Heredity (2010), 1–10

www.nature.com/hdy

ORIGINAL ARTICLE

Population substructure in continuous and fragmented stands of *Populus trichocarpa*

GT Slavov[1,2], S Leonardi[3], WT Adams[4], SH Strauss[4] and SP DiFazio[1,2]

[1]*Department of Biology, West Virginia University, Morgantown, WV, USA;* [2]*U.S. Department of Energy BioEnergy Science Center, Oak Ridge, TN, USA;* [3]*Dipartimento di Scienze Ambientali, Università di Parma, Parma, Italy and* [4]*Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR, USA*

Population substructure has important implications for both basic and applied genetic research. We used 10 microsatellite markers to characterize population substructure in two ecologically and demographically contrasting populations of the model tree *Populus trichocarpa*. The Marchel site was a continuous stand growing in a mesic habitat in western Oregon, whereas the Vinson site consisted of three disjunct and isolated stands in the high desert of eastern Oregon. A previous study revealed that pollen-mediated gene flow is extensive in both populations. Surprisingly, model-based clustering, principal components analysis and analyses of molecular variance provided overwhelming support for the existence of at least two intermingled subpopulations within the continuous Marchel population ($F_{ST} = 0.026$, $P < 0.001$), which occupied an area with a radius of only about 250 m. Genets in these two subpopulations appeared to have different relative clone ages and phenologies, leading us to hypothesize that they correspond to different seedling cohorts, each established from seeds produced by relatively few mothers. As expected, substructure was stronger in the fragmented Vinson population ($F_{ST} = 0.071$, $P = 0.001$), and this difference appeared to result from the more extensive family structure in this population. Using group-likelihood methods, we reconstructed multiple interconnected half-sib families in the Vinson population, with some genets having as many as eight putative siblings. Researchers involved in ongoing and future association studies in *P. trichocarpa* should account for the likely presence of subtle but practically significant substructure in populations throughout the range of this species.
Heredity advance online publication, 9 June 2010; doi:10.1038/hdy.2010.73

Keywords: *Populus trichocarpa*; microsatellites; clonality; population substructure; relatedness; seedling establishment

Introduction

Studies of putatively neutral genetic variation indicate that populations of wind-pollinated forest trees have large effective sizes and relatively weak geographic structure, presumably because of extensive pollen-mediated gene flow (Hamrick *et al*., 1992; Burczyk *et al*., 2004; Smouse and Sork, 2004; Sork and Smouse, 2006). Aspens, poplars and cottonwoods from the genus *Populus* are no exceptions to this trend, and a number of studies have documented their high within-population genetic diversity and low inter-population differentiation for allozyme, restriction fragment length polymorphism, microsatellite and single-nucleotide polymorphism markers (reviewed by Ingvarsson, 2010; Slavov and Zhelev, 2010). Despite the presumed high levels of gene flow, however, population genetic studies in *Populus* often reveal deficiencies of heterozygotes relative to genotype frequencies expected under Hardy–Weinberg equilibrium (reviewed by Slavov and Zhelev, 2010). Although null alleles and other allele detection problems could be driving this pattern (Langley *et al*., 1981; Ewen *et al*., 2000; Dakin and Avise, 2004; Pompanon *et al*., 2005), the life history characteristics of some *Populus* species make the existence of undetected population substructure possible, even at small spatial scales.

Black cottonwoods (*Populus trichocarpa* Torr. & Gray) are typically dioecious, although hermaphroditic trees have been reported (Slavov *et al*., 2009). They typically reach reproductive maturity in 7–10 years, flower in the spring (March to May) and release abundant crops of light seeds with cottony hairs 1–3 months after pollination (DeBell, 1990; Braatne *et al*., 1996). Both pollen and seeds are dispersed by wind and can potentially travel very long distances (Braatne *et al*., 1996; Steinberg, 2001; Slavov *et al*., 2009). Seed crops are frequent and abundant, but successful recruitment (that is, establishment and initial survival) of cottonwood seedlings depends on the availability of recently disturbed and sufficiently moist, alluvial microsites that are free from competing vegetation (Braatne *et al*., 1996; Rood *et al*., 2007). Furthermore, because seeds lose viability 1–2 weeks after dispersal, and because seedlings are extremely drought intolerant, survival is possible only if seed dispersal is synchronized with the recession of the water table after spring floods. Therefore, even under natural hydrologic and hydrogeomorphic conditions, massive seedling recruitment events occur at intervals of 5–10 years or more, with flood control further reducing this frequency (Braatne *et al*., 1996, 2007; Fierke and Kauffman, 2005). *P. trichocarpa* regenerates asexually by root sprouting and the rooting of shoots from broken

Correspondence: Dr GT Slavov, Department of Biology, West Virginia University, 53 Campus Dr, Morgantown, WV 26506, USA.
E-mail: gancho.slavov@mail.wvu.edu
Received 6 December 2009; revised 1 March 2010; accepted 5 May 2010

branches or entire tree trunks that have been toppled during storms and floods and then buried in sediment (Braatne *et al*., 1996; Rood *et al*., 2003, 2007). Thus, black cottonwood stands are characterized by sporadic episodes of seedling establishment after major disturbances, high mortality in the first few growing seasons and expansion of the surviving genets through vegetative propagation. Because of the high phenological variation within populations of *Populus* (Howe *et al*., 2003) and the narrow windows of opportunity for seedling establishment (discussed above), a relatively small number of parents may be contributing to each seedling cohort. Therefore, substructure may be present in populations sampled from relatively small areas because of the periodic establishment of cohorts that are genetically differentiated due to founder effects.

Because of their ecological and economic importance and well-developed propagation protocols and molecular and bioinformatic resources, species of *Populus* have become models for tree genetics and genomics (Jansson and Douglas, 2007; Jansson *et al*., 2010). Two large-scale association genetic studies using range-wide samples of *P. trichocarpa* clones are currently in progress in Canada (C Douglas, personal communication) and the United States (G Tuskan, personal communication). Detailed understanding of the population genetic structure of *P. trichocarpa* is crucially important to correctly interpreting the results of these studies (Balding, 2006; McCarthy *et al*., 2008; Ioannidis *et al*., 2009). Our goal was to characterize the fine-scale genetic structure of two *P. trichocarpa* populations growing in dramatically different habitats. More specifically, we demonstrate that (1) population substructure can be detected even when sampling is performed at small spatial scales (several hundred meters) and (2) the life history of the species and the ecological and demographic characteristics of the sampled populations provide plausible explanations for the existence of population substructure.

Materials and methods

Study sites, plant materials and microsatellite genotyping

The two study sites, Marchel and Vinson, were selected on the basis of their contrasting ecological and demographic characteristics, and were used in a previous study of pollen flow (Slavov *et al*., 2009). The Marchel site is located within a large continuous stand growing in a mesic habitat on the Willamette River, near Corvallis, Oregon (Supplementary Figure S1). We sampled leaf tissue from 292 reproductively mature trees, including all major sources of pollen, in an area with a radius of about 250 m (Slavov *et al*., 2009). After accounting for clonality based on data for 10 microsatellite loci (see below), the number of genets was 184 (177 males and 7 females, Figure 1a). The Vinson site is located in the high desert east of the Cascade Mountains, southwest of Pendleton, Oregon (Supplementary Figure S2) and is considerably warmer and drier than the Marchel site (Slavov *et al*., 2009). Tissue samples were collected from 411 trees, including all reproductively mature males, that formed 24 disjunct stands in an area with a radius of more than 10 km (Figure 3 in Slavov *et al*., 2009). We performed analyses of population substructure on three stands (Rus, Bul and San), which consisted of 97 trees (19 male and 12 female genets) located within an area with a radius of about 1 km (Figure 1b). These stands were chosen because they had the highest number of genets within an area that was roughly comparable in size with the Marchel site.

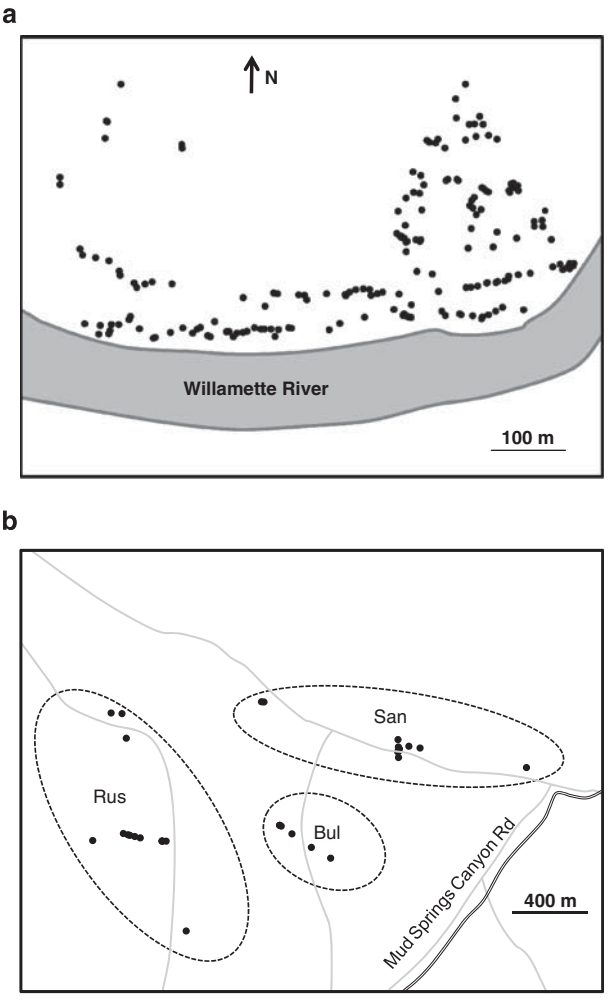

Figure 1 Schematic diagrams of the Marchel (**a**) and Vinson (**b**) study sites. Filled circles indicate the relative positions of the sampled genets (that is, clones with multiple ramets are represented by single circles, whose central positions were calculated as the averages of the Cartesian coordinates of the respective ramets). Rivers and creek beds are shown in light shading, and roads are designated by double lines. The three stands within the Vinson population are delineated with dashed lines.

Tree locations were mapped as described by Slavov *et al*. (2009). Flowering phenology was rated according to a numerical scale representing different stages of anthesis (DiFazio, 2002). The stem diameter of each sampled ramet was measured at breast height using a diameter tape. DNA isolation and genotyping for 10 microsatellite loci were performed as described by Slavov *et al*. (2009).

Data analysis

Clonality, microsatellite diversity and deviations from Hardy–Weinberg equilibrium: The multilocus genotypes of all trees in each population were compared and clonality was inferred using a stepwise procedure. First, individual trees were clustered using the Unweighted Pair

Group Method with Arithmetic Mean based on the sum of the squared differences between the observed allele sizes of their multilocus genotypes (Slatkin, 1995), and clones were identified on the basis of the visual inspection of the resulting dendrogram. Second, the multilocus genotypes of the clones inferred in the first step were analyzed using the Identity Analysis option of v. 3.0.3 of the CERVUS program (Kalinowski *et al*., 2007). Putative clones whose genotypes mismatched at fewer than two loci were merged into a single clone, and the consensus genotype was assigned to that clone. All subsequent analyses were performed using the multilocus genotypes of all inferred genets in each population. Clonal diversity was quantified as $R=(G-1)/(N-1)$, where G is the number of genets and N is the number of ramets sampled (Arnaud-Haond *et al*., 2007).

The number of alleles per locus, allelic richness, observed heterozygosities (H_o) and Nei's unbiased estimates of gene diversity (H_e; Nei, 1978) were calculated using v. 4.05 of the MSA program (Dieringer and Schlötterer, 2003). Allelic richness ($r(10)$) was measured as the number of different alleles that would have been detected if only 10 alleles (that is, five genets) had been sampled in each population (El Mousadik and Petit, 1996; Petit *et al*., 1998).

Exact tests of Hardy–Weinberg equilibrium for each locus in each population were performed using the default parameters of v. 3.11 of the ARLEQUIN program (Excoffier *et al*., 2005). The locus-by-locus analysis of molecular variance option of this program was also used to calculate estimates of the fixation index F_{IS} (Wright, 1965) and to test their statistical significance based on 1000 permutations of alleles among individuals within a population. Global exact tests of Hardy–Weinberg equilibrium across all loci (with the alternative hypothesis of heterozygote deficiency) and exact tests of genotypic disequilibrium for each pair of loci in each population were performed using the default parameters of v. 4.0 of the GENEPOP program (Rousset, 2008).

Population substructure: We used several analytical approaches to test for the presence of population substructure. First, we assessed the extent of spatial genetic structure (SGS) in each population using the approach of Vekemans and Hardy (2004). Briefly, the extent of SGS was assessed by calculating the standardized statistic:

$$Sp=\frac{-b_F}{(1-F_{(1)})} \quad (1)$$

where b_F is the slope of the simple linear regression of the pairwise kinship coefficients (Loiselle *et al*., 1995) for all pairs of genets on the natural logarithm of the physical distances separating pairs of genets, and $F_{(1)}$ is the mean kinship coefficient of genets from the first distance class (0–10, 0–20 or 0–50 m). The Sp statistic can be used to compare the extent of SGS among studies with different sampling designs (Vekemans and Hardy, 2004). The statistical significance of $F_{(1)}$ and b_F was tested based on 1000 permutations of alleles, individuals and individual locations. All analyses of SGS were performed using version 1.2 of the SPAGeDi program (Hardy and Vekemans, 2002).

Second, we used the model-based clustering algorithm implemented in v. 2.2 of the STRUCTURE program (Pritchard *et al*., 2000; Falush *et al*., 2003, 2007) and the empirical statistic ΔK (Evanno *et al*., 2005) to determine the number of sub-populations (K) in each study population. For each population, we ran STRUCTURE using the default model parameters and varying K from one to five. Each run consisted of 50 000 burn-in iterations and 100 000 data collection iterations, and was replicated 20 times. This set of runs produced results that were consistent with those from a smaller set of longer runs (that is, up to 500 000 burn-in iterations and 1 000 000 data collection iterations) and runs allowing for values of K up to 20.

Third, we used v. 3.0 of the EIGENSOFT program to apply the individual-based principal components analysis approach of Patterson *et al*. (2006), which allows formal significance testing for the presence of population substructure. As recommended for microsatellite data (Patterson *et al*., 2006), we omitted the normalization step by setting the *usenorm* parameter of the SMARTPCA program to 'NO.'

Finally, when the analyses using STRUCTURE and EIGENSOFT (described above) indicated the existence of population substructure, we used the spatially explicit, Bayesian clustering algorithm implemented in v. 3.1.4 of the R package GENELAND to delineate the putative sub-populations (Guillot *et al*., 2005, 2008; Coulon *et al*., 2006). We used the spatial model with correlated allele frequencies of GENELAND, set the number of sub-populations to the number inferred using STRUCTURE (see above), assumed uncertainty of spatial coordinates of two meters, and ran the analyses for 1 000 000 iterations, using a thinning value of 100. By default, the null allele model implemented in GENELAND treats all missing data as homozygous null genotypes. This model was not used, because we did not have information reliable enough to distinguish genuine homozygous null genotypes from missing data. Computer simulations indicate that unaccounted null alleles can lead to overestimates of the number of sub-populations (K) using GENELAND, but do not affect the accuracy of clustering (that is, assignment of individual to sub-populations), even when null alleles occur at very high frequencies (that is, up to 20%; Guillot *et al*., 2008). Therefore, we used GENELAND only when there was strong evidence that $K>1$, on the basis of analyses using STRUCTURE and analyses of molecular variance (described below).

Hierarchical population substructure was quantified using Wright's F-statistics (Wright, 1965) as estimated using the locus-by-locus analysis of molecular variance option of the ARLEQUIN program. Tests of differences for diameter and flowering phenology among sub-populations were performed using the *t*-test function of the R package STATS, and tests for differences in the number of ramets per genet were performed using the exact tests implemented in the independence_test function of the R package COIN (http://www.cran.r-project.org/).

Relatedness: To elucidate the underlying causes of population substructure, we used the 'group-likelihood' approach implemented in v. 2.0 of the COLONY program (Wang and Santure, 2009) to infer sibling relationships. This analysis was performed for each population (or sub-population detected as described above) assuming that (1) male and female parents were polygamous, (2) the overall rate of genotyping error and mutations

was 10% (5% null alleles and 5% other types of mutations and genotyping error) and (3) parent–offspring relationships among genets within a population or sub-population are possible. The rates of mutations and genotyping error were chosen conservatively on the basis of empirical data from Mendelian inheritance tests (Slavov *et al*., 2009). Only relationships supported by probabilities of at least 0.9 were assumed to be correct. To evaluate the rates of statistical error, we also formed data sets consisting of seedlings known to be half sibs or full sibs based on known maternal parentage and paternity assignments (Slavov *et al*., 2009). Putative relationships were visualized using the R package NETWORK.

Results

Clonality, microsatellite diversity and deviations from Hardy–Weinberg equilibrium

The extent of clonality ($1-R$) was substantially higher in the arid Vinson site (Table 1), and this difference was statistically significant ($P=0.006$, based on an exact test of the difference in the numbers of ramets per genet in the two populations). However, levels of microsatellite diversity and polymorphism were comparable between the two populations (Table 1). The only substantial difference was in the average number of alleles per locus detected in each population (A), but this appears to result from the different numbers of genets sampled in the two populations (that is, allelic richness, $r(10)$, was practically the same in the two populations).

Statistically significant deviations from Hardy–Weinberg equilibrium (after Bonferroni corrections for multiple testing) were detected for three loci (30%) in the Marchel population and for two loci (20%) in the Vinson population. All of these deviations were caused by heterozygote deficiency. On the basis of the global exact tests of Hardy–Weinberg equilibrium across all loci, heterozygote deficiency was highly significant (one-sided $P<0.0001$) in both populations. The extent of heterozygote deficiency (F_{IS}) appeared to be slightly higher in the Vinson population (Table 1). Significant genotypic disequilibrium (after Bonferroni corrections) was detected for 8 out of the 45 pairs of loci in the Marchel population and 6 out of the 45 pairs of loci in the Vinson population. However, none of these pairs showed significant genotypic disequilibrium in both populations.

Population substructure

The mean pairwise kinship coefficient of genets located within 20 m of one another and the strength of the overall association of pairwise kinship coefficients and physical distances among genets were an order of magnitude higher in the Vinson population than in the Marchel population, resulting in a nearly nine times higher value of Sp in the Vinson population (Table 2). Alternative definitions of the first distance class (for example, 0–10 and 0–50 m) resulted in similar patterns.

Surprisingly, the continuous Marchel population appeared to consist of at least two sub-populations. Log-likelihoods from runs of STRUCTURE with $K=1$–5 had a clear peak at $K=2$ in all 20 replicates (Figure 2).

Table 2 Spatial genetic structure in two *P. trichocarpa* populations with contrasting ecological and demographic characteristics

Population	$F_{(1)}$ *(two-sided* P*-value)*[a]	b_F *(two-sided* P*-value)*[b]	*Sp*[c]
Marchel	0.0085 (0.049)	−0.0019 (0.007)	0.0020
Vinson	0.0993 (0.021)	−0.0154 (0.003)	0.0171

[a]$F_{(1)}$ is the mean pairwise kinship coefficient (Loiselle *et al*., 1995) for pairs of trees separated by up to 20 m.
[b]b_F is the simple linear regression slope of pairwise kinship coefficients on the natural logarithm of pairwise distances among trees.
[c]*Sp* is a standardized measure of SGS (see Materials and methods).

Table 1 Clonality, microsatellite diversity, population substructure, relative clone age and flowering phenology in two populations of *P. trichocarpa*

Population/sub-population	N[a]	G[b]	R[c]	A[d]	r(10)[e]	H_o[f]	H_e[g]	F_{IS} (P*-value)*[h]	*MaxDBH*[i]	*Flowering*[j]
Marchel	292	184	0.63	17.5 (2.6)	5.7 (0.6)	0.734 (0.086)	0.782 (0.088)	0.062 (<0.001)	55.0 (1.8)	89.8 (0.5)
Sub-population 1[k]	43	39	0.90	12.4 (1.6)	5.7 (0.6)	0.763 (0.092)	0.786 (0.088)	0.029 (0.072)	48.1 (3.6)	86.5 (1.1)
Sub-population 2[k]	140	88	0.63	13.7 (1.8)	5.8 (0.3)	0.717 (0.087)	0.762 (0.086)	0.060 (<0.001)	58.4 (2.7)	91.5 (0.5)
Unassigned[k]	109	57	0.52	14.3 (1.9)	5.7 (0.6)	0.740 (0.085)	0.781 (0.087)	0.053 (0.001)	54.7 (3.3)	89.3 (1.0)
Vinson	97	31	0.31	11.1 (1.6)	5.3 (0.6)	0.675 (0.100)	0.737 (0.077)	0.085 (0.003)	59.7 (5.2)	92.7 (1.1)
Sub-population 1[l]	31	11	0.33	6.3 (1.0)	5.0 (0.7)	0.719 (0.106)	0.726 (0.079)	0.008 (0.411)	53.6 (5.5)	92.1 (1.7)
Sub-population 2[l]	30	13	0.41	6.7 (0.9)	4.7 (0.6)	0.660 (0.111)	0.682 (0.094)	0.034 (0.212)	67.8 (10.1)	91.0 (1.5)
Unassigned[l]	36	7	0.17	6.0 (0.8)	5.1 (0.6)	0.652 (0.111)	0.742 (0.070)	0.130 (0.010)	52.4 (6.4)	97.2(1.9)

Standard errors are shown in parentheses unless indicated otherwise.
[a]N is the number of ramets sampled.
[b]G is the number of genets (that is, multilocus genotypes that differ for at least two loci).
[c]$R=(G-1)/(N-1)$ is a measure of clonal diversity (Arnaud-Haond *et al*., 2007).
[d]A is the average number of alleles per locus detected in each population or sub-population.
[e]$r(10)$ is the allelic richness, defined as the number of alleles that would have been detected if 10 alleles (5 genets) had been sampled in each population (El Mousadik and Petit, 1996; Petit *et al*., 1998).
[f]H_o is the observed heterozygosity.
[g]H_e is the unbiased estimate of gene diversity (Nei, 1978).
[h]F_{IS} is the fixation index measuring the correlation of alleles within individuals relative to that within populations (Wright, 1965). Two-sided P-values, based on 1000 permutations of alleles among individuals, are shown in parentheses.
[i]MaxDBH is the breast-height diameter of the largest ramet in each genet (cm). This parameter was used as a surrogate for clone age.
[j]Flowering is the average date of peak flowering (Julian days).
[k]Sub-populations were defined as groups of genets whose posterior probabilities of membership in their respective cluster (based on the GENELAND analysis; Figure 3) was at least 0.8.
[l]Sub-populations were defined based on analyses of relatedness using COLONY (see Figure 4).

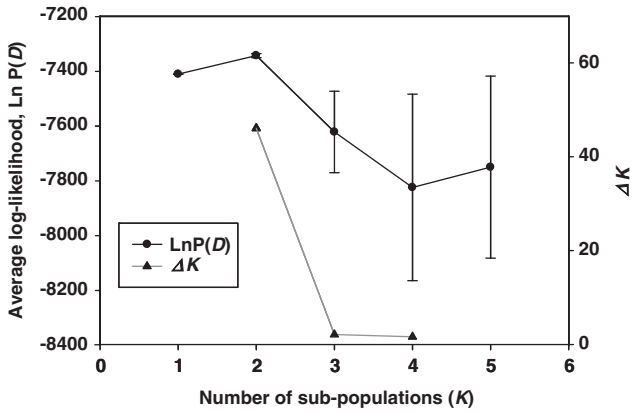

Figure 2 Population substructure in the Marchel population as detected using the STRUCTURE program. Circles indicate average log-likelihoods from 20 program runs (see Materials and methods) for each assumed number of sub-populations (K), and error bars correspond to 1 s.d. Triangles indicate values of the *ad hoc* statistic ΔK, which is based on the rate of change of the log-likelihood as K is increased. ΔK tends to peak at the value of K that corresponds to the highest level of hierarchical substructure (Evanno *et al*., 2005).

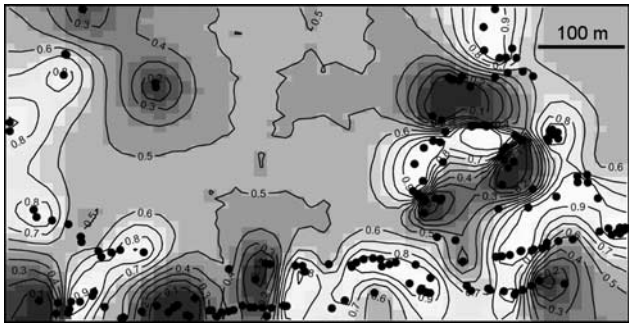

Figure 3 Population substructure in the Marchel population as detected using the R package GENELAND (see Materials and methods). Filled circles indicate the relative positions of the sampled genets (see Figure 1). Darker and lighter shading are proportional to posterior probabilities of membership in clusters (that is, sub-populations) 1 and 2, respectively. Isolines and their corresponding labels represent posterior probabilities of membership in cluster 2.

The posterior probability of $K=2$, calculated based on results from runs with $K=1$–5 using Bayes' formula (as shown in the documentation of STRUCTURE), was $\Pr(K=2\,|\,Data)=1.00$, and most genets (89%) were assigned strongly to one of the sub-populations (Supplementary Figure S3). The existence of two (and not more) sub-populations was also supported by the *ad hoc* statistic ΔK (Figure 2). Furthermore, there was overwhelming statistical support ($P<10^{-10}$ from a test based on the Tracy–Widom distribution; Patterson *et al*., 2006) for the first two axes of variation identified using the principal components analysis approach implemented in EIGENSOFT. Spatially explicit clustering analyses using GENELAND indicated that the two sub-populations form a mosaic of intermingled groups of clones (Figure 3). This pattern was consistent with the clustering results from our analyses using STRUCTURE and EIGENSOFT. The two sub-populations, which were defined as the collection of genets whose posterior probabilities of membership in their respective clusters was at least 0.8 (based on the analyses using GENELAND), had different scores on the first two axes of variation identified using EIGENSOFT ($P=0.050$ and 1.9×10^{-12}, respectively). Furthermore, the posterior probabilities of cluster membership obtained using GENELAND were strongly correlated with those from the STRUCTURE analyses with $K=2$ ($r=0.74$, $P=2\times10^{-16}$). Differentiation between these two sub-populations was relatively weak but statistically significant ($F_{ST}=0.026$, $P<0.001$). Heterozygote deficiency relative to Hardy–Weinberg expectations (F_{IS}) was substantially weaker (and not statistically significant) in one of the two sub-populations compared with the population as a whole (that is, ignoring population substructure; Table 1). Levels of microsatellite diversity and polymorphism were practically identical in the two sub-populations (Table 1). Interestingly, these sub-populations appeared to have different clone ages as measured by the breast-height diameter of the largest ramet in each clone (one-sided $P=0.012$), clone sizes as measured by the number of ramets per clone (one-sided $P=0.027$), and flowering phenologies (one-sided $P<10^{-4}$) (Table 1).

In contrast, runs of STRUCTURE that assumed $K=1$ consistently resulted in the highest log-likelihoods in the low-density, fragmented Vinson population (data not shown). Furthermore, none of the eigenvectors calculated using EIGENSOFT based on data for this population explained a significant proportion of microsatellite variation ($P>0.08$). Analyses of molecular variance, in which trees were grouped into three putative sub-populations corresponding to the stands delineated in Figure 1b, also failed to detect statistically significant population substructure ($F_{ST}=0.041$, $P=0.180$).

Relatedness

Analyses using the data sets that consisted of known siblings (that is, based on data and analyses described by Slavov *et al*., 2009) had moderate-to-high power and relatively low rates of false positives. When the parental genotypes of the seedlings included in the test data set were used in the analyses (but without specifying the known parent–offspring relationships), the reconstruction of sibling relationship was 100% accurate. When parental genotypes were not included, 42 of the 60 (70.0%) known sibling relationships were reconstructed correctly, and sibling relationships were incorrectly inferred for nine pairs of putatively unrelated siblings (2.4% of all pairwise comparisons of putatively unrelated seedlings).

The rate of putative pairwise sibling relationships was higher among the adult genets in the Vinson (11.2%) than in the Marchel population as a whole (4.2%) or the two sub-populations detected in the Marchel population (6.5 and 7.8%, respectively, for sub-populations 1 and 2). Remarkably, 29 of the 31 genets in the Vinson population had at least one putative sibling. Nearly half of the genets (14 out of 31) in this population formed three interconnected half-sib families that consisted of at least five members, and two putative full sibs (San8 and San13clone) had as many as eight putative siblings (Figure 4). Analyses of molecular variance in which the two clusters of sib

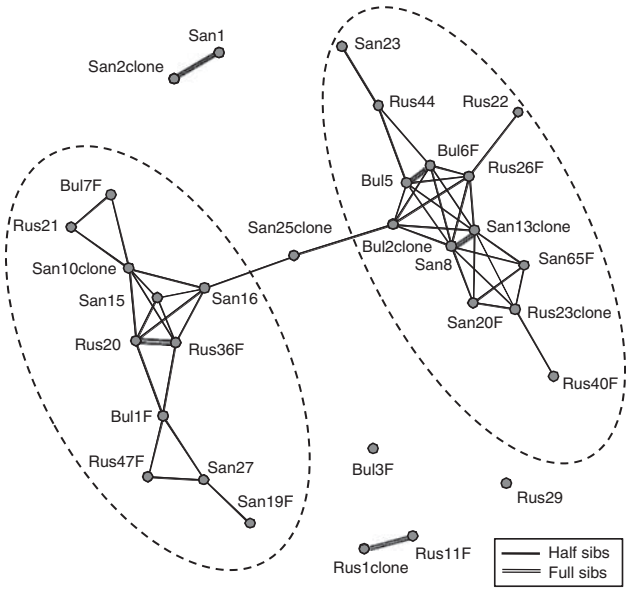

Figure 4 Putative sibling relationships in the Vinson population as determined using the COLONY program (see Materials and methods). Filled circles represent the 31 sampled genets. Relative positions and line lengths were chosen to allow the illustration of putative sibling relationships and do not correspond to the actual spatial situation (see Figure 1b) or to the support for the putative relationships. Genets with names ending in 'clone' had multiple ramets. Single lines in the network plot connect putative half sibs and double lines connect putative full sibs. All putative relationships were supported by probabilities of at least 0.9. The two clusters of sib families circled with dashed lines were significantly differentiated ($F_{ST} = 0.071$, $P = 0.001$).

families delineated in Figure 4 were treated as sub-populations indicated that the heterozygote deficiency detected in the population as a whole ($F_{IS} = 0.085$, Table 1) was caused primarily by sub-population differentiation ($F_{ST} = 0.071$, $P = 0.001$), and heterozygote deficiency was not statistically significant in either of the two sub-populations (Table 1). No statistically significant differences in relative clone age, clone size or flowering phenology were detected between these two sub-populations (Table 1).

Discussion

Clonality, microsatellite diversity and deviations from Hardy–Weinberg equilibrium

Results from our analyses of clonality ($1-R$, Table 1) were consistent with those from earlier studies indicating that asexual propagation is common throughout the genus *Populus* (Rood *et al*., 2003, 2007; Mock *et al*., 2008; Smulders *et al*., 2008; van Loo *et al*., 2008). The extent of clonality in the two study populations was roughly comparable to that determined based on RAPD markers in populations of *P. trichocarpa* sampled along the Yakima, Nisqually and Cowlitz rivers in central and western Washington (Reed, 1995; McKay, 1996). As expected based on the results from the former study, the extent of clonality was approximately twice as high in the drier Vinson site, where successful seedling establishment may be occurring extremely infrequently (discussed below).

Levels of microsatellite diversity and polymorphism (H_o, H_e, A, $r(10)$; Table 1) observed in our study are comparable with those in 47 populations of *P. trichocarpa* sampled throughout British Columbia (M Ismail, personal communication) and are among the highest reported for any species of *Populus* (Slavov and Zhelev, 2010). These levels were very similar in the two study populations, despite the dramatic demographic difference between them. Thus, the extensive long-distance pollen flow detected in the fragmented stands of the Vinson population (Slavov *et al*., 2009) may be sufficient to at least partly mitigate the effects of genetic drift.

As in the vast majority of studies in which microsatellite markers were used to study the population genetic structure of *Populus* species (Slavov and Zhelev, 2010), we detected relatively weak but statistically significant deviations from Hardy–Weinberg equilibrium. These deviations were present at multiple loci in both study populations, and were invariably caused by apparent deficiency of heterozygotes. This heterozygote deficiency probably resulted largely from the occurrence of null alleles and allele drop-outs (that is, alleles that were not detected because of preferential amplification of their homologous alleles in heterozygous genotypes; Ewen *et al*., 2000). Consistent with this, heterozygote deficiency is more common in microsatellite-based studies of the population genetic structure of *Populus* species than in studies based on allozyme and restriction fragment length polymorphism markers (Slavov and Zhelev, 2010). Furthermore, residual deficiency of heterozygotes was present in all of the sub-populations we detected. However, results from our clustering analyses, and especially the statistically significant differentiation (F_{ST}) among the sub-populations we detected, strongly suggest that deviations from Hardy–Weinberg equilibrium were at least partly caused by population substructure.

Population substructure and relatedness

The standardized measures of SGS (*Sp*) we obtained for the two study populations are generally comparable to those observed in other forest trees (Vekemans and Hardy, 2004), including other species of *Populus* and their hybrids (Pospíšková and Šálková, 2006; van Loo *et al*., 2008).

What is the explanation for the substantially stronger SGS in the Vinson population? The different spatial scales sampled in the two study sites may have accentuated the difference between the *Sp* statistics calculated for the two populations (Vekemans and Hardy, 2004). However, this difference most likely results from a combination of several ecological and demographic factors that distinguish the two study populations. First, because the availability of moisture in the first growing season is the most critical factor for seedling survival (Braatne *et al*., 1996; Steinberg, 2001), seedling establishment is probably less frequent in the arid habitats of eastern Oregon (Karrenberg *et al*., 2002; Braatne *et al*., 2007). This may be further exacerbated by the high grazing pressure by domestic livestock and wildlife in these habitats (Case and Kauffman, 1997).

Second, the number of reproductively mature female trees per unit of area is much smaller in the Vinson population than in the Marchel population. This fact, combined with the tremendous phenological variation within *Populus* populations (Howe *et al*., 2003), means that successfully established seedlings in the Vinson population may be produced by the very few mother trees that are shedding seeds during the rare windows of favorable conditions. Finally, paternity analyses performed as part of a previous study revealed that a large proportion of progeny arrays sampled from female trees in the Vinson population consisted of full sibs resulting from mating with only one or two male trees (Slavov *et al*., 2009). Consistent with this, our analyses of relatedness among the adult trees in the Vinson population also revealed extensive family structure (Figure 4). Thus, the strong SGS observed in the Vinson population probably results from infrequent establishment of seedling cohorts produced by mating among relatively few adult trees. Similar patterns of stronger SGS in populations with lower densities are expected under models of isolation by distance and have been detected in multiple plant species (Vekemans and Hardy, 2004).

Analyses using STRUCTURE and EIGENSOFT did not provide evidence for the existence of population substructure in the Vinson population. This was probably because of the small number of genets in this population and the small number of loci we used. Patterson *et al*. (2006), for example, derived an asymptotic expression for the threshold of substructure (that is, minimum F_{ST}) that is detectable using EIGENSOFT as a function of the number of loci and number of individuals sampled. Using computer simulations, Patterson *et al*. (2006) showed that this threshold also applied to analyses using STRUCTURE. After accounting for missing data, and assuming that this relationship is roughly similar for microsatellite markers, we calculated a threshold of detectable substructure of $F_{ST} = 0.076$ for the two Vinson sub-populations we detected using analyses of relatedness (discussed below). The value of F_{ST} that we estimated for these two sub-populations (0.071) falls just below that threshold. Because the power of analyses using EIGENSOFT and STRUCTURE is low for F_{ST} values below or even at the threshold (Patterson *et al*., 2006), the failure to detect substructure in the Vinson population using these approaches is not surprising.

In contrast, our analyses of relatedness using COLONY clearly delineated two groups of inter-related putative half-sib families (Figure 4). The differentiation between these two groups accounted for most of the deviation from Hardy–Weinberg equilibrium detected when the presence of population substructure was ignored. The relatively small number of loci we used probably limited our power to accurately partition cohorts into sibships and discriminate among relationship categories (Blouin, 2003; Weir *et al*., 2006). Our analyses of relatedness, however, depicted a clear pattern of substructure in the Vinson population. Thus, attempting to partition samples into putative sibships or family clusters appears to be a useful supplement to the standard set of analytical approaches used to detect population substructure. This approach may be particularly appropriate when few loci are used ($\leqslant 10$) and sample sizes are small ($N \leqslant 50$). In these situations, model-based clustering and principal components analysis have low power to detect subtle population substructure (Patterson *et al*., 2006), whereas analyses using COLONY can potentially provide informative results and tend to have moderate run times. For example, our run times ranged from a few hours to a few days on a standard personal computer, depending on the complexity of the underlying family structure and the assumed rates of genotyping error. In our experience, the run times of COLONY become unfeasible with substantially larger data sets, and patterns of population substructure are difficult to interpret and summarize when multiple relatively small families are detected (as in the Marchel population).

The more subtle but statistically significant population substructure we detected in the continuous stand at the Marchel site was unexpected. On the basis of the life history characteristics of *P. trichocarpa* and other cottonwoods (Braatne *et al*., 1996, 2007; Karrenberg *et al*., 2002; Rood *et al*., 2007), the structure and successional dynamics of *P. trichocarpa* stands along the Willamette River (Fierke and Kauffman, 2005) and the differences in presumed clone age and flowering phenology we detected between the two Marchel sub-populations (Table 1), we hypothesize that they represent different seedling cohorts.

Flood control has gradually reduced the magnitude of peak flows along the Willamette River, and peak floods after 1980 reached less than 50% of the levels measured a century earlier (Fierke and Kauffman, 2005). Consistent with the 'flood pulse' concept (Samuelson and Rood, 2004; Rood *et al*., 2007), the recent absence of major floods has influenced the age structure of cottonwood stands along the river, with young stands being relatively rare (Dykaar and Wigington, 2000; Fierke and Kauffman, 2005). The mean ages of the five successional classes identified by Fierke and Kauffman (2005) were consistent with the presumed 20-year span between episodes of massive *P. trichocarpa* recruitment along the Willamette River (Dykaar and Wigington, 2000). The mean diameters of the presumed clone progenitors in the two Marchel sub-populations (48.1 ± 3.6 and 58.4 ± 2.7 cm, respectively) matched closely those of the mid (48.0 ± 4.3 cm, mean age = 39 years) and late (62.5 ± 4.3 cm, mean age >65 years) seral stages identified by Fierke and Kauffman (2005). The presumed clone ages of the 57 genets that were not assigned to either cluster appeared to be intermediate to those of the genets in the two sub-populations (Table 1). It is possible that this group consists of a mixture of genets from the two sub-populations we detected, but could not be clustered based on our microsatellite data. Alternatively, this group may also include, or be entirely comprised of, genets from several smaller seedling cohorts.

The two Marchel sub-populations also differed in their flowering phenologies (Table 1). Therefore, it is possible that the two cohorts were established from seeds produced by two relatively small sets of mothers that differed in their timing of seed release, each set of mothers 'taking advantage' of a window of opportunity compatible with its phenology. Under this scenario, the population substructure we detected is merely a consequence of the higher average degree of relatedness within a sub-population than in the population as a whole. Interestingly, the trees from the 'younger' sub-population 1 flowered 5 days earlier than those in

sub-population 2 (Table 1). This is qualitatively and quantitatively consistent with long-term trends of phenological shifts driven by global environmental change that have been observed in a number of plants, including other species of *Populus* (reviewed by Cleland *et al*., 2007; Bertin, 2008). Thus, scenarios that involve interactions of the life history characteristics discussed above, mating patterns (that is, assortative mating with respect to flowering phenology) and climatic selection following establishment are also plausible, but a conclusion about the specific causes of the cohort differentiation we observed is not warranted in the absence of experimental data.

Similar stand dynamics characterized by (1) rare pulses of large-scale recruitment that occurred before dams were built to provide irrigation water, (2) age structure dominated by 40- to 50-year-old cohorts, and (3) lack of stands established after 1980 were depicted in a study of *P. trichocarpa* stands along the Yakima River in Washington (Braatne *et al*., 2007) and are likely to exist in riparian ecosystems throughout the Pacific Northwest. Thus, elucidating the specific causes of population substructure in *P. trichocarpa* populations would be valuable for our basic understanding of the biology of this species.

Implications for association genetic studies

Undetected population substructure can cause spurious phenotype–genotype correlations and is therefore a major concern in association genetic studies (Hirschhorn and Daly, 2005; McCarthy *et al*., 2008; Ioannidis *et al*., 2009). A number of statistical approaches to control for substructure and relatedness when testing phenotype–genotype associations have been developed (reviewed by Balding, 2006; McCarthy *et al*., 2008), but researchers often do not report even simple diagnostics for the presence of substructure (for example, deviations from Hardy–Weinberg equilibrium; Salanti *et al*., 2005; Trikalinos *et al*., 2006). Computer simulations and empirical data show that even very subtle population substructure ($F_{ST} = 0.010$ or less) can cause high rates of spurious associations, unless hundreds of molecular markers have been used to control for its confounding effect (Price *et al*., 2006). Thus, the weak substructure ($F_{ST} = 0.026$) we detected in the continuous and seemingly homogeneous Marchel stand has practical significance. Similar patterns are likely to exist in populations of *P. trichocarpa* throughout the Pacific Northwest of North America. Therefore, it appears that population substructure will be present in any association mapping population of *P. trichocarpa* that was formed by sampling multiple trees from the same stand. Researchers involved in ongoing and future association studies should control for population substructure by using a large number of 'random' markers (*sensu*, Price *et al*., 2006) to detect it and incorporate quantitative measures of its extent in analyses of phenotype–genotype associations. Alternatively, results from association studies whose scope did not allow adequate control for population substructure should be reported with this caveat.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Assistance with field and laboratory work was provided by Shuping Cheng, Gokcin Temel, Jace Carson, Caprice Rosato, Eliza Walthers, Paul Rosenfeld, Rick Meilan, Toby Bradshaw and Brian Watson. Financial support for this study was provided by an EPA STAR Fellowship to SPD, the Tree Biosafety and Genomics Research Cooperative, USDA Biotechnology Risk Assessment Grant 97-39210-5022, the DOE Oak Ridge National Lab Bioenergy Program, the DOE BioEnergy Science Center and NSF FIBR Grant DEB-0425908. We thank Reinhard Stettler, Glenn Howe, three anonymous reviewers, and the Heredity Editor for their comments on an earlier version of this article.

References

Arnaud-Haond S, Duarte CM, Alberto F, Serrao EA (2007). Standardizing methods to address clonality in population studies. *Mol Ecol* **16**: 5115–5139.

Balding DJ (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781–791.

Bertin RI (2008). Plant phenology and distribution in relation to recent climate change. *J Torrey Botanical Society* **135**: 126–146.

Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol Evol* **18**: 503–511.

Braatne JH, Jamieson R, Gill KM, Rood SB (2007). Instream flows and the decline of riparian cottonwoods along the Yakima River, Washington, USA. *River Research and Applications* **23**: 247–267.

Braatne JH, Rood SB, Heilman PE (1996). Life history, ecology, and reproduction of riparian cottonwoods in North America. In: Stettler RF, Bradshaw Jr HD, Heilman PE, Hinckley TM (eds). *Biology of Populus and its Implications for Management and Conservation*. NRC Research Press: Ottawa, Canada. pp 57–85.

Burczyk J, DiFazio SP, Adams WT (2004). Gene flow in forest trees: how far do genes really travel? *Forest Genet* **11**: 179–192.

Case RL, Kauffman JB (1997). Wild ungulate influences on the recovery of willows, black cottonwood and thin-leaf alder following cessation of cattle grazing in northeastern Oregon. *Northwest Sci* **71**: 115–126.

Cleland EE, Chuine I, Menzel A, Mooney HA, Schwartz MD (2007). Shifting plant phenology in response to global change. *Trends Ecol Evol* **22**: 357–365.

Coulon A, Guillot G, Cosson JF, Angibault JMA, Aulagnier S, Cargnelutti B *et al.* (2006). Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Mol Ecol* **15**: 1669–1679.

Dakin EE, Avise JC (2004). Microsatellite null alleles in parentage analysis. *Heredity* **93**: 504–509.

DeBell DS (1990). *Populus trichocarpa* Torr. & Gray, Black Cottonwood. In: Burns RM, Honkala BH (eds). *Silvics of North America Vol. 2. Hardwoods. Agriculture Handbook 654*. US Department of Agriculture, Forest Service: Washington, DC.

Dieringer D, Schlötterer C (2003). Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes* **3**: 167–169.

DiFazio SP (2002). Measuring and Modeling Gene Flow from Hybrid Poplar Plantations: Implications for Transgenic Risk Assessment. PhD Thesis, Oregon State University, Corvallis, OR. http://www.esd.ornl.gov/PGG/difaz_thesis.pdf.

Dykaar BB, Wigington PJ (2000). Floodplain formation and cottonwood colonization patterns on the Willamette River, Oregon, USA. *Environ Manage* **25**: 87–104.

El Mousadik A, Petit RJ (1996). Chloroplast DNA phylogeography of the argan tree of Morocco. *Mol Ecol* **5**: 547–555.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611–2620.

Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW *et al.* (2000). Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet* **67**: 727–736.

Excoffier L, Laval G, Schneider S (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform* **1**: 47–50.

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.

Falush D, Stephens M, Pritchard JK (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* **7**: 574–578.

Fierke MK, Kauffman JB (2005). Structural dynamics of riparian forests along a black cottonwood successional gradient. *Forest Ecol Manage* **215**: 149–162.

Guillot G, Mortier F, Estoup A (2005). GENELAND: a computer package for landscape genetics. *Mol Ecol Notes* **5**: 712–715.

Guillot G, Santos F, Estoup A (2008). Analysing georeferenced population genetics data with GENELAND: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics* **24**: 1406–1407.

Hamrick JL, Godt MJW, Sherman-Broyles SL (1992). Factors influencing levels of genetic diversity in woody plant species. *New Forests* **6**: 95–124.

Hardy OJ, Vekemans X (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**: 618–620.

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.

Howe GT, Aitken SN, Neale DB, Jermstad KD, Wheeler NC, Chen THH (2003). From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Can J Botany* **81**: 1247–1266.

Ingvarsson PK (2010). Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in *Populus*. In: Jansson S, Bhalerao R, Groover AT (eds). *Genetics and Genomics of Populus*. Springer: NY. pp 91–112.

Ioannidis JPA, Thomas G, Daly MJ (2009). Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* **10**: 318–329.

Jansson S, Bhalerao R, Groover AT (2010). *Genetics and Genomics of Populus*. Springer: NY.

Jansson S, Douglas CJ (2007). *Populus*: a model system for plant biology. *Annu Rev Plant Biol* **58**: 435–458.

Kalinowski ST, Taper ML, Marshall TC (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol Ecol* **16**: 1099–1106.

Karrenberg S, Edwards PJ, Kollmann J (2002). The life history of Salicaceae living in the active zone of floodplains. *Freshwater Biol* **47**: 733–748.

Langley CH, Voelker RA, Brown AJL, Ohnishi S, Dickson B, Montgomery E (1981). Null allele frequencies at allozyme loci in natural populations of *Drosophila melanogaster*. *Genetics* **99**: 151–156.

Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Botany* **82**: 1420–1425.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356–369.

McKay SJ (1996). The Impact of River Regulation on Establishment Processes of Riparian Black Cottonwood. MS Thesis. University of Washington, USA.

Mock KE, Rowe CA, Hooten MB, DeWoody J, Hipkins VD (2008). Clonal dynamics in western North American aspen (*Populus tremuloides*). *Mol Ecol* **17**: 4827–4844.

Nei M (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: 2074–2093.

Petit RJ, El Mousadik A, Pons O (1998). Identifying populations for conservation on the basis of genetic markers. *Conservation Biol* **12**: 844–855.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005). Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**: 847–859.

Pospíšková M, Šálková I (2006). Population structure and parentage analysis of black poplar along the Morava River. *Can J Forest Res* **36**: 1067–1076.

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

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

Reed JP (1995). Factors Affecting the Genetic Architecture of Black Cottonwood Populations. MS Thesis. University of Washington, USA.

Rood SB, Goater LA, Mahoney JM, Pearce CM, Smith DG (2007). Flood, fire, and ice: disturbance ecology of riparian cottonwoods. *Can J Botany* **85**: 1019–1032.

Rood SB, Kalischuk AR, Polzin ML, Braatne JH (2003). Branch propagation, not cladoptosis, permits dispersive, clonal reproduction of riparian cottonwoods. *Forest Ecol Manage* **186**: 227–242.

Rousset F (2008). GENEPOP′ 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resources* **8**: 103–106.

Salanti G, Amountza G, Ntzani EE, Ioannidis JP (2005). Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *Eur J Hum Genet* **13**: 840–848.

Samuelson GM, Rood SB (2004). Differing influences of natural and artificial disturbances on riparian cottonwoods from prairie to mountain ecoregions in Alberta, Canada. *J Biogeography* **31**: 435–450.

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.

Slavov GT, Leonardi S, Burczyk J, Adams WT, Strauss SH, DiFazio SP (2009). Extensive pollen flow in two ecologically contrasting populations of *Populus trichocarpa*. *Mol Ecol* **18**: 357–373.

Slavov GT, Zhelev P (2010). Salient biological features, systematics, and genetic variation of *Populus*. In: Jansson S, Bhalerao R, Groover AT (eds). *Genetics and Genomics of Populus*. Springer: NY. pp 15–38.

Smouse PE, Sork VL (2004). Measuring pollen flow in forest trees: an exposition of alternative approaches. *Forest Ecol Manage* **197**: 21–38.

Smulders MJM, Cottrell JE, Lefèvre F, Van Der Schoot J, Arens P, Vosman B *et al.* (2008). Structure of the genetic diversity in black poplar (*Populus nigra* L.) populations across European river systems: Consequences for conservation and restoration. *Forest Ecol Manage* **255**: 1388–1399.

Sork VL, Smouse PE (2006). Genetic analysis of landscape connectivity in tree populations. *Landscape Ecol* **21**: 821–836.

Steinberg PD (2001). *Populus balsamifera* subsp. *trichocarpa*. In: *Fire Effects Information System [Online]*, US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory. http://www.fs.fed.us/database/feis/.

Trikalinos TA, Salanti G, Khoury MJ, Ioannidis JPA (2006). Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. *Am J Epidemiol* **163**: 300–309.

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

Vekemans X, Hardy OJ (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol Ecol* **13**: 921–935.

Wang J, Santure AW (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**: 1579–1594.

Weir BS, Anderson AD, Hepler AB (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**: 771–780.

Wright S (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**: 295–420.

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