New Phytologist

Research

Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*

Gancho T. Slavov[1,2], Stephen P. DiFazio[1], Joel Martin[3], Wendy Schackwitz[3], Wellington Muchero[4], Eli Rodgers-Melnick[1], Mindie F. Lipphardt[1], Christa P. Pennacchio[3], Uffe Hellsten[3], Len A. Pennacchio[3], Lee E. Gunter[4], Priya Ranjan[4], Kelly Vining[5], Kyle R. Pomraning[6], Larry J. Wilhelm[7], Matteo Pellegrini[8], Todd C. Mockler[9], Michael Freitag[6], Armando Geraldes[10], Yousry A. El-Kassaby[11], Shawn D. Mansfield[12], Quentin C. B. Cronk[10], Carl J. Douglas[10], Steven H. Strauss[5], Dan Rokhsar[3] and Gerald A. Tuskan[4]

[1]Department of Biology, West Virginia University, Morgantown, WV 26506-6057, USA; [2]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3EB, UK; [3]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; [4]BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; [5]Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331-5752, USA; [6]Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331-7305, USA; [7]Oregon Health Sciences University, Beaverton, OR 97006, USA; [8]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095-1606, USA; [9]Donald Danforth Plant Science Center, Saint Louis, MO 63132, USA; [10]Department of Botany, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4; [11]Department of Forest Sciences, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4; [12]Department of Wood Science, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

Author for correspondence:
Gancho T. Slavov
Tel: +44 1970 823094
Email: gts@aber.ac.uk

Received: *26 April 2012*
Accepted: *27 June 2012*

New Phytologist (2012)
doi: 10.1111/j.1469-8137.2012.04258.x

Key words: allele frequency gradients, black cottonwood (*Populus trichocarpa*), genome resequencing, linkage disequilibrium (LD), population structure, recombination.

Summary

- Plant population genomics informs evolutionary biology, breeding, conservation and bioenergy feedstock development. For example, the detection of reliable phenotype–genotype associations and molecular signatures of selection requires a detailed knowledge about genome-wide patterns of allele frequency variation, linkage disequilibrium and recombination.
- We resequenced 16 genomes of the model tree *Populus trichocarpa* and genotyped 120 trees from 10 subpopulations using 29 213 single-nucleotide polymorphisms.
- Significant geographic differentiation was present at multiple spatial scales, and range-wide latitudinal allele frequency gradients were strikingly common across the genome. The decay of linkage disequilibrium with physical distance was slower than expected from previous studies in *Populus*, with r^2 dropping below 0.2 within 3–6 kb. Consistent with this, estimates of recent effective population size from linkage disequilibrium ($N_e \approx 4000$–6000) were remarkably low relative to the large census sizes of *P. trichocarpa* stands. Fine-scale rates of recombination varied widely across the genome, but were largely predictable on the basis of DNA sequence and methylation features.
- Our results suggest that genetic drift has played a significant role in the recent evolutionary history of *P. trichocarpa*. Most importantly, the extensive linkage disequilibrium detected suggests that genome-wide association studies and genomic selection in undomesticated populations may be more feasible in *Populus* than previously assumed.

Introduction

Forest tree species are assumed to have large effective population sizes because of their high levels of genetic diversity, weak inter-population differentiation for neutral loci and rapid decay of linkage disequilibrium (LD) with physical distance (Hamrick *et al.*, 1992; Neale & Ingvarsson, 2008; Neale & Kremer, 2011). However, because most previous population genetic studies were based on small numbers of statistically independent loci (González-Martínez *et al.*, 2006b; Savolainen & Pyhäjärvi, 2007), relatively little is known about the genome-wide patterns of allele frequency variation, LD and recombination in these ecologically and economically important organisms. Detailed information on these patterns is indispensable for understanding the evolutionary history of tree populations and for designing, analyzing and interpreting data from association studies and selection scans. Thus, in addition to providing unprecedented opportunities for marker-assisted breeding through genomic

selection (i.e. the prediction of genetic value from high-density single-nucleotide polymorphism (SNP) genotype data; Meuwissen & Goddard, 2010), genome-wide surveys of DNA polymorphism can yield novel biological insights.

Because of their role as foundation species in a number of ecosystems (Whitham *et al.*, 2006), wide geographic distribution, rapid growth and potential as a bioenergy crop (Rubin, 2008), species of *Populus* have become models for tree genomics (Jansson *et al.*, 2010) and have well-developed molecular resources, including a whole-genome sequence (Tuskan *et al.*, 2006) integrated with genetic and physical maps (Kelleher *et al.*, 2007). Several association studies in *P. trichocarpa*, *P. deltoides*, *P. nigra* and *P. tremula* are targeting traits related to biomass productivity, cell wall characteristics and climatic adaptation (Stanton *et al.*, 2010).

Black cottonwood (*Populus trichocarpa*), which is one of the fastest growing species of the genus, inhabits riparian areas in western North America from Baja California to Alaska (DeBell, 1990). To characterize genome-wide patterns of allele frequency variation, LD and recombination in *P. trichocarpa*, we resequenced the genomes of 16 trees (Fig. 1a, Supporting Information Table S1) selected as a 'range-wide' sample (i.e. spanning a large proportion of the species' native range). Results based on genome resequencing data were corroborated by extensive sequence and SNP data generated using traditional technologies (e.g. genotypic data for 29 213 SNPs in 120 trees from a similar geographic area, Fig. 2). In addition, we used information on DNA sequence features and data from whole-genome methylated DNA immunoprecipitation (MeDIP) resequencing to identify the strongest correlates of fine-scale recombination rates estimated using our genome resequencing data.

Materials and Methods

Plant materials and DNA extraction

We assembled a clonally replicated population of 1100 black cottonwood (*Populus trichocarpa* Torr. & Gray) genotypes that were established in multiple field trials. The vast majority of these genotypes (n = 1052) were sampled in the core of the range of *P. trichocarpa*, west of the Cascade Mountains in northern Oregon, Washington and southern British Columbia. This set of genotypes is expected to be appropriate for association mapping because *P. trichocarpa* grows best in this area (DeBell, 1990) and is characterized by high levels of molecular and phenotypic variation, with relatively weak interpopulation differentiation for neutral markers (Weber & Stettler, 1981; Weber *et al.*, 1985). An additional set of 48 genotypes (22 from northern California, 15 from the Mid-Willamette Valley in central Oregon and 11 from central British Columbia) was included to allow a preliminary assessment of levels of molecular and phenotypic variation on a broader geographic scale. To generate genome resequencing data, we selected 16 *P. trichocarpa* trees spanning the entire range of our experimental population and a large proportion of the range of *P. trichocarpa* (Fig. 1a). Five to ten micrograms of high-molecular-weight DNA was extracted from 3–5 g of roots or leaves from plants grown in the glasshouse or in hydroponic systems using a protocol available at http://my.jgi.doe.gov/general/protocols/Populus_nuclear_DNA_extraction.doc. To generate corroborating Infinium SNP data (described below), we selected 12 trees from each of 10 subpopulations, which were chosen to roughly match the geographic distribution of the 16 resequenced trees (Figs 1a, 2a), and included 10 of these trees to allow SNP genotype comparisons between the two datasets.

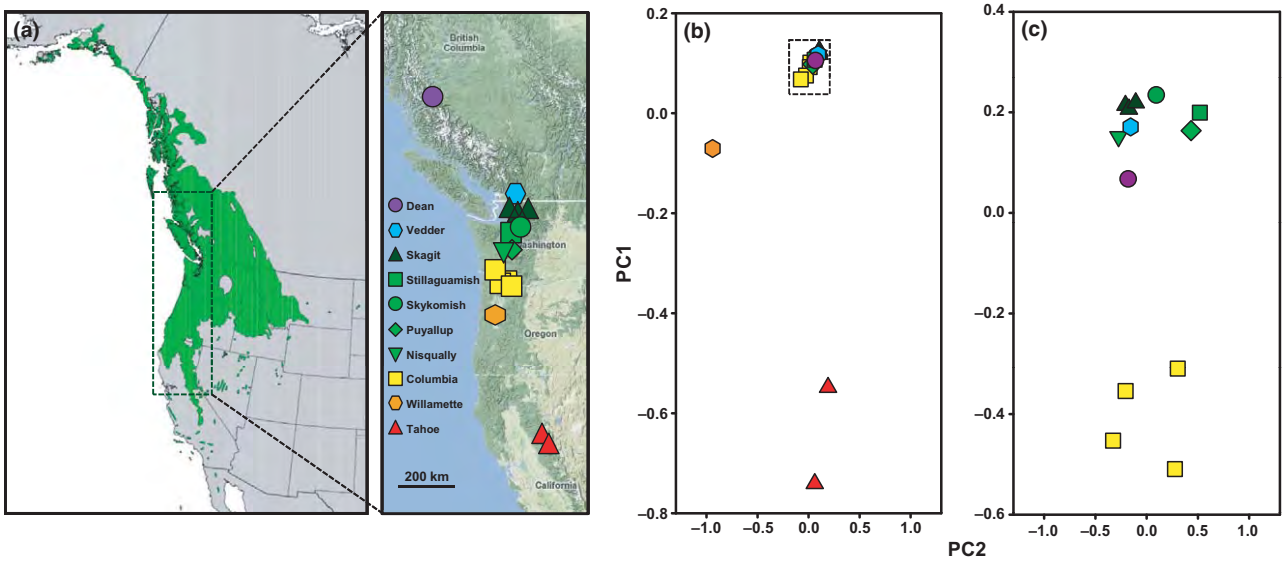

Fig. 1 Population structure in *Populus trichocarpa* based on genome resequencing data. (a) Sampling locations of the 16 *P. trichocarpa* trees that were resequenced, relative to the species range (Little, 1971). (b, c) Major axes of variation based on principal component analysis (PCA) of 235 259 single-nucleotide polymorphisms (SNPs) on a range-wide scale (b) and after excluding the clearly differentiated Tahoe and Willamette trees (c). The cluster which includes trees from multiple locations (dotted frame in b) was re-analyzed at a finer spatial scale (c).

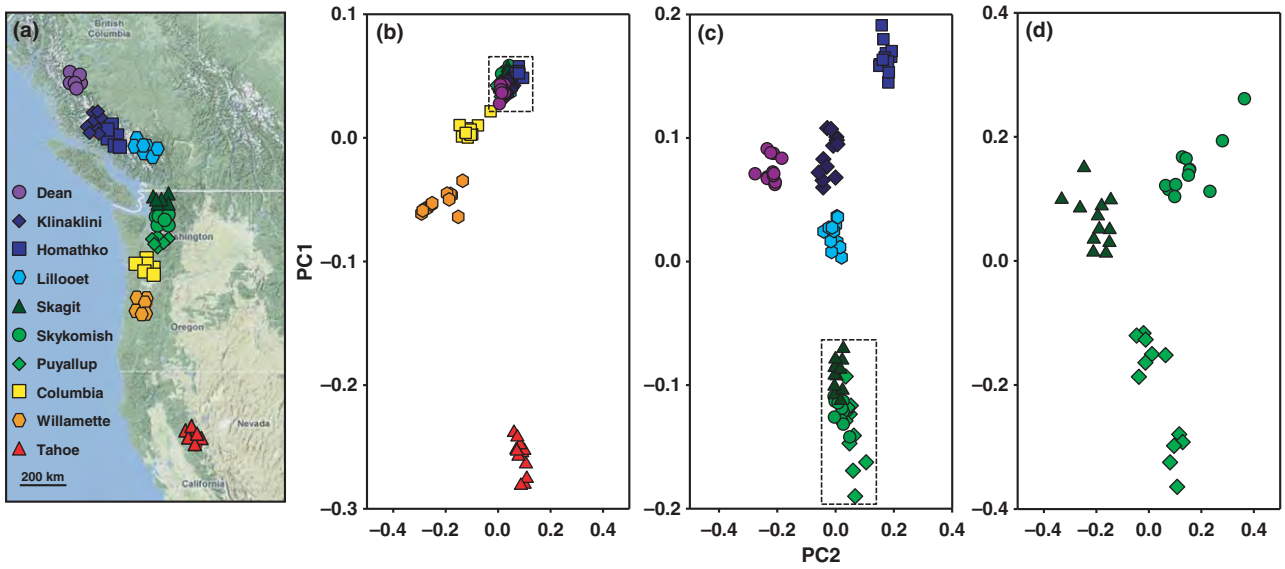

Fig. 2 Population structure in *Populus trichocarpa* based on Infinium single-nucleotide polymorphism (SNP) data. (a) Sampling locations of 120 *P. trichocarpa* trees from 10 subpopulations chosen to span the geographic distribution of the 16 resequenced trees. (b–d) Major axes of variation based on principal component analysis (PCA) of 22 280 SNPs on a range-wide scale (b), after excluding trees from the clearly differentiated Tahoe, Willamette and Columbia subpopulations (c), and within the most homogeneous set of subpopulations in western Washington (d). Clusters which include trees from multiple subpopulations (dotted frames in b and c) were re-analyzed at finer spatial scales after excluding clearly differentiated trees.

Library construction and genome resequencing

DNA was randomly sheared into small fragments (200–300 bp) using a Covaris E210 ultrasonicator (Covaris, Inc., Woburn, MA, USA), according to the manufacturer's recommendations. The overhangs created by fragmentation were converted into blunt ends using T4 DNA polymerase and DNA polymerase I Klenow fragment. Illumina adaptors (Illumina, San Diego, CA, USA) were then ligated to the DNA fragment using DNA ligase. Finally, a polymerase chain reaction (PCR) was performed using DNA Phusion Polymerase to selectively enrich those DNA fragments that have adaptor molecules on both ends and to amplify the amount of DNA in the library. Short-read sequence data were generated for each individual tree (i.e. without pooling) using an Illumina Genome Analyzer.

Short-read sequence alignment

Sequence reads were aligned to assembly v2 of the reference genome of *Nisqually-1* (http://www.phytozome.net/poplar.php; Tuskan *et al.*, 2006) using the MAQ software package (Li *et al.*, 2008) with default options, except for increasing the number of allowable mismatches in the first 24 bp to three, and the resulting SNPs were filtered using the default options of SNPFilter from the 'maq.pl' script. The rationale of this step was to potentially align as many reads as possible, but also to capture mapping quality and other alignment statistics that can be used for downstream filtering (see SNP filtering below). The 'map' alignment file from MAQ was converted to Binary Alignment/Map (BAM) format using the SAMtools package (Li *et al.*, 2009), and subsequent filtering was based on the BAM representation of the alignments (Table S1).

SNP filtering

To guide the selection of SNP filtering criteria, we generated Sanger sequence data by PCR amplification of fragments spanning 10 candidate genes (Table S2) in 47 *P. trichocarpa* trees, including 15 of the 16 trees in the genome resequencing data. Fragments were then direct-sequenced by Beckman Coulter (Beckman Coulter, Danvers, MA, USA) on an ABI 3730XL sequencer (Life Technologies Corporation, Carlsbad, CA, USA). Sequences were assembled using the Phred/Phrap (Gordon *et al.*, 1998) programs and polymorphisms were scored using PolyPhred (Nickerson *et al.*, 1997).

Biallelic SNPs that passed the initial sequence alignment filtering (see Short-read sequence alignment above) were further screened using nine additional filtering criteria (Table 1) based on sequence alignment statistics (Li *et al.*, 2008), the observed number of minor alleles, the availability of data for all trees and conformity of genotype frequencies to Hardy–Weinberg expectations. We searched a large parameter space ($n > 10^6$ scenarios) and evaluated each set of filtering criteria by comparing the set of SNPs obtained after filtering the genome resequencing data with a set obtained from Sanger resequencing data for the same trees (Table S2). Based on the tradeoff between the estimated rates of false positives and false negatives with respect to SNP detection, we selected two sets of filtering criteria (Table 1), both of which included only SNPs that were genotyped in all 16 resequenced trees and had minor allele frequencies (MAFs) of at least 0.094 (i.e. 3/32). The Hardy–Weinberg equilibrium (HWE) set included relatively relaxed alignment criteria, but required genotype frequencies to be roughly concordant with Hardy–Weinberg proportions (i.e. $F_{IS} \leq |0.4|$) and observing at least one homozygous genotype for the minor allele. F_{IS} was calculated as

Table 1 Filtering criteria for genome resequencing data from 16 *Populus trichocarpa* trees

Filtering criteria	HWE	QS
Q[a]	10	30
Min depth[b]	5	5
Max depth[c]	275	NA
Min hit rate[d]	0.8	0.7
Max hit rate[e]	1.6	1.0
Missing[f]	0	0
Minor alleles[g]	3	3
Min homozygotes[h]	1	NA
Max $\|F_{IS}\|$[i]	0.4	NA
Statistics		
False positives (%)[j]	0	17
False negatives (%)[k]	79	54
MAF[l]	0.27	0.22
No. of loci[m]	456 079	1 453 752

HWE, Hardy–Weinberg equilibrium; QS, Quality Score.
[a]Q, minimum consensus genotype quality score (Li *et al.*, 2008).
[b]Min depth, minimum number of reads.
[c]Max depth, maximum number of reads.
[d]Min hit rate, minimum approximate copy number of the sequence flanking a single-nucleotide polymorphism (SNP) (Li *et al.*, 2008).
[e]Max hit rate, maximum approximate copy number of the sequence flanking an SNP (Li *et al.*, 2008).
[f]Missing, maximum number of trees for which missing data are allowed.
[g]Minor alleles, minimum number of copies of the minor allele.
[h]Min homozygotes, minimum number of homozygous genotypes for the minor allele.
[i]Max $|F_{IS}|$, maximum deviation of observed genotype frequencies from Hardy–Weinberg expectations.
[j]False positives, percentage of filtered SNPs that were not found in Sanger sequencing data for the same trees (Table S2).
[k]False negatives, percentage of SNPs that were found in Sanger sequencing data for the same trees, but were not detected in the Illumina data or did not pass one or more of the filtering criteria.
[l]MAF, average minor allele frequency. For comparison, the average MAF for common SNPs (MAF ≥ 0.10) genotyped through Sanger resequencing (Table S2) was 0.24 and that for the Infinium SNP data was 0.27.
[m]No. of loci, number of SNPs that passed all filtering criteria.

$1 - H_o/H_e$, where H_o and H_e were the observed and expected heterozygosity based on genotypes from the core of the range of *P. trichocarpa* (i.e. excluding trees from the strongly differentiated Tahoe and Willamette subpopulations; see Results). We did not use a more stringent threshold value of F_{IS} in order to avoid the elimination of high-quality SNPs with genotype frequencies deviating from HWE because of sampling variance or because they reflected population substructure within the core of the range of *P. trichocarpa* (see the Results section). In contrast, the Quality Score (QS) set relied on stricter alignment criteria, but did not use any information on genotype frequencies.

Infinium SNP data

To verify that the patterns observed from analyses of the genome resequencing data were not a sampling artifact caused by the relatively small number of resequenced trees, we genotyped 120 trees for 29 213 SNPs using the methods described in Supporting Information Methods S1. Briefly, we used information from previous studies to identify 3704 candidate genes that are being targeted in ongoing association studies of traits related to cell wall characteristics and climatic adaptation. After combining our genome resequencing data with transcriptome resequencing data from developing xylem of 20 *P. trichocarpa* trees (Geraldes *et al.*, 2011), we identified 169 626 potential target SNPs within or near these candidate genes. We then used information on pairwise tagging, design scores and assay types for Infinium iSelect HD Custom Genotyping (Illumina), SNP annotation and spacing to select 38 000 target SNPs. Infinium assays were successfully developed for 34 131 SNPs, 29 213 of which were successfully genotyped.

Population structure and allele frequency gradients

We performed principal component analyses (PCAs) at multiple scales using v. 3.0 of the EIGENSOFT package (Patterson *et al.*, 2006) after removing one SNP from each pair of loci located within 10 kb of one another and linked at $r^2 \geq 0.8$. This was done in order to avoid artifacts caused by large blocks of tightly linked markers (Patterson *et al.*, 2006; Nelson *et al.*, 2008). The statistical significance of the relationship between geographic and genetic (F_{ST}; Wright, 1965; Weir & Cockerham, 1984) distances among the 10 subpopulations represented in the Infinium SNP data was assessed through Mantel tests (10 000 permutations) using GenAlEx (Peakall & Smouse, 2006). We quantified allele frequency gradients in the genome resequencing data by calculating Pearson's correlation coefficients between the source latitude and the number of copies of an arbitrarily chosen allele in each tree. To assess the robustness of these measures, we used the Infinium SNP data to calculate Pearson's correlation coefficients between latitudes and allele frequencies in the 10 subpopulations.

LD and recombination

For genome resequencing data, we estimated haplotype frequencies using version 2.1 of the PHASE program (Stephens *et al.*, 2001; Li & Stephens, 2003; Stephens & Scheet, 2005) and calculated r^2 (Slatkin, 2008) for all pairs of SNPs in windows consisting of a fixed number of SNPs. To verify that estimates of r^2 were not biased by inaccurate haplotype frequency estimation, values of r^2 obtained as described above were compared with those calculated using genotypic correlations (Marchini *et al.*, 2006). For data filtered using the HWE filtering criteria (Table 1), windows consisted of 50 consecutive SNPs (median size, 35.5 kb), with adjacent windows overlapping by 20 SNPs. For data filtered using the QS filtering criteria (Table 1), we employed 200-SNP windows (median size, 43.6 kb) overlapping by 20 SNPs. To quantify LD for pairs of unlinked loci, r^2 was estimated through genotypic correlations for 10^7 randomly selected pairs of SNPs located on different chromosomes. For comparisons of LD between different datasets (e.g. core of range vs range-wide), mean values of r^2 in each distance class were corrected for small sample size by subtracting $1/n$, where n is the number of chromosomes used to estimate r^2 (Tenesa *et al.*, 2007). The same

approach was used for the Infinium SNP data, with windows corresponding to candidate genes and the 2-kb flanking regions up- and downstream from them.

We obtained estimates of the scaled recombination rate ($\rho = 4N_e c$) for each 50- or 200-SNP window using the default options of PHASE (i.e. using the –MR0 model). Based on the haplotype frequencies estimated as part of this analysis, we also calculated the nucleotide diversity (π; Nei & Miller, 1990) for each window. We then calculated 100-kb values of ρ and π as weighted averages of the values for the 50- or 200-SNP windows overlapping with each 100-kb window (i.e. using the lengths of overlap as weights). Based on information on correlates of recombination in other organisms (Myers *et al.*, 2005; Drouaud *et al.*, 2006; Kim *et al.*, 2007; Liu *et al.*, 2009), we assessed 17 potential explanatory variables reflecting: the position along chromosomes relative to putative centromeres and telomeres; nucleotide diversity; DNA sequence features (e.g. GC content, occurrence of CpG-rich regions, repeat elements and genes); and epigenetic patterns revealed using whole-genome MeDIP resequencing (Methods S2). Estimates of effective population size (N_e) were obtained by relating values of r^2 or estimates of ρ to direct estimates of local recombination rates (c) obtained from genetic linkage maps (Methods S3).

Recombination hotspots were identified using the –MR1 option of PHASE, with all other parameters set to their default values. In order to identify robust recombination hotspots, we ran PHASE on the HWE-filtered SNP data three times, setting the seed of the pseudorandom number generator to a different number for each run, and combined results with a run based on SNPs filtered using the QS criteria (Table 1). Subsequent analyses were performed only for putative hotspots that consistently (i.e. in results from all runs) had: a median intensity (λ) of at least 20, a posterior probability that $\lambda > 10$ of at least 0.9 (Crawford *et al.*, 2004) and border discrepancies of no more than 5 kb across runs. We also eliminated hotspots detected on scaffolds that were not mapped to chromosomes and in windows that were longer than 100 kb or had < 50 SNPs (i.e. at the ends of chromosomes). All of these criteria were satisfied for 606 hotspots, whose median Bayes factor (Crawford *et al.*, 2004) was 293 (range, 27–1468). For each of these hotspots, we attempted to identify a matching 'coldspot' (Myers *et al.*, 2005) in a window that was located within 500 kb of the window in which the respective hotspot was detected, had $\lambda < 10$ and a posterior probability that $\lambda > 10$ of at most 0.5 across all runs of PHASE and was no longer than 100 kb and contained no fewer than 50 SNPs. We identified 589 coldspot windows that met these criteria and randomly selected the start and end coordinates of coldspots within these windows, matching the lengths of coldspots to their respective hotspots. To test whether hotspot locations relative to genes followed a nonrandom pattern, we repeated the random selection of coldspot borders 10 000 times and calculated empirical one-sided P values by comparing the observed distribution of distances between hotspots and genes to the one of the simulated coldspots. We compared DNA sequence and methylation characteristics between hotspots and coldspots through paired t-tests (Wackerly *et al.*, 2002) using logarithmic transformations when diagnostic box plots suggested widely unequal variances or severe departures from normality of the data. We tested for over- or under-representation of different types of repeat elements by comparing the proportional occurrence of each type of repeat element relative to all repeat elements identified in hotspots to that in coldspots using contingency table χ^2 tests (Wackerly *et al.*, 2002).

Results

SNP discovery and genotyping

Resequencing was based on 6.4 billion Illumina Genome Analyzer reads with an average length of 59 bp (379 Gb of sequence). Approximately 84% of these reads mapped to the reference genome of *Nisqually-1* (Tuskan *et al.*, 2006), covering 95% of its length to an average depth of 39× (Table S1). The HWE set of SNP filtering criteria produced no false positives (Table 1) and resulted in genotype calls that matched those based on the Sanger data 96% of the time and those from the Infinium SNP data 97% of the time. Furthermore, 98% of the heterozygous SNP calls for *Nisqually-1* based on genome resequencing data filtered using the HWE criteria were supported by Sanger reads used to generate the reference genome sequence (Tuskan *et al.*, 2006). Because of the high reliability of SNP detection and genotype calls obtained using the HWE criteria, we used them to filter the genome resequencing data for all subsequent analyses. However, the HWE criteria also resulted in a high rate of false negatives (79%). In contrast, the QS set of filtering criteria had a high rate of false positives (17%), but also a substantially lower rate of false negatives (54%), resulting in genome coverage which was approximately three times more dense (Table 1). This set was used to assess the robustness of our results to varying SNP density. In addition, we employed the Infinium SNP data (120 trees genotyped for 29 213 SNPs) to verify all patterns detected using genome resequencing data. Average MAFs for common SNPs (MAF ≥ *c.* 0.10) were 0.27, 0.22, 0.24 and 0.27 for genome resequencing data filtered using the HWE and QS criteria, the Sanger resequencing data and the Infinium SNP data, respectively.

Population structure and allele frequency gradients

PCAs of SNP genotypes based on genome resequencing and the Infinium array revealed clear patterns of geographic structure at multiple spatial scales (Figs 1, 2). The primary axes of variation (PC1) were highly significant ($P \leq 0.0005$), explained between 2.9% (Fig. 2c) and 18.2% (Fig. 1b) of the total variance and were associated with the latitudinal origins of the resequenced trees at all of these spatial scales. For example, correlations between PC1 scores and source latitudes were $r = 0.87$, 0.88 and 0.71, respectively, for analyses presented in Fig. 2(b–d) ($P < 10^{-5}$). Similar patterns were detected using model-based clustering of trees based on a subset of the Infinium SNP data

(Supporting Information Fig. S1). Consistent with the spatial pattern of genetic differentiation revealed by PCA, genetic and geographic distances among subpopulations were linearly associated (Fig. S2, Table S3).

Quantitative measures of latitudinal allele frequency gradients based on genome resequencing and Infinium SNP data were strongly correlated ($r = 0.76$, $P < 10^{-15}$), verifying that allele frequency patterns detected using the genome resequencing data were robust. Analyses of *c.* 450 000 SNPs from the genome resequencing data revealed that latitudinal allele frequency gradients were: strikingly common at the range-wide scale (*c.* 23%, Fig. 3a), and less common, but still substantially more frequent than expected by chance (*c.* 9%, Fig. 3b), within the core of the range (i.e. excluding trees from the strongly differentiated Tahoe and Willamette subpopulations). The relative frequencies of latitudinal gradients across different categories of SNPs also differed between these two spatial scales. At the range-wide scale, clinal allele frequency variation was more common among intergenic than among genic SNPs (Fig. 3a). In contrast, latitudinal allele frequency gradients within the core of the range of *P. trichocarpa* were more common among genic than among intergenic SNPs (Fig. 3b). Interestingly, the frequency of latitudinal gradients at both spatial scales was similar in SNPs located in untranslated regions, introns and coding sequences.

The observed heterozygosities of the two resequenced trees from the Tahoe subpopulation ($H_o = 0.226$ and 0.263 based on the HWE filtering criteria) were substantially lower than those of the trees sampled from the core of the range (average $H_o = 0.333$, SD = 0.024). The same pattern was detected using the Infinium SNP data ($H_o = 0.219$ for the Tahoe subpopulation vs average $H_o = 0.314$, SD = 0.011, across subpopulations in the core of the range). After excluding the Tahoe subpopulation, however, the observed heterozygosity and nucleotide diversity were not significantly correlated with latitude ($P > 0.40$).

LD and recombination

Values of r^2 obtained based on the estimated haplotype frequencies were concordant with those calculated using genotypic correlations (Fig. S3), and were therefore used for all subsequent analyses, except for the quantification of LD among physically unlinked loci (see next paragraph). The genome-wide average r^2 dropped below 0.2 within 3–6 kb, depending on the dataset and filtering criteria used (Figs 4, 5). Filtering of the genome resequencing data using the stringent HWE set of criteria (Table 1) resulted in estimates that were very similar to those obtained based on the Sanger resequencing data, and slightly higher, but consistent, with those obtained based on the Infinium SNP data (Fig. 5). In contrast, estimates based on the more relaxed QS set of criteria appeared to be downwardly biased (Fig. 5).

Population structure had a two-fold effect on LD. First, r^2 among physically linked loci decayed at a slightly faster rate (i.e. at *c.* 500-bp shorter distances) when all trees were analyzed as one homogeneous population than when strongly differentiated trees were excluded (Fig. S4). Consistent with this, the estimate of effective population size from LD was 24% higher using the range-wide sample ($N_e = 5592$) than that using only the data from trees sampled in the core of the range ($N_e = 4506$, $P < 10^{-12}$ from a paired *t*-test across chromosomes) (Fig. S4). Furthermore, rates of decay of r^2 with physical distance and LD-based estimates of N_e varied dramatically among subpopulations, with N_e estimates ranging from 193 (Tahoe) to 2278 (Willamette) (Fig. 6). Second, in contrast with its effect on local LD, population subdivision caused elevated levels of genome-wide LD among physically unlinked loci (Fig. S5).

Fine-scale rates of recombination estimated indirectly from the resequencing data varied by orders of magnitude across the genome (Figs 7, S6) and were correlated with recombination

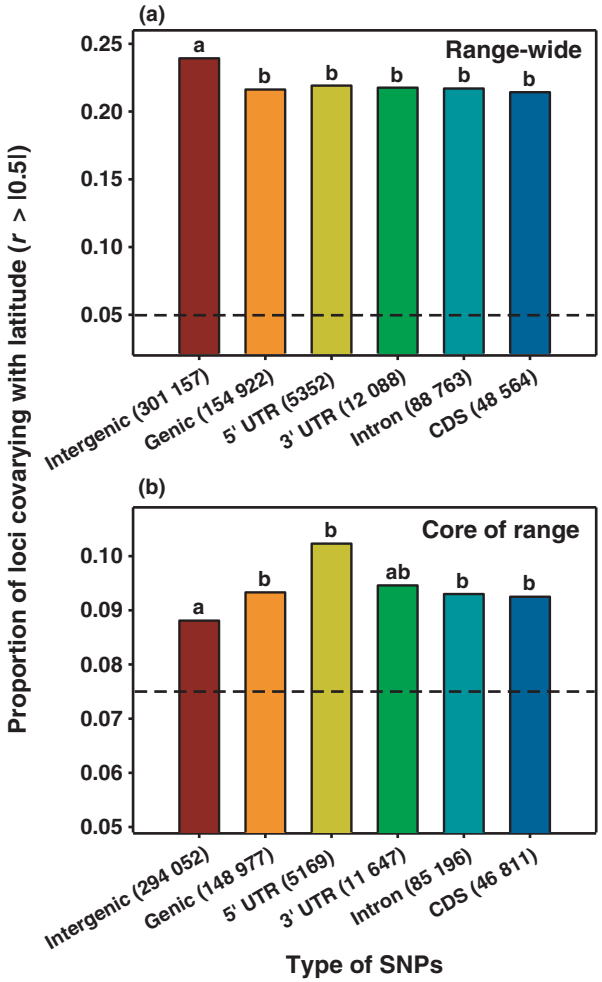

Fig. 3 Latitudinal allele frequency gradients in *Populus trichocarpa* based on genome resequencing data. Proportions of single-nucleotide polymorphisms (SNPs) that covary moderately or strongly with latitude (i.e. Pearson's $|r| > 0.50$). (a) 456 079 SNPs that passed Hardy–Weinberg equilibrium (HWE) filtering based on data for all trees ($n = 16$). (b) 443 029 SNPs that passed HWE filtering based on data for 13 trees, excluding the strongly differentiated trees sampled from the Tahoe and Willamette subpopulations (Figs 1, 2). In each graph, bars annotated with different letters correspond to significantly different proportions ($P < 0.002$ based on contingency table χ^2 tests). Dashed lines indicate the proportions of SNPs expected to covary with latitude (i.e. $|r| > 0.50$) based on 1000 permutations of SNP genotypes and source latitudes. The number of loci in each category is shown in parentheses. CDS, coding DNA sequence; UTR, untranslated region.

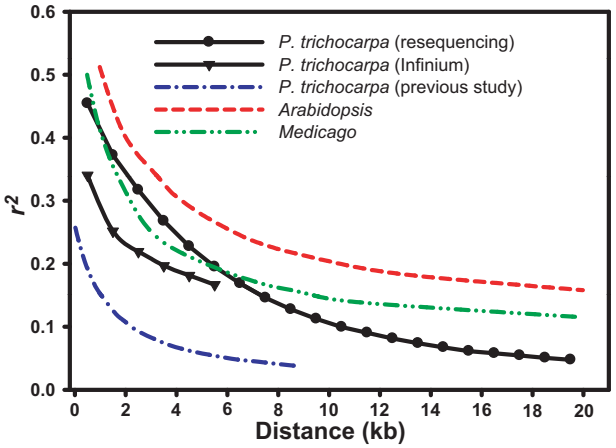

Fig. 4 Genome-wide linkage disequilibrium (LD) in *Populus trichocarpa* based on common single-nucleotide polymorphisms (SNPs) (i.e. minor allele frequency (MAF) ≥ *c.* 0.10). Black circles and triangles correspond to average values of r^2 for interlocus distances grouped in 1-kb bins based on analyses of genome resequencing (n = 16 trees, MAF ≥ 0.094) and analyses of Infinium SNP data for 3704 candidate genes (n = 120 trees, MAF ≥ 0.10), respectively. Data from a previous small-scale study in *P. trichocarpa* (Wegrzyn *et al.*, 2010) and for the primarily self-pollinating plants *Arabidopsis thaliana* (Kim *et al.*, 2007) and *Medicago truncatula* (Branca *et al.*, 2011) are illustrated using dashed and/or dotted lines.

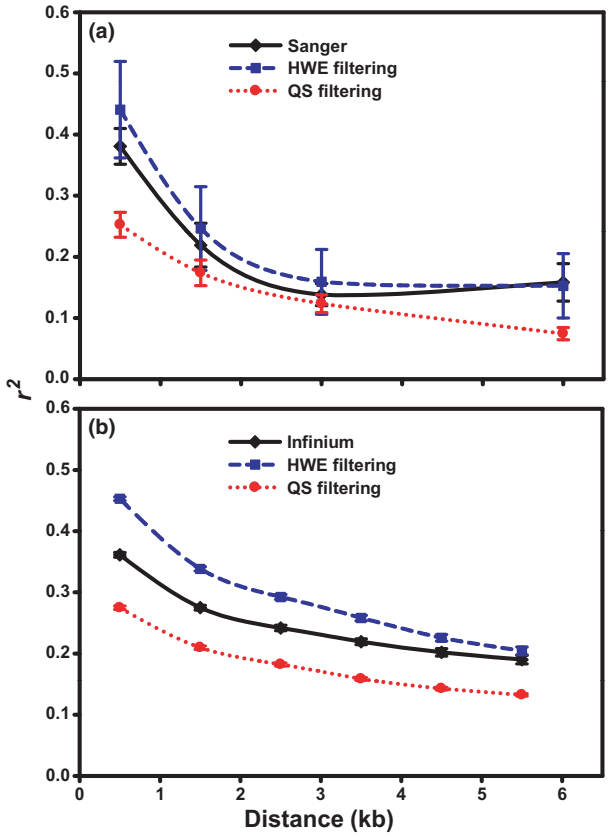

Fig. 5 Effect of filtering criteria on estimates of linkage disequilibrium (LD; r^2) based on genome resequencing data relative to estimates based on Sanger resequencing data (a) and Infinium single-nucleotide polymorphism (SNP) data (b). All estimates are based on data for trees from the core of the range of *Populus trichocarpa* (i.e. excluding the strongly differentiated trees from the Tahoe and Willamette subpopulations). Error bars correspond to 95% confidence intervals based on 1000 bootstrap samples. (a) Average values of r^2 for SNPs located inside or within 2 kb of 10 candidate genes resequenced using Sanger technology (Table S2, n = 40 trees, minor allele frequency (MAF) ≥ 0.10, 1243 SNP pairs) and based on SNPs from the same regions that were genotyped by genome resequencing (n = 13 trees, MAF ≥ 0.115) and filtered using the Hardy–Weinberg equilibrium (HWE; 304 SNP pairs) and Quality Score (QS; 3760 SNP pairs) sets of criteria (Table 1). Pairwise r^2 values were grouped into 0–1, 1–2, 2–4 and 4–8 kb distance classes to ensure that each class contained at least 50 values in each dataset. (b) Average values of r^2 in 1-kb distance classes for loci inside or within 2 kb of 3704 candidate genes represented on the Infinium genotyping array (n = 96 trees, MAF ≥ 0.10, 99 025 SNP pairs) and based on SNPs from the same regions that were genotyped by genome resequencing (n = 13 trees, MAF ≥ 0.115) and filtered using the HWE (170 186 SNP pairs) and QS (1 801 953 SNP pairs) sets of criteria.

rates estimated directly from genetic maps (Spearman's ρ = 0.50, $P < 10^{-10}$) (Fig. S7). As expected, several chromosomal, DNA sequence and epigenetic features were moderately correlated with recombination rates, as well as with one another (Fig. 7, Table S4). The overall patterns were consistent across analyses based on 100-kb and 1-Mb windows, but correlations tended to be stronger for 1-Mb windows (Table S5). Multicollinearity made the identification of a single most appropriate set of predictors challenging, but a forward selection procedure with a Bonferroni adjustment to control for the large number of potential explanatory variables resulted in models with R^2 = 0.65 and 0.77 for 100-kb and 1-Mb windows, respectively (Tables S6, S7). Models including only the three strongest predictors (the number of MeDIP reads and the occurrence of long terminal repeat (LTR) Gypsy retrotransposons and A-, T- or AT-rich low-complexity sequences; Tables S4, S5) explained between one-half and three-quarters of the genome-wide variation in log-transformed recombination rates (R^2 = 0.53 and 0.76 for 100-kb and 1-Mb windows, respectively).

We identified over 6000 putative recombination hotspots with an average intensity (i.e. fold change relative to background rates of recombination) of 261 (range, 20–5918). Approximately 10% of these hotspots (n = 606) passed stringent screening, including detection using both HWE and QS filtering criteria (Table 1). The distribution and sequence characteristics of these hotspots were compared with those of control regions (i.e. 'coldspots'; Myers *et al.*, 2005), which were chosen as close as possible to each hotspot (i.e. the median distance between hotspots and coldspots was *c.* 74 kb). Recombination hotspots tended to be located closer to genes than expected by chance, preferentially outside of genes and preferentially upstream of genes ($P < 10^{-4}$ for all three tests; Table 2). As expected on the basis of the correlates of recombination identified (Tables S4–S7), DNA in hotspots tended to be less methylated ($P < 10^{-15}$), contained fewer LTR Gypsy retrotransposons ($P < 10^{-15}$), contained more AT-rich low-complexity sequences ($P < 10^{-15}$) and had a lower GC content ($P < 10^{-15}$), but a larger number of bases in CpG-rich regions ($P < 10^{-6}$), compared with DNA in coldspots (Table 2).

Discussion

The high reliability of SNP detection and genotyping based on our genome resequencing data was illustrated through

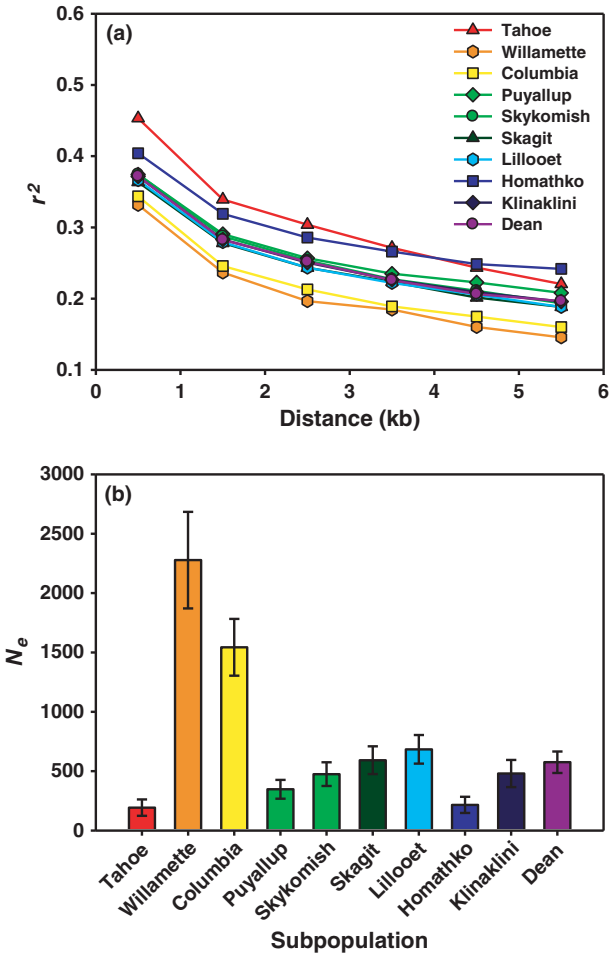

Fig. 6 Variation of linkage disequilibrium (LD; r^2) and LD-based estimates of effective population size (N_e) across 10 subpopulations of *Populus trichocarpa*. (a) Average values of r^2 for interlocus distances grouped in 1-kb bins based on analyses of Infinium single-nucleotide polymorphism (SNP) data (minor allele frequency (MAF) ≥ 0.10) for 12 trees in each subpopulation. (b) Estimates of N_e based on relating estimates of the scaled recombination rate ($4N_ec$) from the Infinium SNP data to direct estimates of recombination from a dense SNP linkage map. Subpopulations are ordered by latitude from south (left) to north (right).

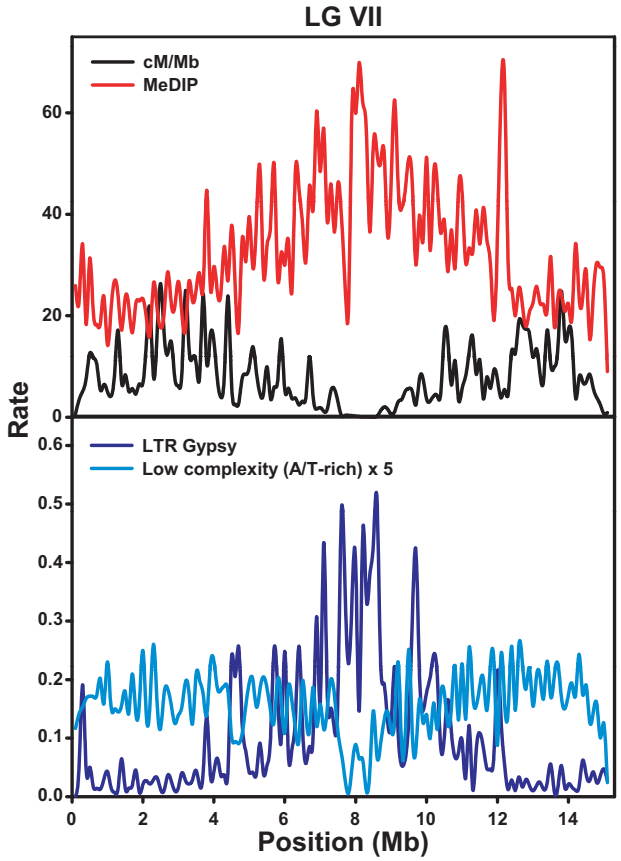

Fig. 7 Correlates of recombination in *Populus trichocarpa*. Chromosome-wide patterns of variation of: recombination rates (cM/Mb) estimated from linkage disequilibrium (LD); relative degree of methylation (methylated DNA immunoprecipitation, MeDIP), measured as the number of MeDIP reads per kilobase of target sequence per million reads mapped (Supporting Information Methods S2); the proportion of bases in long terminal repeat (LTR) Gypsy retrotransposons; and the proportion of bases in A-, T- or AT-rich low-complexity sequences (multiplied by five to equalize scales). All data are summarized in 100-kb windows. Values of $4N_ec$ were converted to recombination frequencies (cM/Mb) using the LD-based estimate of N_e for Linkage Group VII (Fig. S4).

comparisons with Sanger data (e.g. Table 1, comparisons with *Nisqually-1* reference genome sequence), consistency with transcriptome resequencing data used for SNP discovery (Geraldes *et al.*, 2011; Methods S1) and transferability to the Infinium SNP genotyping assay (Methods S1). This suggests that the prospects of high-density genotyping in *Populus* are excellent, especially considering the rapidly decreasing cost and increasing throughput of sequencing technologies (Metzker, 2010). However, our results also highlight the importance of filtering genome resequencing data appropriately (Table 1). Naïve filtering based entirely on sequence depth is likely to result in high rates of false positives and may influence downstream data analyses and interpretation (Table 1). However, the stringent filtering criteria required to achieve low rates of false positives limited our inference to common SNPs (MAF ≥ *c.* 0.10), prevented us from calculating reliable absolute estimates of nucleotide diversity (i.e. because of the consistently high rates of false negatives across all filtering scenarios considered) and possibly resulted in biased genome sampling. Further technological and bioinformatic advances will probably help to mitigate these limitations. For example, the availability of resequencing data from a larger number of individuals will allow the inclusion of lower frequency SNPs (i.e. even after requiring the detection of multiple minor alleles) in population genetic analyses, association studies and genomic selection.

Despite the great potential for long-distance seed and pollen dispersal in *P. trichocarpa* (Slavov *et al.*, 2009; DiFazio *et al.*, 2012), population genetic structure appears to be present from the range-wide (Figs 1, 2) to the local stand (Slavov *et al.*, 2010) scale. The genetic differentiation of trees sampled from the Tahoe subpopulation (pairwise $F_{ST} > 0.151$; Table S3) was stronger than expected based on previous population genetic studies in *Populus* (Slavov & Zhelev, 2010), but comparable with that observed in a range-wide sample of *P. balsamifera* (Keller *et al.*, 2010), a close relative of *P. trichocarpa*. Even the subtler genetic

Table 2 Contrasts between 589 recombination hotspots and their matched coldspots (*P* is the two-sided *P* value, unless indicated otherwise)

Characteristic	Hotspots	Coldspots	*P*
Location relative to genes			
Number overlapping with genes	240	285[a]	$< 10^{-4}$[b]
Number upstream of genes	222	180[a]	$< 10^{-4}$[b]
Number downstream of genes	127	124[a]	0.37[b]
Average distance upstream of genes (bp)	1974	3278[a]	$< 10^{-4}$[b]
Average distance downstream of genes (bp)	1654	3034[a]	$< 10^{-4}$[b]
DNA characteristics and methylation			
Average GC content (%)	30.8	33.1	$< 10^{-15}$
Average number of CpG bases	92	83	5.1×10^{-7}
Average number of MeDIP reads per kb of sequence	677	1267	$< 10^{-15}$
DNA repeat composition[c]			
Proportion of LTR Gypsy retrotransposons (%)	5.9	12.3	$< 10^{-15}$
Proportion of AT-rich low-complexity sequences (%)	22.7	13.9	$< 10^{-15}$
Proportion of LINEs (%)	0.3	0.9	0.0002

[a]Average across 10 000 simulated coldspot locations.
[b]One-sided *P* values.
[c]Proportional occurrence of each type of repeat element relative to all repeat elements identified in hotspots or coldspots.

differentiation observed in the core of the range of *P. trichocarpa* (pairwise F_{ST} = 0.013–0.048; Table S3) needs to be statistically accounted for in association studies (Price *et al.*, 2006, 2010).

The observed continuous patterns of differentiation and allele frequency variation (Figs 1, 2) match those expected under population genetic models of isolation by distance (IBD) (Rousset, 1997). For example, our results are consistent with both hypotheses proposed by Soltis *et al.* (1997) to explain the phylogeographic patterns detected in plants growing in the Pacific Northwest of North America. However, the lack of a clear trend of decreasing diversity with latitude makes a scenario of recolonization from multiple glacial refugia (i.e. 'north–south recolonization hypothesis') more plausible than a scenario of recolonization from a single group of refugia located in the mountains of northern California and southern Oregon (i.e. 'leading edge hypothesis'). Regardless of the specifics of the underlying demographic scenario, the remarkable abundance and ubiquity of latitudinal SNP gradients at the range-wide scale (i.e. about one of four SNPs with MAF ≥ *c.* 0.10), and their higher frequency among intergenic than among genic SNPs, suggest that allele frequency gradients caused by local adaptation may be completely confounded with neutral differentiation under IBD (Vasemägi, 2006; Novembre & Di Rienzo, 2009). The reversed pattern of higher frequency of latitudinal gradients among genic than intergenic SNPs in the core of the range indicates that spatial patterns of SNP variation could potentially be instrumental for the detection of molecular signatures of selection. However, performing this robustly will probably require the use of appropriately defined study populations and neutral differentiation models consistent with IBD, as well as conservatively combining multiple sources of evidence (e.g. allele frequency differentiation, local levels of LD and nucleotide diversity vs divergence, and allele frequency spectra for ancestral vs derived alleles and haplotypes) (Grossman *et al.*, 2010; Hernandez *et al.*, 2011).

The decay of LD (r^2) with physical distance between loci was substantially slower than expected based on previous studies in *P. trichocarpa* (Gilchrist *et al.*, 2006; Wegrzyn *et al.*, 2010) and other forest trees (Krutovsky & Neale, 2005; Heuertz *et al.*, 2006; Ingvarsson, 2008; Neale & Ingvarsson, 2008). Even more strikingly, the observed extent of LD was within the same order of magnitude as that in the predominantly selfing plants *Arabidopsis thaliana* (Kim *et al.*, 2007) and *Medicago truncatula* (Branca *et al.*, 2011) (Fig. 4). What factors could explain this unexpected result?

First, estimates of r^2 are upwardly biased in small samples (Terwilliger & Hiekkalinna, 2006) and very sensitive to the distribution of allele frequencies across loci (VanLiere & Rosenberg, 2008). However, both the number of trees used to generate the genome resequencing data (n = 16 trees or $2n$ = 32 chromosomes) and the MAF thresholds used (≥ *c.* 0.10) were very similar or identical to those employed in previous studies in *Populus* and other forest trees (Krutovsky & Neale, 2005; González-Martínez *et al.*, 2006a; Heuertz *et al.*, 2006; Ingvarsson, 2008; Wegrzyn *et al.*, 2010), as well as those used in the highly selfing plants employed for reference in Fig. 4 (Kim *et al.*, 2007; Branca *et al.*, 2011). Thus, neither of these factors seems to be a likely explanation for the unexpectedly slow decay of LD in *P. trichocarpa*.

Second, estimates of LD decay were affected by the criteria used to filter the genome resequencing data (Fig. 5). Estimates based on the stringent HWE criteria were very similar to those based on Sanger resequencing data (Fig. 5a) and roughly comparable to those based on Infinium SNP data (Fig. 5b). The slight difference in the latter case probably resulted from the much larger sample size used to generate the Infinium SNP data (i.e. $2n$ = 192 chromosomes sampled in the core of the range, making the small sample bias in this dataset negligible) and/or the exclusion of nearly 50 000 strongly linked (i.e. $r^2 \geq 0.8$) SNPs through the pairwise tagging analyses performed to select target loci for Infinium genotyping (Methods S1). In contrast, analyses of genome resequencing data filtered using the more relaxed QS criteria appeared to systematically result in underestimates of the

mean r^2 values by distance class (Fig. 5). This is probably caused primarily by the high frequency of false-positive SNPs (i.e. which would generally be expected to behave as unlinked) in data filtered using these criteria (Table 1). However, even the decay of r^2 based on data filtered using the QS criteria was an order of magnitude slower than previously estimated for *Populus*. Taken together, the comparisons among datasets generated using different SNP genotyping assays indicate that the extensive genome-wide LD detected was probably not caused by technological artifacts.

Third, at 100-kb and 1-Mb window scales, rates of recombination were weakly to moderately correlated with gene density (Tables S4, S5), and recombination hotspots were located closer to genes than expected by chance (Table 2). These trends suggest that LD within or near genes may be weaker than the genome average. Consistent with this, estimates of r^2 based exclusively on SNPs located within or near genes (Infinium SNP data) were lower than those based on the genome resequencing data (Fig. 4), but the difference was relatively small.

Finally, LD estimates based on our genome resequencing data (i.e. with one to three SNPs per kilobase; Table 1) may not be directly comparable to those from previous studies in *P. trichocarpa* and other forest trees, which were typically based on up to several kilobases of sequence for a relatively small number of candidate genes (Krutovsky & Neale, 2005; González-Martínez *et al.*, 2006a; Heuertz *et al.*, 2006; Ingvarsson, 2008; Wegrzyn *et al.*, 2010). We hypothesize that data from these studies may not have been sufficiently extensive to develop accurate expectations on genome-wide patterns of LD. Consistent with this, analyses of resequencing data for 372 randomly sampled gene fragments in *P. trichocarpa*'s close relative *P. balsamifera* resulted in dramatically variable LD estimates that were, on average, higher than previously reported for other *Populus* species (Olson *et al.*, 2010). More importantly, estimates of LD based exclusively on very closely spaced SNPs may not be appropriate for the prediction of longer range LD (Kim *et al.*, 2007). This is presumably because gene conversion may have a substantial effect on short-range but not on long-range LD (Andolfatto & Nordborg, 1998), and because LD estimates at different distances reflect N_e over different time scales (i.e. longer range LD is expected to reflect more recent N_e; Tenesa *et al.*, 2007). The simple regression model that is commonly used to summarize the decay of r^2 with physical distance (Hill & Weir, 1988) does not account for these two factors, and predictions of long-range LD based on fitting this model to short-range LD data should be treated with caution.

As a reflection of the surprisingly slow decay of r^2 with physical distance between loci, estimates of N_e from LD (i.e. *c.* 4000–6000, depending on the genetic map and estimation method used) are remarkably low, considering the vast census sizes of the *P. trichocarpa* populations sampled. These N_e estimates are also over 20 times lower than those obtained for European aspen (*P. tremula*) based on putatively neutral nucleotide diversity (Ingvarsson, 2008), and three to four times lower than those that would be obtained for *P. trichocarpa* based on similar assumptions and estimates of neutral nucleotide diversity from previous studies (Gilchrist *et al.*, 2006; Tuskan *et al.*, 2006). The dramatic contrast between N_e estimates for *P. tremula* and *P. trichocarpa* may be a reflection of the substantial differences in life history between aspens and cottonwoods (e.g. greater propensity for asexual reproduction in aspens, possibly resulting in a 'Meselson effect'; Balloux *et al.*, 2003; Slavov & Zhelev, 2010). Furthermore, because LD-based estimates of N_e presumably reflect more recent population history than those from nucleotide diversity (Tenesa *et al.*, 2007), and because *P. trichocarpa* populations are likely to have experienced severe bottlenecks and/or founder effects over the last few hundred generations (Soltis *et al.*, 1997), our low estimates of N_e from LD are not implausible. Similar to the geographic patterns of allele frequency variation and heterozygosity, N_e estimates by subpopulation (Fig. 6) are consistent with a scenario of recolonization from multiple glacial refugia. These estimates also indicate that the Willamette and Columbia subpopulations may be located in a 'melting pot' (Petit *et al.*, 2003) of haplotype diversity from multiple refugia.

In addition to its dramatic variation among subpopulations, LD also varied substantially across the genome. The epigenetic and DNA sequence variables identified as best predictors of historical rates of recombination ($4N_ec$) were similar to those in other plants (Drouaud *et al.*, 2006; Kim *et al.*, 2007; Gore *et al.*, 2009; Liu *et al.*, 2009; Branca *et al.*, 2011). However, the correlations estimated tended to be stronger and the combined explanatory power of the variables considered higher than in these studies. Furthermore, the clear pattern of nonrandom occurrence of recombination hotspots relative to genes (Table 2) was consistent with that detected based on detailed studies in humans (McVean *et al.*, 2004; Myers *et al.*, 2005). The ability to detect these patterns attests to the precision of our estimates of fine-scale recombination rates, and suggests that this information will probably be useful in both fundamental studies of DNA- and chromosome-level evolution and in applied molecular breeding.

In summary, our analyses of extensive genome resequencing and SNP genotyping data for a broad collection of *P. trichocarpa* trees led to two main novel findings. First, significant population genetic structure appears to be present at multiple spatial scales, and latitudinal allele frequency gradients are surprisingly common across the genome (Figs 1–3). Second, genome-wide LD extends over much larger physical distances than expected on the basis of previous smaller scale studies (Figs 4–6). These results have several implications. First, geographic patterns of population structure, allele frequency gradients and estimates of N_e from LD consistently suggest that genetic drift has played a significant role in the recent evolutionary history of *P. trichocarpa*, with post-glacial recolonization from multiple refugia appearing more plausible than recolonization from a single southern refugium. Second, the striking genome-wide ubiquity of high-LD regions and SNPs with latitudinal allele frequency gradients indicates that reliable selection scans in *P. trichocarpa* will require the development of *ad hoc* statistical tests integrating multiple molecular signatures. Finally, the commonly made assumptions that population structure is weak and LD decays within 1–2 kb in forest trees (Neale & Ingvarsson, 2008; Neale & Kremer, 2011) may not hold universally across

species of *Populus*. This will need to be reflected in the analyses of data from ongoing and in the design of future association studies. For example, statistically significant marker–phenotype associations will need to be interpreted in the specific context of local LD (i.e. the actual causative polymorphism(s) may, in some cases, be kilobases away from the markers used to detect the association). Most significantly, the extensive LD we detected makes genome-wide association studies and genomic selection in *P. trichocarpa* much more feasible than previously assumed (Neale & Ingvarsson, 2008; Neale & Kremer, 2011), especially given the increasing cost-effectiveness of SNP genotyping.

Acknowledgements

Funding was provided by the BioEnergy Science Center, a US Department of Energy (DOE) Bioenergy Research Center (Office of Biological and Environmental Research in the DOE Office of Science) and by the Province of British Columbia through Genome British Columbia Applied Genomics Innovation Program project 103BIO. The work conducted by the DOE Joint Genome Institute was supported by the Office of Science of the DOE under Contract No. DE-AC02-05CH11231. Reinhard Stettler, Jon Johnson, Brian Stanton, Richard Shuren, Nancy Engle, Xiaohan Yang and Stan Wullschleger provided assistance with the collection of plant materials. We thank Reinhard Stettler, Glenn Howe, the *New Phytologist* Editor and three anonymous reviewers for their comments on earlier versions of the manuscript.

References

Andolfatto P, Nordborg M. 1998. The effect of gene conversion on intralocus associations. *Genetics* **148**: 1397–1399.

Balloux F, Lehmann L, de Meeûs T. 2003. The population genetics of clonal and partially clonal diploids. *Genetics* **164**: 1635–1644.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L *et al.* 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences, USA* **108**: E864–E870.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* **36**: 700–706.

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*. Washington, DC, USA: US Department of Agriculture, Forest Service, 570–576.

DiFazio SP, Leonardi S, Slavov GT, Garman S, Adams WT, Strauss SH. 2012. Gene flow and simulation of transgene dispersal from hybrid poplar plantations. *New Phytologist* **193**: 903–915.

Drouaud J, Camilleri C, Bourguignon PY, Canaguier A, Bérard A, Vezon D, Giancola S, Brunel D, Colot V, Prum B *et al.* 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Research* **16**: 106–114.

Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao YJ, Tam A, Wang SC, Friedmann M, Birol I *et al.* 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources* **11**: 81–92.

Gilchrist EJ, Haughn GW, Ying CC, Otto SP, Zhuang J, Cheung D, Hamberger B, Aboutorabi F, Kalynyak T, Johnson L *et al.* 2006. Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* **15**: 1367–1378.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. 2006a. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* **172**: 1915–1926.

González-Martínez SC, Krutovsky KV, Neale DB. 2006b. Forest-tree population genomics and adaptive evolution. *New Phytologist* **170**: 227–238.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195–202.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J *et al.* 2009. A first-generation haplotype map of maize. *Science* **326**: 1115–1117.

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O *et al.* 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**: 883–886.

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

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**: 54–78.

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

Jansson S, Bhalerao R, Groover AT. 2010. *Genetics and genomics of Populus*. New York, NY, USA: Springer.

Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin TM, DiFazio SP, Ali J *et al.* 2007. A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. *Plant Journal* **50**: 1063–1078.

Keller SR, Olson MS, Silim S, Schroeder W, Tiffin P. 2010. Genomic diversity, population structure, and migration following rapid range expansion in the balsam poplar, *Populus balsamifera*. *Molecular Ecology* **19**: 1212–1226.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **39**: 1151–1155.

Krutovsky KV, Neale DB. 2005. Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. *Genetics* **171**: 2029–2041.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**: 1851–1858.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.

Little EA Jr. 1971. *Atlas of United States trees, volume 1, conifers and important hardwoods*. US Department of Agriculture Miscellaneous Publication 1146, 9 p., 200 maps. [WWW document] URL http://esp.cr.usgs.gov/data/atlas/little/poputric.pdf [accessed on 13 June 2012].

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, Tang HM, Fu Y, Nettleton D, Schnable PS. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genetics* **5**: e1000733.

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR *et al.* 2006. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* **78**: 437–450.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581–584.

Metzker ML. 2010. Sequencing technologies – the next generation. *Nature Reviews Genetics* **11**: 31–46.

Meuwissen T, Goddard M. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* **185**: 623–631.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324.

Neale DB, Ingvarsson PK. 2008. Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**: 149–155.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**: 111–122.

Nei M, Miller JC. 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* **125**: 873–879.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G *et al.* 2008. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American Journal of Human Genetics* **83**: 347–358.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research* **25**: 2745–2751.

Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics* **10**: 745–755.

Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186**: 526–536.

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

Peakall R, Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**: 288–295.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M *et al.* 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**: 1563–1565.

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

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**: 459–463.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**: 1219–1228.

Rubin EM. 2008. Genomics of cellulosic biofuels. *Nature* **454**: 841–845.

Savolainen O, Pyhäjärvi T. 2007. Genomic diversity in forest trees. *Current Opinion in Plant Biology* **10**: 162–167.

Slatkin M. 2008. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**: 477–485.

Slavov GT, Leonardi S, Adams WT, Strauss SH, DiFazio SP. 2010. Population substructure in continuous and fragmented stands of *Populus trichocarpa*. *Heredity* **105**: 348–357.

Slavov GT, Leonardi S, Burczyk J, Adams WT, Strauss SH, DiFazio SP. 2009. Extensive pollen flow in two ecologically contrasting populations of *Populus trichocarpa*. *Molecular Ecology* **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*. New York, NY, USA: Springer, 15–38.

Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS. 1997. Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution* **206**: 353–373.

Stanton BJ, Neale DB, Li S. 2010. *Populus* breeding: from the classical to the genomic approach. In: Jansson S, Bhalerao R, Groover AT, eds. *Genetics and genomics of Populus*. New York, NY, USA: Springer, 309–348.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**: 449–462.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**: 978–989.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**: 520–526.

Terwilliger JD, Hiekkalinna T. 2006. An utter refutation of the 'Fundamental Theorem of the HapMap'. *European Journal of Human Genetics* **14**: 426–437.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

VanLiere JM, Rosenberg NA. 2008. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology* **74**: 130–137.

Vasemägi A. 2006. The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics* **173**: 2411–2414.

Wackerly DD, Mendenhall W, Scheaffer RL. 2002. *Mathematical statistics with applications. 6th edn.* Pacific Grove, CA, USA: Duxbury.

Weber JC, Stettler RF. 1981. Isoenzyme variation among ten populations of *Populus trichocarpa* Torr. et Gray in the Pacific Northwest. *Silvae Genetica* **30**: 82–87.

Weber JC, Stettler RF, Heilman PE. 1985. Genetic variation and productivity of *Populus trichocarpa* and its hybrids. I. Morphology and phenology of 50 native clones. *Canadian Journal of Forest Research* **15**: 376–383.

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai CJ, Neale DB. 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist* **188**: 515–532.

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

Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM *et al.* 2006. A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* **7**: 510–523.

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

Supporting Information

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

Fig. S1 Population structure in *Populus trichocarpa* based on 1000 randomly chosen Infinium single-nucleotide polymorphism (SNP) loci (one per candidate gene) and model-based clustering using v. 2.2 of the STRUCTURE program.

Fig. S2 Linear associations between genetic (F_{ST}, based on Infinium single-nucleotide polymorphism (SNP) data, 22 280 loci) and geographic distances among subpopulations of *Populus trichocarpa*.

Fig. S3 Correspondence of r^2 values calculated on the basis of estimated haplotype frequencies (i.e. using PHASE) to those based on genotypic correlations.

Fig. S4 Effect of population structure on estimates of linkage disequilibrium (LD; r^2) based on genome resequencing and

Infinium single-nucleotide polymorphism (SNP) data (a) and LD-based estimates of effective population size (N_e) from genome resequencing data (b).

Fig. S5 Linkage disequilibrium for physically unlinked loci.

Fig. S6 Correlates of recombination in *Populus trichocarpa*.

Fig. S7 Recombination rates in 1-Mb windows estimated from a dense single-nucleotide polymorphism (SNP) linkage map and from genome resequencing data.

Table S1 Summary of genome resequencing data for 16 *Populus trichocarpa* trees

Table S2 Summary statistics for Sanger resequencing data for 10 candidate genes in 47 *Populus trichocarpa* trees

Table S3 Genetic (below diagonal, F_{ST}, based on Infinium single-nucleotide polymorphism (SNP) data, 22 280 loci) and geographic (above diagonal, km) distances among 10 subpopulations of *Populus trichocarpa*

Table S4 100-kb window correlation matrix of recombination correlates in *Populus trichocarpa*

Table S5 1-Mb window correlation matrix of recombination correlates in *Populus trichocarpa*

Table S6 Multiple linear regression model for $\log_{10}(4N_ec)$ analyzed in 100-kb windows

Table S7 Multiple linear regression model for $\log_{10}(4N_ec)$ analyzed in 1-Mb windows

Methods S1 Infinium single-nucleotide polymorphism (SNP) data.

Methods S2 Correlates of recombination.

Methods S3 Estimates of effective population size from linkage disequilibrium.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.