

## ORIGINAL ARTICLE

Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differencesC Lexer<sup>1</sup>, CA Buerkle<sup>2</sup>, JA Joseph<sup>1</sup>, B Heinze<sup>3</sup> and MF Fay<sup>1</sup><sup>1</sup>Jodrell Laboratory, Royal Botanic Gardens, Richmond, Surrey, UK; <sup>2</sup>Department of Botany, University of Wyoming, Laramie, WY, USA and <sup>3</sup>Department of Genetics, Federal Research Centre for Forests, Hauptstrasse, Vienna, Austria

The use of admixed human populations to scan the genome for chromosomal segments affecting complex phenotypic traits has proved a powerful analytical tool. However, its potential in other organisms has not yet been evaluated. Here, we use DNA microsatellites to assess the feasibility of this approach in hybrid zones between two members of the 'model tree' genus *Populus*: *Populus alba* (white poplar) and *Populus tremula* (European aspen). We analyzed samples of both species and a Central European hybrid zone ( $N=544$  chromosomes) for a genome-wide set of 19 polymorphic DNA microsatellites. Our results indicate that allele frequency differentials between the two species are substantial (mean  $\delta = 0.619 \pm 0.067$ ). Background linkage disequilibrium (LD) in samples of the parental gene pools is moderate and should respond to sampling schemes that minimize drift and

account for rare alleles. LD in hybrids decays with increasing number of backcross generations as expected from theory and approaches background levels of the parental gene pools in advanced generation backcrosses. Introgression from *P. tremula* into *P. alba* varies strongly across marker loci. For several markers, alleles from *P. tremula* are slightly over-represented relative to neutral expectations, whereas a single locus exhibits evidence of selection against *P. tremula* genotypes. We interpret our results in terms of the potential for admixture mapping in these two ecologically divergent *Populus* species, and we validate a modified approach of studying genotypic clines in 'mosaic' hybrid zones.

*Heredity* (2007) 98, 74–84. doi:10.1038/sj.hdy.6800898; published online 20 September 2006

**Keywords:** hybrid zone; admixture; introgression; linkage disequilibrium; rare alleles; *Populus*

## Introduction

'Admixture mapping' as suggested by Chakraborty and Weiss (1988) and Briscoe *et al.* (1994) utilizes linkage disequilibrium (LD) induced by the mixing of genes from two divergent gene pools. In an outcrossing species and in the absence of confounding population structure, LD will decay with increasing genetic map or chromosomal distance (Lynch and Walsh, 1998), because the chance that stretches of DNA are broken up by recombination becomes greater the further two loci are apart. Admixture will effectively widen the region of a genome that is affected by LD, because recombination will take several/many generations to break up the chromosome blocks derived from each parental population (Briscoe *et al.*, 1994; Chapman and Thompson, 2002). Hence, admixture potentially facilitates molecular marker-based 'genome-scans' to narrow in on genomic regions conferring trait differences between two divergent source gene pools (Chakraborty and Weiss, 1988; Briscoe *et al.*, 1994; McKeigue *et al.*, 2000; Pfaff *et al.*, 2001). This prediction has been verified recently by

successful admixture genome-scans for two complex traits in humans – hypertension and susceptibility for multiple sclerosis (Reich *et al.*, 2005; Zhu *et al.*, 2005).

The requirements for genome-scans through admixture in humans have been carefully evaluated by geneticists for years (e.g., McKeigue *et al.*, 2000; Pfaff *et al.*, 2001; Hoggart *et al.*, 2004). In addition to setting the stage for association mapping in human medicine, these studies have encouraged the development of similar approaches in natural admixed populations or 'hybrid zones' of wild animals or plants (Rieseberg *et al.*, 1999; Rieseberg and Buerkle, 2002). In 'non-human' organisms, admixture mapping holds enormous potential for studies addressing the genetic changes that occur during divergence of populations, ecotypes, or species. This may allow geneticists to address some of the biggest issues in evolutionary biology, for example the number and effect sizes of genes involved in adaptation and speciation (Fisher, 1930; Wright, 1931), the genetic architecture of barriers that keep previously diverged genomes from merging upon secondary contact (Barton and Hewitt, 1985; Barton and Gale, 1993), or the likelihood of spread of advantageous alleles (Morjan and Rieseberg, 2004). These questions are often addressed by quantitative trait locus (QTL) mapping of trait differences in 'mapping populations' derived from experimental crosses (Orr, 2001). However, crosses are often difficult to obtain in wild taxa, for example in species that are long-lived or

Correspondence: Dr C Lexer, Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK.

E-mail: c.lexer@kew.org

Received 17 September 2005; revised 28 July 2006; accepted 14 August 2006; published online 20 September 2006

otherwise of limited tractability – these are the taxa for which admixture mapping of complex traits would be most useful. An important aspect is that in evolutionary biology interest often will be directed toward identifying genomic segments subject to natural selection, rather than segments associated with a particular trait alone (Rieseberg *et al.*, 1999; Wu, 2001). For simplicity, variation in such segments is referred to as ‘adaptive or detrimental variation’ from here onwards.

*Populus* has been suggested as a ‘model forest tree’ for studying tree form, function, and evolution (Taylor, 2002), interactions between ecological carrier species and their communities (Whitham, 1989), and how ecology interacts with plant development (Cronk, 2005). The favorable genetic attributes of *Populus* such as small genome size (550 Mb;  $2C = 1.1$  pg), diploidy throughout the genus ( $2n = 38$ ), porous species barriers (Rajora and Dancik, 1992; Martinsen *et al.*, 2001; Floate, 2004), and a near-complete genome sequence with thousands of expressed sequence tags and markers (<http://www.ornl.gov/sci/ipgc>) make this genus an ideal candidate for evaluating the ‘admixture mapping’ approach in a ‘non-human’ organism. Here, we focus on two species that hybridize frequently in Europe, *Populus alba* (white poplar) and *Populus tremula* (European aspen).

*P. alba* and *P. tremula* share a wide sympatric or parapatric distribution across large parts of Central and Southern Europe and hybrid zones often form between them (Rajora and Dancik, 1992; Fossati *et al.*, 2004; Lexer *et al.*, 2005). A recent sequencing survey of nuclear genes in one of the two species (*P. tremula*; Ingvarsson, 2005) indicates that LD may often not extend beyond a single gene, suggesting that the detection of genetic associations in wild *intra*-specific populations may be difficult. *P. alba* and *P. tremula* exhibit several features that render them potentially suitable for admixture mapping of detrimental or adaptive trait variation. Firstly, they are ecologically divergent: *P. alba* is restricted mainly to lowland flood-plain forests whereas *P. tremula* occurs in mixed upland communities (Adler *et al.*, 1994), and traits potentially involved in these divergent ecological preferences have been identified (Karrenberg *et al.*, 2002). Second, species barriers in *Populus* are likely to be genetic rather than chromosomal (Cervera *et al.*, 2001; Lexer *et al.*, 2005), and thus variation at genetic factors involved in species isolation should segregate in hybrids. Also, the two species differ in numerous diagnostic morphological characters (Adler *et al.*, 1994), hybrids display large phenotypic variances in multiple traits, and the proportion of recombinant (backcrossed) genotypes in hybrid zones is high (Rajora and Dancik, 1992; Lexer *et al.*, 2005).

Here, we ask the following questions regarding the potential of hybrid zones between *P. alba* and *P. tremula* for admixture genome-scanning: (1) How large are DNA microsatellite allele frequency differentials between the two species? (2) How large is background LD among unlinked loci in different hybrid genotypic classes and each parental species and what causes it? (3) How large is the variation in introgression rates among highly informative marker loci? We use our data to assess the feasibility of admixture mapping-related studies in European *Populus*, and we discuss a modified approach of interpreting genotypic clines that may be useful for studying the evolution of ‘mosaic’ hybrid zones.

## Materials and methods

### Sampling of hybrid zone and parental populations

A large ‘mosaic’ hybrid zone between *P. alba* and *P. tremula* (hybrids commonly known as *P. x canescens*) was sampled in the Danube valley near Vienna, Austria. The sampling area covered a linear distance of approximately 110 km of the river valley between Krems and Hainburg, Austria, and included lowland flood-plain ‘gallery’ forest located within the Danube Floodplain National Park (<http://www.donauauen.at/>) and adjacent areas. *P. x canescens* hybrid morphotypes were sampled in such a way as to maximize geographic coverage within the hybrid zone (sampling a transect was not feasible because of the patchy distribution of remnant forests and suitable habitats within forests). Leaf material was collected for DNA extraction.

Two neighboring ‘populations’ were sampled for each of the two parental species. These samples were described in Lexer *et al.* (2005) and are referred to as ‘subpopulations’ here, since molecular analyses indicate that levels of gene exchange are high ( $N_e m > 3.0$  in both species; Lexer *et al.*, 2005) and that they form one panmictic unit for each species. Subpopulations of *P. alba* were sampled in the Danube valley in Austria (within the zone of sympatry; sampling mid-point: 48.26°N, 16.27°E) and Romania (outside the zone of sympatry; sampling mid-point: 43.77°N, 23.96°E). *P. tremula* was sampled from the Austrian Danube (within the zone of sympatry; sampling mid-point: 48.28°N, 15.89°E) and from the Eastern Alps in Austria (outside the zone of sympatry; sampling mid-point: 46.62°N, 13.85°E). The sample sizes were: 378 chromosomes for *P. x canescens* hybrids, 88 chromosomes for *P. alba* and 78 chromosomes for *P. tremula*. In effect, the sampling of the hybrid zone was increased by ~100% since the initial characterization of the hybrid zone (Lexer *et al.*, 2005).

### Within-genome sampling

The 19 microsatellite markers used in this study were developed by Tuskan *et al.* (2004) and Van der Schoot *et al.* (2000) as indicated on the web-site of the International *Populus* Genome Consortium: [http://www.ornl.gov/sci/ipgc/ssr\\_resource.htm](http://www.ornl.gov/sci/ipgc/ssr_resource.htm). These 19 markers, listed in Table 1, can be considered genetically independent, since they are located either on different chromosomes or widely spaced on the same linkage group of the *P. trichocarpa* x *P. deltoides* genetic map (Yin *et al.*, 2004); levels of synteny in *Populus* are high as indicated by comparative genetic mapping (Cervera *et al.*, 2001). Six of the markers used in this study have not yet been placed on the *Populus* genetic map, but exact tests for LD in *P. tremula* (the species with the larger effective population size  $N_e$ ) indicated no genetic association among any of these loci. Our choice of using unlinked markers for this study reflects the situation encountered by many students of non-model organisms, where a limited number of unlinked loci becomes available first and a larger number of linked loci is employed at a later stage.

### DNA extractions and microsatellite genotyping

DNA extractions were carried out as described previously (Lexer *et al.*, 2005), except that the initial step of

tissue disruption of silicagel-dried leaves was automated using a Retsch MM301 mixer mill. Microsatellites were PCR-amplified following Lexer *et al.* (2005), making use of a three-primer PCR to achieve economic fluorescent labeling of PCR products via a labeled 'universal' M13 primer. PCR products were resolved on an AB 3100 automated sequencer (Applied Biosystems) using the fluorescent dyes 6-FAM and JOE as well as molecular size differences for multiplexing. Molecular sizes in base pairs were determined using the GENSCAN-500 ROX (Applied Biosystems) size standard, and microsatellite genotypes were scored by two people independently (CL and JAJ) using GENSCAN and GENOTYPER software (Applied Biosystems, Foster City, CA, USA).

### Genetic data analysis

Microsatellite allele frequency differentials were determined as

$$\delta = \sum_i |f_{i1} - f_{i2}|/2$$

where  $f_{i1}$  and  $f_{i2}$  represent the  $i$ th allele frequencies in the two parental populations, respectively, following Zhu *et al.* (2005). Rare alleles in the hybrid zone and each parental species were identified based on a threshold population frequency of 8%. To test whether samples of each parental species represented one or more panmictic population(s), a Bayesian analysis was carried out with STRUCTURE version 2 (Pritchard *et al.*, 2000), using the admixture model allowing for correlated allele frequencies, a burn-in phase of 50 000 followed by a run length of 100 000. These parameter settings were identified as appropriate based on the diagnostic tools available in STRUCTURE.

Maximum-likelihood estimates (MLEs) as well as upper and lower bounds of the molecular hybrid index

$h$  for each plant from the hybrid zone were calculated from codominant microsatellite data using the computer program HINDEX (Buerkle, 2005). The genomic composition of plants from the hybrid zone was presented previously (Lexer *et al.*, 2005), and the Bayesian genetic structure analysis presented there allowed the identification of pure parental reference populations for studying admixture. Estimation of parental allele frequencies was thus straight-forward, so simple hybrid index MLE's ( $h$ ) rather than Bayesian admixture proportions ( $Q$ ) were chosen for the present study. Although  $Q$  should in theory yield equivalent results, more experience exists with  $h$  in interspecific hybrid zones (Rieseberg *et al.*, 1998, 1999; Buerkle, 2005). The purpose of re-calculating  $h$  here was to create criteria for sorting hybrids into different genotypic classes so that two-locus disequilibria could be computed for each class. The lower (*P. tremula*) bound of  $h$  was used for this purpose. Four genotypic classes with approximately equal sample size of  $N=45 \pm 2$  individuals were defined from the hybrid zone data. Ranked from low hybrid index values (early generation hybrids) to high hybrid index values (advanced generation backcrosses to *P. alba*), the four genotypic classes were defined as: hybrid index (HI) group 1 (0.256–0.771), HI group 2 (0.771–0.854), HI group 3 (0.854–0.908), and HI group 4 (0.908–0.933). Seven apparent  $F_1$  genotypes were omitted from the dataset because the  $F_1$  generation is typically not suitable for admixture mapping.

Two-locus disequilibria were calculated in the form of the standardized LD  $D'$ . Haplotype frequencies for this purpose were estimated from microsatellite genotype data using the Expectation-Maximization (EM) algorithm as implemented in ARLEQUIN (Excoffier and Slatkin, 1995), and  $D'$  was calculated from haplotype data using the HAPLOXT program distributed within the GOLD software package (Abecasis and Cookson,

**Table 1** Microsatellite markers employed in this study, including repeat type, linkage group and map position on the *P. trichocarpa* × *P. deltoides* genetic map, number of alleles/frequency ( $f$ ) of most informative allele in *P. alba* and *P. tremula*, allele frequency differential  $\delta$ , and number of rare alleles ( $f < 8\%$ ) in *P. alba*, *P. tremula* and *P. x canadensis* hybrids, respectively

Locus <sup>a</sup>	Repeat type	Linkage group	Position (cM)	No. of alleles/ $f$ of most informative allele <i>P. alba</i>	No. of alleles/ $f$ of most informative allele <i>P. tremula</i>	$\delta$	No. of rare alleles ( $f < 8\%$ ) <sup>b</sup>
PMGC 2852 <sup>c</sup>	Di	I	157.3	10/0.500	12/0.244	0.514	5/7/15
WMPS 15 <sup>c</sup>	Tri	V	167.7	7/0.286	7/0.014	0.386	4/3/4
ORPM 312 <sup>c</sup>	Tri	VII; V <sup>d</sup>	58.7; 83.8 <sup>d</sup>	9/0.013	7/0.359	0.724	6/4/8
ORPM 344 <sup>c</sup>	Di	X	86.5	4/0.814	5/0.057	0.908	2/2/7
ORPM 206 <sup>c</sup>	Tri	XIX	28.2	2/0.989	4/0.353	0.647	1/1/6
ORPM 127 <sup>c</sup>	Di	IV	49.1	3/0.939	5/0.484	0.461	2/2/4
ORPM 202 <sup>c</sup>	Tri	VIII	140.9	3/0.560	4/0.028	0.639	0/2/3
ORPM 30_1	Di	I	184.2	2/0.955	3/0.776	0.224	1/1/4
ORPM 30_2 <sup>c</sup>	Di	III	53	14/0.000	8/0.316	0.469	10/3/19
ORPM 220 <sup>c</sup>	Tetra	NA <sup>e</sup>	NA <sup>e</sup>	1/1.000	6/0.013	0.987	0/3/4
ORPM 28 <sup>c</sup>	Di	XVIII	61.1	3/0.366	4/0.000	0.403	1/3/2
ORPM 137 <sup>c</sup>	Di	NA <sup>e</sup>	NA <sup>e</sup>	7/0.081	8/0.597	0.836	4/5/4
ORPM 14	Tetra	XVI	14.8	1/0.000	2/0.064	0.064	0/1/1
ORPM 21	Di	IX	30.1	2/0.000	1/0.114	0.114	0/0/2
ORPM 60 <sup>c</sup>	Tri	NA <sup>e</sup>	NA <sup>e</sup>	7/0.182	4/0.897	0.818	4/3/8
ORPM 149 <sup>c</sup>	Di	NA <sup>e</sup>	NA <sup>e</sup>	4/0.000	4/0.595	0.841	1/1/4
ORPM 167 <sup>c</sup>	Di	NA <sup>e</sup>	NA <sup>e</sup>	2/0.000	2/0.988	0.988	1/1/3
ORPM 214 <sup>c</sup>	Di	NA <sup>e</sup>	NA <sup>e</sup>	3/0.000	2/0.847	0.847	0/0/2
WMPS 5 <sup>c</sup>	Di	XII; XV <sup>d</sup>	43; 49.8 <sup>d</sup>	7/0.274	12/0.000	0.917	1/8/10

<sup>a</sup>Markers developed by Tuskan *et al.* (2004b) or Van der Schoot *et al.* (2000) as indicated at [http://www.ornl.gov/sci/ipgc/ssr\\_resource.htm](http://www.ornl.gov/sci/ipgc/ssr_resource.htm).

<sup>b</sup>Given in the order '*P. alba*/*P. tremula*/*P. x canadensis* hybrids', sample sizes were  $N=88$ , 78 and 378 chromosomes, respectively.

<sup>c</sup>Markers included in analyses of introgression frequencies and departures from neutrality.

<sup>d</sup>Duplicated locus mapped on two different linkage groups Yin *et al.*, 2004.

<sup>e</sup>Map position not yet determined, but not in LD with any other marker in *P. tremula*, the species with the larger effective population size ( $N_e$ ).

2000). This method was used to estimate  $D'$  for all pairs of marker loci in each of the four hybrid genotypic classes (HI groups 1–4) and in the parental populations of *P. alba* and *P. tremula*. The calculations were carried out both with and without rare alleles ( $f < 8\%$ ) in the datasets, resulting in a total of 1026 two-locus comparisons. The significance of pairwise disequilibria was tested using the EM algorithm and associated permutation procedure in ARLEQUIN, and significance thresholds were corrected for multiple tests using sequential Bonferroni (Rice, 1989). Descriptive statistics such as means, standard errors, and medians for  $D'$  were calculated and compared in SPSS (SPSS Inc.).

In order to investigate the likely causes for background LD in the two parental populations, variance components of LD were calculated following Ohta (1982) using the computer program LINKDOS (Garniere-Gere and Dillmann, 1992). This method uses Wright's island model to partition the variance in LD into within- and between subpopulation components, in analogy to the partitioning of inbreeding coefficients in subdivided populations. The most intuitive variance components are  $D'_{IS}$ , the variance of LD within subpopulations,  $D'_{ST}$ , the variance component due to genetic drift between subpopulations, and  $D'_{IT}$ , the variance of LD in the total population. Differences between variance components contain information regarding the likely cause(s) for LD, for example, if  $D'_{IS} < D'_{ST}$  then LD may be due to drift and limited migration, and if  $D'_{IS} > D'_{ST}$  then epistatic natural selection is a more likely cause. The hypothesis that  $D'_{IS}$  and  $D'_{ST}$  differed from one another was tested with non-parametric Wilcoxon signed-rank tests using SPSS.

For the analysis of introgression frequencies, 16 marker loci with the greatest information content were chosen ( $\delta > 0.25$ , see Table 1). In order to maximize the accuracy with which clines could be estimated, alleles at each locus were combined into two allelic classes with frequency differences between species ( $\delta$ ) equal to those observed when alleles were utilized separately (see Supplementary material). Allelic classes were created by an exhaustive search of all possible combinations of alleles in two classes (code written in R version 2.1.1; R Development Core Team, 2005). This simple pooling of alleles into classes reduces multiallelic data to a biallelic classification, without a loss of information or distortion of the relationship between parental species. Below we refer to genotypes, which in this case are genotypes of allelic classes rather than of individual microsatellite alleles.

Introgression of *P. tremula* alleles into the *P. alba* background was described using estimated clines of microsatellite genotype frequencies. The more common approach to the analysis of introgression is to calculate clines across a spatial transect (Barton and Gale, 1993). Our approach also differs from human admixture mapping, where locus-specific excess ancestry is used as a predictor variable in a logistic regression with disease status as the response variable (Reich et al., 2005; Zhu et al., 2005). In contrast to these methods, the approach utilized here involves estimating genotype frequencies as a function of hybrid index. For each locus, the allelic class with the highest frequency in *P. tremula* was taken as the focal class and separate clines were estimated for homozygous and heterozygous genotypes that included the allelic class. Multinomial logistic

regression was used to estimate genotype frequencies as a function of hybrid index, with genotypes that did not include the focal allelic class serving as the reference category. Regressions were performed in R (R Development Core Team, 2005), using the R packages *nnet* (package version 7.2–16) and *genetics* (G Warnes and F Leisch, package version 1.1.3).

Given that the hybrid index is an estimate of the genome-wide average frequency of *P. alba* alleles within an individual, it can be utilized to generate expected genotype frequencies under the assumption of neutral introgression. For any given  $h$ , the expected frequency of an allele is given by:  $E(a) = a_{\text{trem}} + (a_{\text{alba}} - a_{\text{trem}}) \cdot h$ , where  $a_{\text{alba}}$  and  $a_{\text{trem}}$  are the allele frequencies in the parental populations of *P. alba* and *P. tremula* (with a diagnostic, diallelic locus, this reduces to  $E(a) = h$ ). The expected frequencies for the genotypes that include a particular allele are given by  $E(a)^2$  for the homozygote and  $2 \cdot E(a) \cdot (1 - E(a))$  for the heterozygote, and the frequency for all other genotypes is expected to be  $(1 - E(a))^2$ . To test for departures from neutrality, the slope and intercept of the regressions for observed data were compared to those obtained in 1000 replicate simulations of genotypic data for each locus. In simulations, genotypes were sampled randomly according to frequencies expected under neutrality, using a population of the same size and with the same distribution of hybrid index as the empirical data. Preliminary analyses indicated that in all simulation replicates, parameter estimates for the regression converged without difficulty.

## Results

### Allele frequency differentials and rare alleles

The 19 microsatellites exhibited large allele frequency differentials between the two parental species, *P. alba* and *P. tremula* (up to 0.909, mean of  $\delta = 0.619 \pm 0.067$  s.e., median = 0.647). There was no obvious relationship between microsatellite repeat type and  $\delta$ , for example, frequency differentials varied from 0.064 to 0.987 for tetra-repeat marker loci and from 0.114 to 0.919 for di-repeat markers. Estimates for  $\delta$  for each marker, along with information about the number of alleles observed in each species and the parental frequencies of the most informative allele for each locus, are given in Table 1.

The number of 'rare alleles' (alleles with population frequencies  $< 8\%$ ) was larger in the hybrid zone compared to samples of the two parental species for most loci (15 out of 19 microsatellites; Table 1). Probabilistic analysis of this result was not intended since sampling strategies differed between the hybrid zone and the two parental species, the two parental gene pools being represented by a smaller number of gene copies sampled over a much larger geographic range. Nevertheless, the high proportion of low frequency alleles present in the hybrid zone indicates that rare alleles must be taken into account when interpreting patterns of LD in admixture mapping studies.

Patterns of background LD in the parental genepools Bayesian genetic structure analysis indicated that the neighboring subpopulations sampled for each of the two parental species, *P. alba* and *P. tremula*, behaved like one single panmictic unit in each case. The natural logarithm

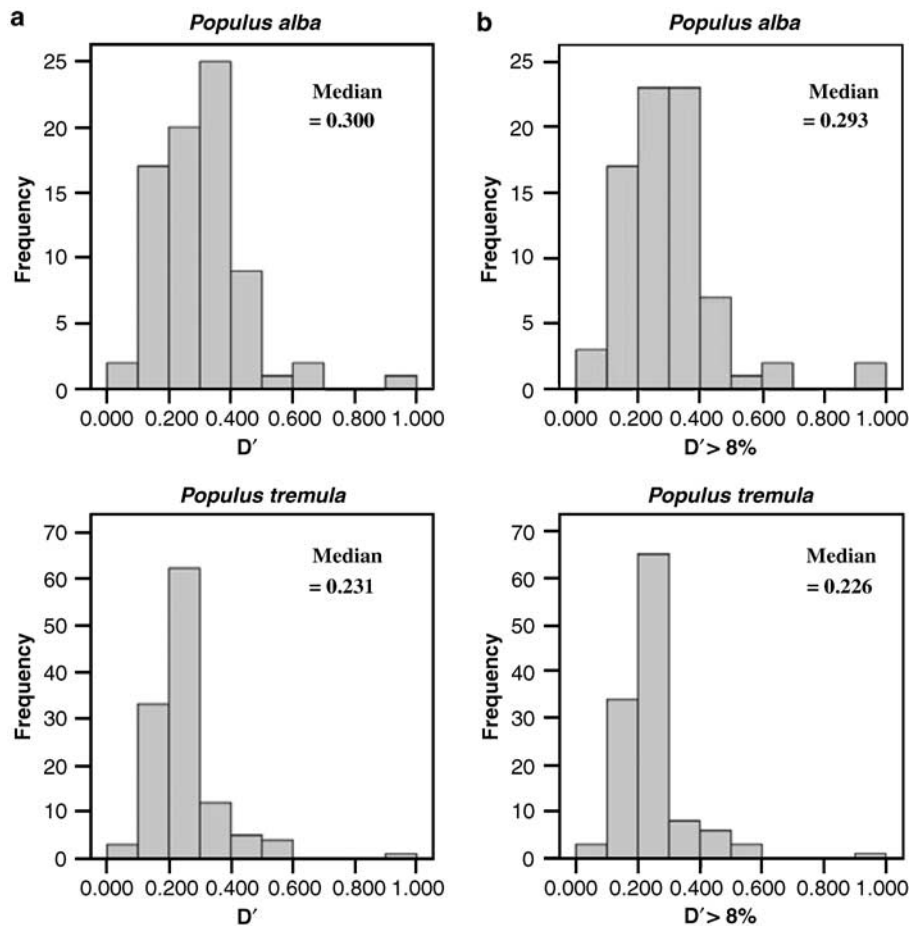
(ln) of the probability of pairs of subpopulations of *P. alba* representing one panmictic unit was  $-1412$ , whereas it was  $-1890$  for a model including two populations for this species. Likewise, the statistics in *P. tremula* were  $-1545$  for the 'one-population' genetic structure model and  $-1614$  for the 'two-population' model, again indicating a single panmictic genepool. These results confirm previous estimates of high levels of gene exchange ( $N_e m$ ) among neighboring subpopulations of each species. Based on the data, subpopulations were combined for analyses of LD in each parental genepool.

Standardized estimates of LD ( $D'$ ) for the two parental genepools were larger for *P. alba* than for *P. tremula* (mean =  $0.300 \pm 0.016$  s.e., median =  $0.300$  for *P. alba*; mean =  $0.251 \pm 0.011$  s.e., median =  $0.231$  for *P. tremula*; Figure 1a). A similar difference between species was observed when rare alleles were excluded from the analysis, but the median of  $D'$  was smaller for both species in this case (Figure 1b). Analysis of the variance components of LD in the parental genepools revealed striking differences between the two species. In *P. alba*,  $D_{ST}^2$ , the variance of LD due to genetic drift between subpopulations was significantly larger than  $D_{IS}^2$ , the variance of LD within subpopulations (Z of Wilcoxon sign rank test =  $-8.043$ ,  $P < 0.005$ ; Figure 2a). This difference was not observed in *P. tremula* (Figure 2b). These

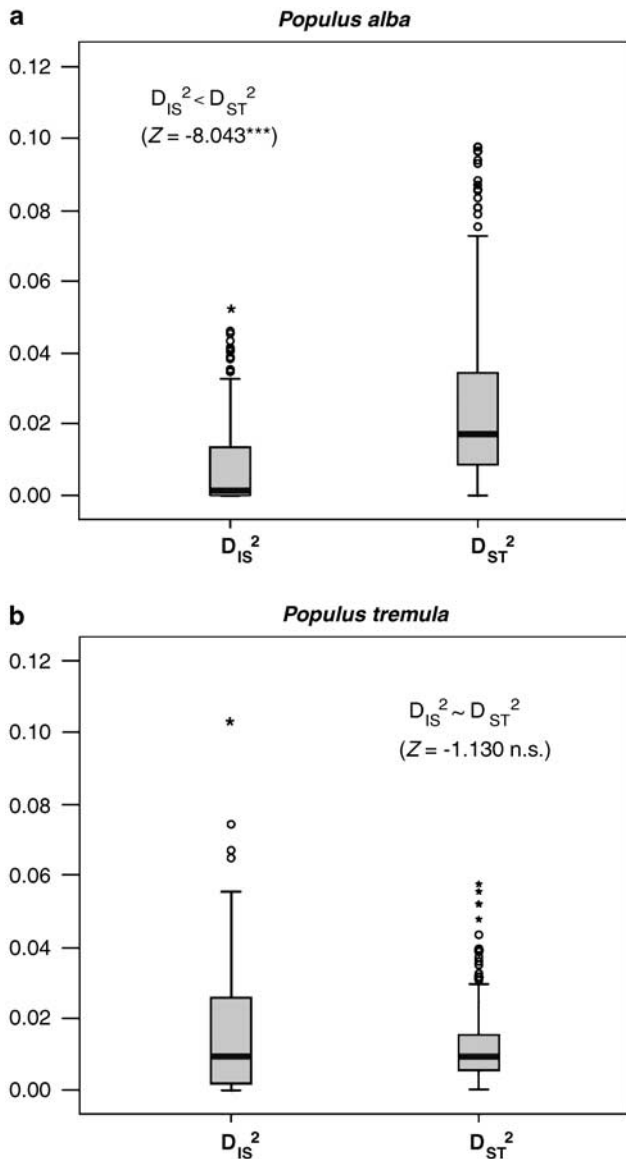
inter-specific differences provide information about the likely causes of background LD in these two divergent genepools.

#### Two-locus disequilibria in different hybrid generations

Linkage disequilibria in four different genotypic classes of hybrids as defined by their molecular hybrid index (HI groups 1–4) were generally larger in early generation hybrids (HI group 1) than in highly introgressed advanced generation backcrosses (HI groups 3 and 4; see Supplementary material). The proportion of significant tests for LD followed the same overall pattern: 15.8% of exact tests between pairs of loci were significant in HI group 1 (early generation hybrids), whereas 5.3, 4.2 and 3.8% were significant in HI groups 2, 3 and 4 (later generation hybrids), respectively. For comparison, only 1.4% of exact tests for LD yielded a significant result in the parental population of *P. alba* and no significant test at all was observed in *P. tremula* (not shown). Note, however, that slight differences in allelic diversities between study groups may lead to differences in the power to detect LD. As expected, the median of  $D'$  in different genotypic classes of hybrids was generally smaller when rare alleles were excluded from the analysis (see Supplementary material). Our data reflect



**Figure 1** Distribution of standardized two-locus disequilibria ( $D'$ ) for 171 comparisons among 19 multi-allelic microsatellite loci in Central European populations of *P. alba* and *P. tremula*. (a) Rare alleles ( $f < 8\%$ ) included. (b) Rare alleles excluded. For overall allele numbers and numbers of rare alleles at each locus in each species see Table 1.



**Figure 2** Distribution of  $D_{IS}^2$ , the within-subpopulation variance component of LD, and  $D_{ST}^2$ , the between-subpopulation variance component due to genetic drift, in Central European populations of *Populus alba* and *P. tremula*. Medians are indicated as thick black lines within boxplots. Z-values and significance levels from non-parametric tests for differences between the two variance components are indicated in each graph. (a) *P. alba*. (b) *P. tremula*.

the decay of LD in successive hybrid generations as predicted from theory (Briscoe *et al.*, 1994; Martinsen *et al.*, 2001; Chapman and Thompson, 2002).

#### Variation in introgression frequencies among DNA microsatellite loci

Sixteen loci that exhibited the greatest level of differentiation between parental taxa ( $\delta > 0.25$ ; Table 1) were selected for introgression analysis. The markers displayed a variety of patterns of parental genotype frequencies and patterns of introgression (Figure 3). Several loci exhibited departures from neutral expectations and can be placed into three categories: a small excess of *P. tremula* genotypes in the hybrid and *P. alba*

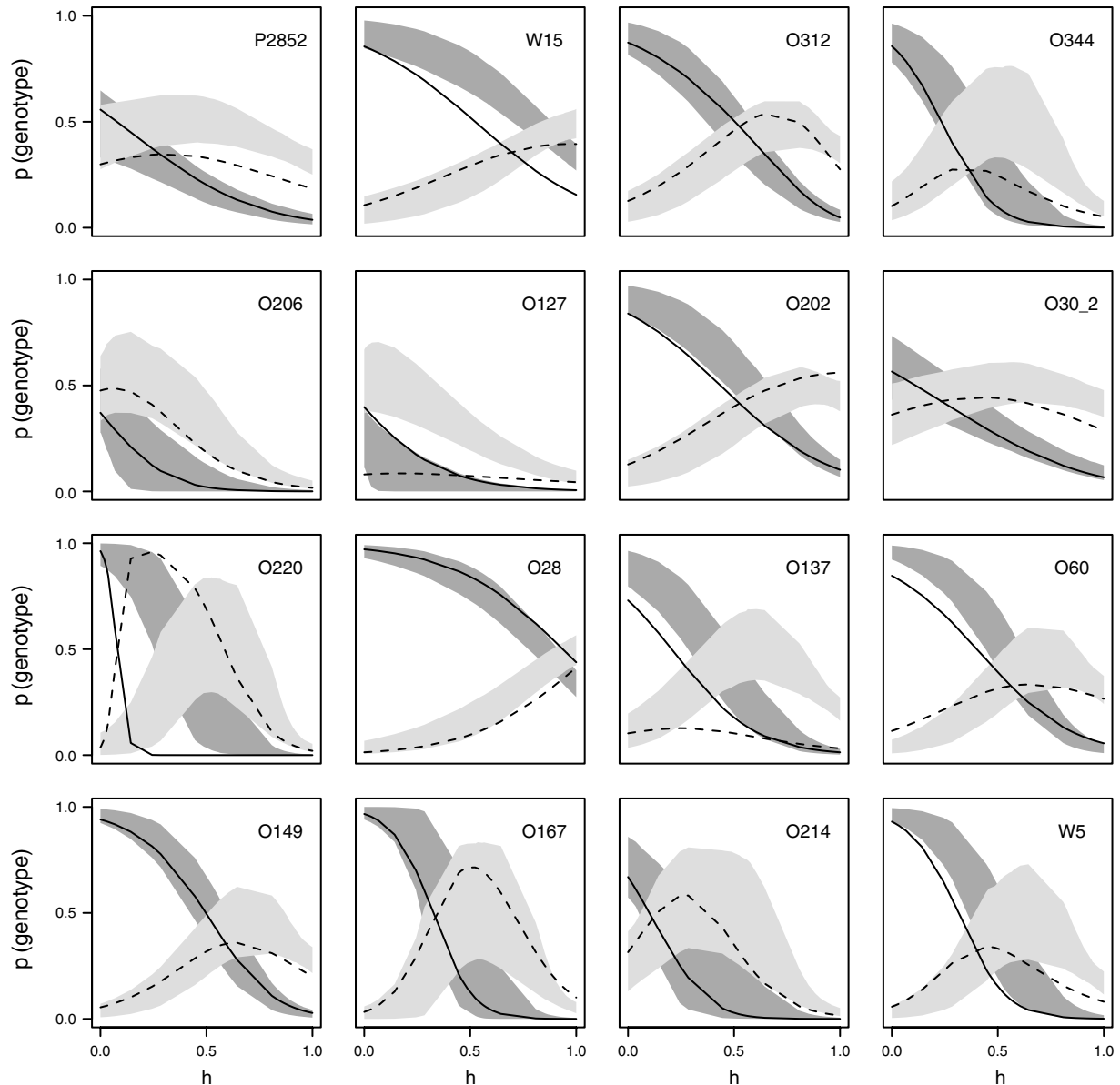
genetic backgrounds, a large deficit of *P. tremula* genotypes, and shifts in the centrality of the cline (primarily toward *P. tremula*; Figure 4; see Supplementary material). Homozygous genotypes at four loci (ORPM 202, ORPM 28, ORPM 137, ORPM 60) and heterozygotes at two loci (ORPM 344, ORPM 127) exhibited introgression frequencies that were slightly higher than neutral expectations (slope of smaller magnitude). Homozygotes at ORPM220 were significantly underrepresented in hybrids, with a slope estimate that was much steeper than any in replicate simulations. The functions for homozygotes at six loci and heterozygotes at four loci had intercepts that were significantly smaller than those in 95% of simulations. For all but one of these loci, a smaller intercept along with the negative slope, means that relative to expectations the curves were shifted toward lower hybrid index and *P. tremula* genotype frequencies (Figures 3 and 4). The exception is ORPM 28, for which the frequency of heterozygotes increases with hybrid index (positive slope), so that the smaller intercept corresponds to a shift in the cline toward *P. alba*.

#### Discussion

Natural hybrid zones are likely to represent extreme cases for admixture genome-scanning compared to admixed human populations: allele frequency differences between divergent populations or species will be larger in most cases (Rieseberg *et al.*, 1999; Rieseberg and Buerkle, 2002), genetic architectures of parental gene pools may differ due to differences in breeding systems, and the excess of rare alleles often found in hybrid zones (Schilthuizen *et al.*, 1999) may contribute to background LD. Nevertheless, the most critical factors for admixture mapping are likely to hold across a wide range of organisms. These include: (1) the magnitude of allele frequency differences between the hybridizing populations, (2) levels of background LD (associations among unlinked markers) in the parental gene pools, (3) patterns of admixture and their effects on LD in the admixed population (Pfaff *et al.*, 2001), (4) the range of variation in marker ancestry (variation in introgression frequencies or cline shapes) across the genome. Here, we have tested these factors one-by-one in a natural hybrid zone between two ecologically divergent members of the ‘model tree’ genus *Populus*, *P. alba* and *P. tremula*. We focused on genomic segments subject to natural selection rather than segments linked to a particular trait. We use our data to assess the feasibility of admixture genome scanning in *Populus* and to discuss the potential and caveats of admixture mapping-related studies in hybrid zones of wild species.

#### Background LD in *P. alba*, *P. tremula*, and a natural hybrid zone

Assessing the strength of background LD is critical to genetic association studies because the strength of LD in the target gene pool will dictate both the power and error of association tests (Lynch and Walsh, 1998). Knowing the strength of background LD is especially important in ‘admixture mapping’ studies. LD induced by admixture will decay with each generation following the initial hybridization event (Chakraborty and Weiss, 1988; Briscoe *et al.*, 1994; Pfaff *et al.*, 2001) as a function of

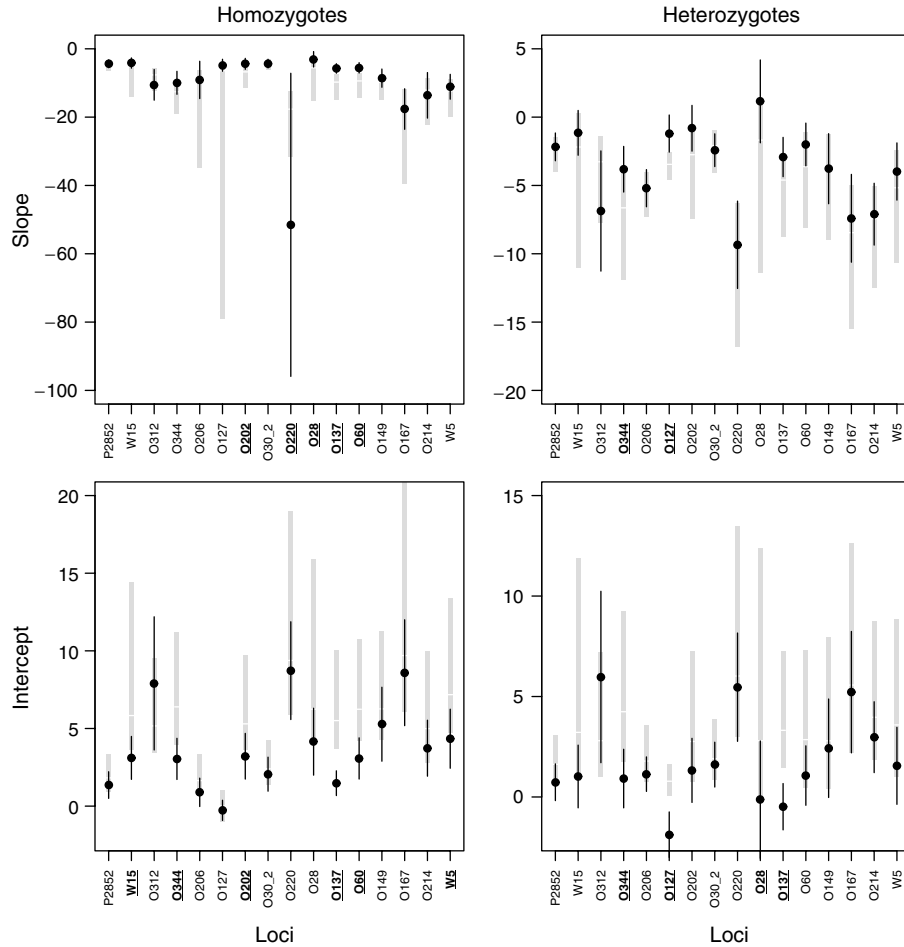


**Figure 3** The introgression of the predominant *P. tremula* allelic class (black solid lines – homozygotes, black dashed lines – heterozygotes) from individuals with low hybrid index (*P. tremula*) to those with high hybrid index (*P. alba*) varies among the sixteen most informative loci. Probabilities of observing homozygotes and heterozygotes are derived from multinomial logistic regressions. Gray regions indicate the 95% confidence envelope for logistic regressions fit to simulated data under the assumption of neutral introgression (dark gray – homozygotes, light gray – heterozygotes). Departure of the observed introgression rates from neutrality is evident graphically when fitted lines deviate from the confidence envelopes.

the recombination rate ( $\theta$ ) (Briscoe *et al.*, 1994), which is proportional to genetic map distance in an idealized situation. However, recombination rates may vary across the target genome, and numerous other factors may affect LD in addition to  $\theta$ , most notably population structure (Lynch and Walsh, 1998; Flint-Garcia *et al.*, 2003). Hence, it is important to estimate the strength of background LD in parental and admixed populations before carrying out a full-blown genome-scan experiment. Whereas it would be desirable to include markers with different degrees of linkage (different physical distances along the chromosomes) in studies of LD, the necessary high-density genotypic data for linked markers are usually not available until an admixture genome

scan has been completed. Thus, the present study utilized unlinked markers to obtain first data on background LD in two *Populus* species and their hybrids.

As expected based on the outbreeding mating system of our study taxa (wind-dispersed pollen and seeds), LD among unlinked markers was generally small to moderate in both species (Figure 1). Levels of LD were somewhat stronger in *P. alba* than in *P. tremula* (Figure 1a), in concordance with smaller effective population sizes in the former species (*P. alba*:  $N_e$  c. 500–550; *P. tremula*:  $N_e$  c. 550–700; Lexer *et al.*, 2005). The variance component of LD due to genetic drift between subpopulations ( $D_{ST}^2$ ) was significantly larger than the within-subpopulation component ( $D_{IS}^2$ ) in *P. alba*, but not in *P. tremula* (Figure 2).



**Figure 4** Estimates of the slope or intercept (black circles with 95% confidence intervals as black bars) from multinomial logistic regressions for several loci (underlined) fall outside of the 95% confidence intervals (CI) derived from 1000 replicate simulations of neutral introgression at each locus (gray boxes, median indicated by white line). Black vertical lines indicate 95% CIs for parameter estimates from observed data.

This indicates that, although no genetic structure was detectable among subpopulations with conventional and Bayesian-based methods, the patchy and fragmented distribution of the flood-plain pioneer *P. alba* may have allowed the build-up of small to moderate levels of LD due to drift, whereas the more continuous distribution of the mixed-forest-species *P. tremula* appears to have resulted in linkage equilibrium among independent markers. Our results indicate that levels of background LD in the two parental gene pools are low to moderate, and that most of the observed LD is due to drift rather than epistatic selection (Ohta, 1982).

An even more relevant question is whether LD in hybrids is conducive to admixture mapping experiments. Our results from a hybrid zone between *P. alba* and *P. tremula* in Central Europe indicate that LD among independent marker loci is pronounced in early generation hybrids (median = 0.303 with rare alleles included in the dataset; 15.8% of tests significant at an experiment-wide level), and that LD decays markedly in later hybrid generations (see Supplementary material). In the most advanced backcross generations (hybrid index group 4), the median of LD is approximately 0.25, which is within the range of the two parental species. Of course, this decrease in LD with progressive hybrid generations is

expected based on simulation studies of LD as a function of generations since admixture (Briscoe *et al.*, 1994), and studies that model genomic block size in hybrid populations (Martinsen *et al.*, 2001; Chapman and Thompson, 2002). Also, pollen records for warmth-loving tree species like *P. alba* (Huntley and Birks, 1983) suggest that the studied hybrid zone should already be 100–200 tree generations old, and thus LD should approach that of the pure parental gene pools for a substantial proportion of genotypes sampled in nature.

A regression-based method for detecting departures from neutral introgression at multi-allelic codominant markers. Admixture mapping as applied to humans has focused on genomic regions with an excess of ancestry in the direction of either parental population, for example, marker-location specific excess ancestry in cases and controls (Pfaff *et al.*, 2001; Zhu *et al.*, 2005). Similarly, the few existing admixture mapping-related studies in non-human taxa have focused on regions of the genome that may introgress more or less frequently than expected under neutrality (Rieseberg *et al.*, 1999; Rieseberg and Buerkle, 2002). In either case, methods are needed to estimate ancestry at marker loci against the remainder of

the genome and across a range of different genetic backgrounds, and these methods need to be applicable to a wide range of markers, population structures, and sampling schemes.

Excellent analytical tools for the genetic analysis of admixed populations have been created by evolutionary biologists studying 'clines' in hybrid zones (Barton and Hewitt, 1985; Barton and Gale, 1993). However, cline parameters are often difficult to measure where habitats are discontinuous and hybrid zones are patchy, as is the case for 'mosaic hybrid zones' (Harrison, 1990). Here, we have explored an alternative, multinomial regression-based approach. Our method uses models of the genotypes of *individuals* along a cline, as opposed to clinal models of pooled genotypes (i.e., genotype frequencies in populations) at a geographic location (Barton and Gale, 1993). Of course, variability in cline shape may also arise due to genealogical/spatial variation among neutral loci (Ibrahim *et al.*, 1996; Klopstein *et al.*, 2006). In our approach, genealogical variation among loci is encompassed by computer simulations under the null hypothesis of neutrality. We estimated genotype frequencies at microsatellite loci as a function of the overall genomic composition of individual plants. We utilized logistic regression as in human admixture mapping (McKeigue *et al.*, 2000; Hoggart *et al.*, 2004; Reich *et al.*, 2005; Zhu *et al.*, 2005). In contrast to human studies, however, we did *not* use locus-specific excess ancestry as a predictor variable for a phenotypic trait such as disease status. Rather, we examined genotypes at individual marker loci as a function of genome-wide admixture, that is hybrid index. Computer simulations for each locus provided confidence intervals for *expected* introgression across the full range of genetic backgrounds (estimated via the hybrid index) and were compared to *observed* introgression for each locus. Our approach is thus equivalent to a genomic scan for chromosomal segments that move across the species barrier more or less frequently than expected under neutrality, which is a topic of major interest in evolutionary biology (Wu, 2001).

Application of this approach to our data revealed large variation in introgression frequencies among individual marker loci (Figure 3), and the introgression patterns of several loci deviated from neutral expectations (see departures of point estimates from simulation envelopes in Figure 4). This included one locus for which homozygotes for *P. tremula* alleles are much rarer than expected in hybrids (ORPM 220, Figure 4). In several cases (e.g., ORPM 220 and 127), departures from neutral expectations were evident for homozygotes, but not heterozygotes, or *vice versa* (Figure 4). The differences may have a biological explanation, or may result from sample size and distributional differences between homozygotes and heterozygotes, and differences in the power to estimate functions with narrow confidence limits.

With respect to biological interpretations, it would be premature to interpret departures from neutrality in terms of linked candidate genes before a genome-wide marker scan has been completed. However, three conclusions may be drawn from our current data: Firstly, different regions of the *P. tremula* genome do introgress into *P. alba* at different rates, and these differences can be detected with the methods devised here. Secondly, alleles

from *P. tremula* may sometimes introgress into *P. alba* at a rate slightly higher than that expected under neutrality, as exemplified by homozygotes at locus ORPM 202 and 60 (Figure 4). Much of speciation genetics has been focused on genetic factors that are negatively selected in hybrids (Dobzhansky, 1937; Butlin, 1989), but recent conceptual and empirical work suggests that chromosomal blocks within recombinant hybrids may sometimes experience positive selection (Barton, 2001; Rieseberg *et al.*, 2003; Seehausen, 2004). Our results suggest that it will be feasible to detect positive selection on individual genomic blocks in hybrid zones of *P. alba* and *P. tremula*, if present, through the use of multi-allelic codominant markers. Thirdly, our approach can detect significant under-representation of *P. tremula* genotypes in hybrids and in the *P. alba* genetic background (ORPM 220; Figure 4). In addition, by examining genotype rather than allele frequencies, the regression method can reveal significant contrasts between the introgression of homozygotes and heterozygotes (e.g. ORPM 220), which may be due to dominance relationships among alleles at linked loci. One interpretation of the difference between homozygotes and heterozygotes at ORPM 220 is that *P. tremula* alleles linked to the marker are recessive to the *P. alba* alleles and experience negative selection in hybrids.

#### Outlook for 'admixture mapping' of adaptive or detrimental variation in *Populus*

Hybrid zones between *P. alba* and *P. tremula* meet the most critical requirements for admixture-based genetic analyses: microsatellite allele frequency differentials ( $\delta$ ) at most loci are substantial (Table 1;  $\delta > 0.3$  is needed for admixture mapping), background LD in the parental gene pools is low to moderate (Figure 1) and should respond to sampling schemes that minimize drift and account for rare alleles, and LD in hybrids decays rapidly with increasing hybrid index, i.e., increasing number of backcross generations. The latter observation is of special interest.

We have shown previously (Lexer *et al.*, 2005) that both early and late generation hybrids are present in this Central European hybrid zone. Hence, preferential sampling of recombinant *early generation hybrids* from the population would provide us with a sample of genotypes suitable for low-resolution mapping. This could, for example, include 80–100 markers to conduct a first admixture genome-scan at 15–20 cM intervals, based on studies of genomic block size in hybrid populations (Briscoe *et al.*, 1994; Martinsen *et al.*, 2001; Pfaff *et al.*, 2001), age estimates for the hybrid zone based on fossil records for temperate trees (Huntley and Birks, 1983), and genome-length estimates for *Populus* (Bradshaw and Stettler, 1994; Frewen *et al.*, 2000; Cervera *et al.*, 2001).

In contrast, preferential sampling of advanced generation hybrids should provide us with material for high-resolution analyses. We note that ancestry from the donor genome (*P. tremula*) in the studied population is roughly 20% if advanced generations are included (Lexer and co-workers, unpublished data), which is comparable to admixture proportions utilized in human genetics (Zhu *et al.*, 2005). Given the size of the *Populus* genome ( $2n = 38$ ; 550 Mb;  $2C = 1.1$  pg), however, high-resolution

mapping experiments are likely to require hundreds if not thousands of markers. This limitation can probably not be overcome with microsatellite loci (roughly 4000 markers are available for *P. trichocarpa*, but only a fraction of these cross-amplify in *P. alba* and *P. tremula*). It is therefore likely that high-density SNP (Single Nucleotide Polymorphism) data will be required for high resolution 'admixture mapping' in *Populus*.

Admixture LD among highly informative codominant markers in *Populus* hybrid zones should permit the precise estimation of marker ancestry in hybrids in a genomic context. This should make it possible to estimate the selective value (fitness effects) of individual chromosome blocks in admixed populations. Hybrid populations in other European river valleys may serve as independent 'replicates'. Admixture should also allow the detection of associations among markers and quantitative phenotypic traits or ecological habitat factors. This would be of great interest not only for evolutionary biology but also for applied breeding programs, and first attempts to model the dependence of quantitative traits upon individual admixture (McKeigue et al., 2000; Hoggart et al., 2004) are encouraging.

## Acknowledgements

We thank Hans Herz and Wilfried Nebenfuehr of BFW Vienna, Austria, Christian Fraissl and Franz Kovacs of the Danube Floodplain National Park, the Forstverwaltung Lobau of the Vienna City Council, Herbert Tiefenbacher and several other private landowners for support during field work in Austria, Marius-Sorin Nica for field work in Romania, Loren Rieseberg, Keith Gardner, and Jon Slate for valuable discussions during various stages of the work, and Robyn Cowan for help in the lab. This work was supported by grants of the Natural Environment Research Council (NERC; grant NE/C507037/1) and the Royal Society (grant round 2004/R1) to CL.

## References

- Abecasis G, Cookson W (2000). GOLD – graphical overview of linkage disequilibrium. *Bioinformatics* **16**: 182–183.
- Adler W, Oswald K, Fischer R (1994). *Exkursionsflora von Oesterreich*. Verlag Eugen Ulmer: Stuttgart und Wien.
- Barton N (2001). The role of hybridization in evolution. *Mol Ecol* **10**: 551–568.
- Barton N, Gale K (1993). Genetic analysis of hybrid zones. In: Harrison RG (ed). *Hybrid Zones and the Evolutionary Process*. Oxford University Press: Oxford. pp 13–45.
- Barton N, Hewitt G (1985). Analysis of hybrid zones. *Ann Rev Ecol Syst* **16**: 113–148.
- Bradshaw HD, Stettler RF (1994). Molecular genetics of growth and development in *Populus*. II. Segregation distortion due to genetic load. *Theor Appl Genet* **89**: 551–558.
- Briscoe D, Stephens J, O'Brien S (1994). Linkage disequilibrium in admixed populations: applications in gene mapping. *J Hered* **85**: 59–63.
- Buerkle CA (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Mol Ecol Notes* **5**: 684–687.
- Butlin R (1989). Reinforcement of premating isolation. In: Otte D, Endler J (eds). *Speciation and its Consequences*. Sinauer Associates: Sunderland, MA. pp 158–179.
- Cervera M-T, Storme V, Ivens B, Gusmao J, Liu BH, Hostyn V 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.
- Chakraborty R, Weiss K (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* **85**: 9119–9123.
- Chapman N, Thompson E (2002). The effect of population history on the lengths of ancestral chromosome lengths. *Genetics* **162**: 449–458.
- Cronk Q (2005). Plant eco-devo: the potential of poplar as a model organism. *New Phytologist* **166**: 39–48.
- Dobzhansky T (1937). *Genetics and the origin of species*. Columbia University Press: New York.
- Excoffier L, Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921–927.
- Fisher R (1930). *The genetical theory of natural selection*. Oxford University Press: Oxford.
- Flint-Garcia S, Thornsberry J, Buckler E (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357–374.
- Floate K (2004). Extent and patterns of hybridisation among the three species of *Populus* that constitute the riparian forest of southern Alberta, Canada. *Can J Bot* **82**: 253–264.
- Fossati T, Patrignani G, Zapelli I, Sabatti M, Sala F, Castiglione S (2004). Development of molecular markers to assess the level of introgression of *P. tremula* into *P. alba* natural populations. *Plant Breeding* **123**: 382–385.
- Frewen BE, Chen THH, Howe GT, Davis J, Rohde A, Boerjan W et al. (2000). Quantitative trait loci and candidate gene mapping of bud set and bud flush in *Populus*. *Genetics* **154**: 837–845.
- Garniere-Gere P, Dillmann C (1992). A computer program for testing pairwise linkage disequilibria in subdivided populations. *J Hered* **83**: 239.
- Harrison R (1990). Hybrid zones: windows on evolutionary processes. *Oxford Sur Evol Biol* **7**: 69–128.
- Hoggart C, Shriver M, Kittles R, Clayton D, McKeigue P (2004). Design and analysis of admixture mapping studies. *Am J Hum Genet* **74**: 965–978.
- Huntley B, Birks H (1983). *An atlas of past and present pollen maps of Europe: 0-13 000 years ago*. Cambridge University Press: Cambridge.
- Ibrahim KM, Nichols RA, Hewitt GM (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* **77**: 282–291.
- Ingvarsson PK (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**: 945–953.
- Karrenberg S, Edwards P, Kollmann J (2002). The life history of Salicaceae living in the active zone of floodplains. *Freshwater Biol* **47**: 733–748.
- Klopfstein S, Currat M, Excoffier L (2006). The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol* **23**: 482–490.
- Lexer C, Fay M, Joseph J, Nica M-S, Heinze B (2005). Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. *Mol Ecol* **14**: 1045–1057.
- Lynch M, Walsh B (1998). *Genetics and analysis of quantitative traits*. Sinauer Associates: Sunderland, MA.
- Martinsen G, Whitham T, Turek R, Keim P (2001). Hybrid populations selectively filter gene introgression between species. *Evolution* **55**: 1325–1335.
- McKeigue P, Carpenter J, Parra E, Shriver M (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* **64**: 171–186.

- Morjan C, Rieseberg L (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol* **13**: 1341–1356.
- Ohta T (1982). Linkage disequilibrium with the island model. *Genetics* **101**: 139–155.
- Orr H (2001). The genetics of species differences. *Trends Ecol Evol* **16**: 343–350.
- Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI *et al.* (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet* **68**: 198–207.
- Pritchard J, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Rajora O, Dancik B (1992). Genetic characterization and relationships of *Populus alba*, *P. tremula*, and *P. x canescens*, and their clones. *Theor Appl Genet* **84**: 291–298.
- Reich D, Patterson N, De Jager PL, McDonald GJ, Waliszewska A, Tandon A *et al.* (2005). A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet* **37**: 1113–1118.
- Rice W (1989). Analyzing tables of statistical tests. *Evolution* **43**: 223–225.
- Rieseberg L, Buerkle C (2002). Genetic mapping in hybrid zones. *Am Nat* **159**: S37–S49.
- Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T *et al.* (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**: 1211–1216.
- Rieseberg L, Whitton J, Gardner K (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**: 713–727.
- Rieseberg LH, Baird SJE, Desrochers AM (1998). Patterns of mating in wild sunflower hybrid zones. *Evolution* **52**: 713–726.
- Schilthuizen M, Hoekstra R, Gittenberger E (1999). Selective increase of a rare haplotype in a land snail hybrid zone. *Proc R Soc Lond B* **266**: 2181–2185.
- Seehausen O (2004). Hybridization and adaptive radiation. *Trends Ecol Evol* **19**: 198–207.
- Taylor G (2002). *Populus: Arabidopsis* for forestry. Do we need a model tree? *Ann Bot – London* **90**: 681–689.
- Tuskan G, Gunter L, Yang Z, Yin T, Sewell M (2004). Microsatellite discovery and genetic mapping in *Populus trichocarpa* × *deltoides* hybrids. *Can J For Res* **34**: 85–93.
- van der Schoot J, Pospiskova M, Vosman B, Smulders M (2000). Development and characterisation of microsatellite markers in black poplar (*Populus nigra* L.). *Theor Appl Genet* **101**: 317–322.
- Whitham T (1989). Plant hybrid zones as sinks for pests. *Science* **244**: 1490–1493.
- Wright S (1931). Evolution in Mendelian populations. *Genetics* **16**: 97–159.
- Wu C-I (2001). The genic view of the process of speciation. *J Evol Biol* **14**: 851–865.
- Yin T, DiFazio S, Gunter L, Riemenschneider D, Tuskan G (2004). Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theor Appl Genet* **109**: 451–463.
- Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T *et al.* (2005). Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet* **37**: 177–181.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)