applied Genetics*, **109**, 451–463.

Yin T, DiFazio S, Gunter LE *et al.* (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Research*, **18**, 422–430.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.

This study arose from K.N.S.'s postdoctoral work in C.L.'s group at University of Fribourg within C.L.'s larger research programme on the ecological and evolutionary genomics of 'porous' species boundaries in *Populus*. D.L., C.C. and S.W. are group members with interests in the molecular ecology and evolutionary genomics of speciation. R.N. is Senior Vice President of Plant Genomics at Floragenex, Inc. S.C.'s interests are primarily in forest tree biology, conservation and breeding.

Data accessibility

The microsatellite data set is made available in Table S3 (Supporting information) and SNP data can be found in Table S4 (Supporting information).

Supporting information

Fig. S1 Influence of genomic features on SNP distributions.

Fig. S2 Autocorrelation (Moran's I) analysis for all 19 *Populus* chromosomes (drawn to scale) based on the proportion of SNPs fixed between species.

Table S1 RAD sequencing results for 14 samples of *Populus alba* and *P. tremula*.

Table S2 Thirty-two microsatellite loci ('Locus') with detailed information on genomic location of each marker (Chr, chromosome; Pos, location of the locus in base pairs on the chromsome).

Table S3 Introgress analyses are based on data from 32 microsatellite loci ('locus').

Table S4 Details on 38 525 SNPs passing quality control.

Table S5 We report windowed statistics for the reference genome of *Populus trichocarpa* and the reference mapped SNPs obtained from RAD sequencing.

Table S6 Pair-wise Pearson correlation coefficients between windowed (non-overlapping windows sized 100 kb, Table S5, Supporting information) estimates of differentiation between the genomes of *Populus alba* and *P. tremula*, and windowed genomic features affecting SNP distributions: average FST (avgFST), number of fixed SNPs (fixed), number of variable SNPs (variable), number of SNPs/100 kb (density), proportion of fixed over informative SNPs (fixratio), fraction of repetitive DNA elements (repfract), ratio of GC basepairs (gcratio), and fraction of ambiguous (non-ACGT) basepairs (xratio).