MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1023–1036 doi: 10.1111/1755-0998.12513

New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array

P. FAIVRE-RAMPANT,*[1] G. ZAINA,†[1] V. JORGE,‡ S. GIACOMELLO,§ V. SEGURA,‡ S. SCALABRIN,§ V. GUÉRIN,‡ E. DE PAOLI,§ C. ALUOME,*‡ M. VIGER,¶ F. CATTONARO,§ A. PAYNE,¶ P. PAULSTEPHENRAJ,* M. C. LE PASLIER,* A. BERARD,* M. R. ALLWRIGHT,¶ M. VILLAR,‡ G. TAYLOR,¶ C. BASTIEN‡ and M. MORGANTE†§

*INRA, US1279 EPGV, CEA-IG/CNG, F-91057, Evry, France, †DI4A, University of Udine, via delle Scienze 206, 33100 Udine, Italy, ‡INRA, UR 0588 AGPF, Centre INRA Val de Loire, 2163 avenue de la Pomme de Pin, CS 40001 – Ardon, 45075, Orléans, France, §IGA, Parco Scientifico e Tecnologico Luigi Danieli, via Jacopo Linussio 51, 33100 Udine, Italy, ¶Centre For Biological Sciences, University of Southampton, Life Sciences, SO17 1BJ, Southampton, UK

Abstract

Whole genome resequencing of 51 *Populus nigra* (L.) individuals from across Western Europe was performed using Illumina platforms. A total number of 1 878 727 SNPs distributed along the *P. nigra* reference sequence were identified. The SNP calling accuracy was validated with Sanger sequencing. SNPs were selected within 14 previously identified QTL regions, 2916 expressional candidate genes related to rust resistance, wood properties, water-use efficiency and bud phenology and 1732 genes randomly spread across the genome. Over 10 000 SNPs were selected for the construction of a 12k Infinium Bead-Chip array dedicated to association mapping. The SNP genotyping assay was performed with 888 *P. nigra* individuals. The genotyping success rate was 91%. Our high success rate was due to the discovery panel design and the stringent parameters applied for SNP calling and selection. In the same set of *P. nigra* genotypes, linkage disequilibrium throughout the genome decayed on average within 5–7 kb to half of its maximum value. As an application test, ADMIXTURE analysis was performed with a selection of 600 SNPs spread throughout the genome and 706 individuals collected along 12 river basins. The admixture pattern was consistent with genetic diversity revealed by neutral markers and the geographical distribution of the populations. These newly developed SNP resources and genotyping array provide a valuable tool for population genetic studies and identification of QTLs through natural-population based genetic association studies in *P. nigra*.

Keywords: HT-genotyping design, large-scale SNP discovery, population genetics, *Populus nigra*

Received 22 September 2015; revision received 17 December 2015; accepted 22 December 2015

Introduction

Black poplar (*Populus nigra* L., Salicaceae) is an Eurasian native species distributed within riparian corridors in lowland, piedmont and mountainous zones from Morocco and Ireland at the western limit of its natural range to Russia and China in the East (Dickmann & Kuzovkina 2013). As a pioneer species, *P. nigra* plays an important role in the establishment of riparian ecosystems (Imbert & Lefèvre 2003), where it can be found as isolated trees and in pure or mixed stands. Black poplar is threatened throughout its natural range by anthropogenic disturbances of the river bank; gene introgression from cultivars (*P. deltoides* x *P. nigra*) and from the worldwide spread of fastigiated form *P. nigra* var *italica*, (Cagelli & Lefèvre 1997; Vanden Broeck *et al.* 2005). As such black poplar deserves special attention in terms of conservation at national and European levels (Lefèvre *et al.* 2001). Microsatellite genetic variation analyses showed high genetic diversity within populations and weak but significant genetic differentiation across river basins suggesting high levels of gene flow (Smulders *et al.* 2008; DeWoody *et al.* 2015).

Ease of vegetative propagation, good coppicing ability, resistance and tolerance to several bio-aggressors (Benetka *et al.* 2012); a long growing season (Rohde *et al.* 2011) and high plasticity in response to environmental changes (Chamaillard *et al.* 2011) are important adaptive characteristics that have promoted black poplar as a par-

Correspondence: Patricia Faivre-Rampant, Fax: 33 1 60 87 84 55; E-mail: faivre@versailles.inra.fr

[1]These authors contributed equally to this work.

ental pool in interspecific breeding programs world-wide (Stanton *et al.* 2013). The first common garden experiments performed with natural populations of black poplar have revealed locally adapted populations for bud set phenology (Rohde *et al.* 2011) and leaf traits (DeWoody *et al.* 2015; Guet *et al.* 2015). Local adaptation has also been reported in other poplar species (Ingvarsson *et al.* 2006; Keller *et al.* 2010; Viger *et al.* 2013) as well as in other temperate, widespread forest trees (Savolainen *et al.* 2007). Past adaptation processes have most likely generated wide reservoirs of standing genetic variation for many other adaptive traits in black poplar.

One main challenge is to identify loci/genes that underlie this phenotypic variation. Such information can then be used to access and manage genetic diversity and develop adapted marker-assisted selection schemes (Harfouche *et al.* 2012). Association genetics is a promising method for achieving this goal in woody species with a long life cycle; late expression of important traits and considerable population genetic diversity (Neale & Savolainen 2004; Neale & Kremer 2011). The development of High Throughput (HT) genotyping tools is undoubtedly a prerequisite for such an approach. Single nucleotide polymorphisms (SNPs) are a suitable and very attractive genetic markers for this purpose. It is now well established that HT DNA sequencing technologies are powerful tools enabling the rapid discovery of large numbers of SNPs. Different options have been deployed in tree species and other plants including RNA sequencing (i.e. HT-sequencing at the transcriptome level (Parchman *et al.* 2010; Geraldes *et al.* 2011; Howe *et al.* 2013; Mantello *et al.* 2014)) and targeted sequencing (i.e. HT-sequencing of particular, captured genomic regions such as the gene-enriched portion (Zhou & Holiday 2012) or restricted genomic DNA (Grattapaglia *et al.* 2011; Schilling *et al.* 2014). For species with a relatively small genome like *Populus* sp. (500 Mb) whole genome HT-sequencing can be feasibly achieved (Slavov *et al.* 2012; Evans *et al.* 2014). Recently, studies have demonstrated the usefulness of both HT-sequencing and SNP arrays to allow candidate gene association genetics in natural populations of *Populus trichocarpa* (Porth *et al.* 2013; McKown *et al.* 2014). The success of association studies mainly depends on the availability of SNPs; the extent of linkage disequilibrium (LD); the extent of variation in the phenotype of interest and the genetic structure of the association population. In *P. nigra*, these determinants are poorly documented however, previous studies (limited to relatively few SNPs identified within 2–39 genes) reported LD to decay within 300–1000 bp (Chu *et al.* 2009, 2014; Marroni *et al.* 2012; Guerra *et al.* 2013).

In order to perform association studies in *P. nigra*, our aims were to identify SNPs at the whole genome scale and develop a SNP bead-chip array. Due to the expected rapid decay of LD in most undomesticated tree species, we opted for a candidate-genomic-region approach that focused on leaf rust resistance, bud phenology, water-use efficiency and wood chemistry in both QTL intervals identified in *P.nigra* mapping pedigrees (Rohde *et al.* 2011; Fabbrini *et al.* 2012; El-Maki 2013; Guet *et al.* 2015) and candidate genes underlying QTLs in other poplar species (Rae *et al.* 2008; Novaes *et al.* 2009; Ranjan *et al.* 2010; Monclus *et al.* 2012; Viger *et al.* 2013). SNPs outside the candidate regions were also selected to provide genomic control tools to characterize neutral genetic diversity and detect population structure. To reach this objective, we first created a *P. nigra* reference sequence using the *P. trichocarpa* genome sequence as a template (Tuskan *et al.* 2006) and identified a large set of SNPs at the whole genome scale by HT-resequencing of 51 *P. nigra* genomes. Second, we defined a SNP selection strategy in order to design a useful SNP array for candidate-based association studies in natural populations. Third, the usefulness of the array was evaluated by genotyping 888 *P. nigra* individuals. Data analysis focused on LD decay with distance and on the genetic structure of a large *P. nigra* association population sampled in 12 river basins across Western Europe.

Material and methods

SNP discovery and selection

Discovery panel and whole genome resequencing. A SNP discovery panel of 51 individuals selected as representative of the genetic diversity of an association population covering the range of the black poplar in Western Europe was used for HT-genome sequencing (Table S1, Supporting information).

Nuclear DNA was isolated from young leaves as described by Zhang *et al.* (1995) and Chalhoub *et al.* (2004). Whole genome resequencing was performed at the Institute of Applied Genomics (IGA, Udine, Italy) and the INRA-EPGV/CEA-IG/CNG (Institut National de la recherche Agronomique-Etude du Polymorphisme des Génomes Végétaux/Commissariat à l'Energie Atomique-Institut de Génomique/Centre National de Génotypage, Evry, France) facilities using either the GAII analyser or Hiseq 2000 Illumina platforms (Inc. San Diego, CA, USA). Paired-end sequencing libraries were prepared following the "Illumina Paired-End Sample Preparation" protocol using an insert size spanning from 300 to 600 bp. Paired-end runs were performed for 76, 100, or 114 cycles following Illumina's operating instructions (Table 1). The Illumina sequencer analyser provided a quality score (Qscore) for each base and an average Qscore value was assigned to each read. Reads with Qscore values >30 were considered as good sequences.

Table 1 Raw sequence data used for SNP detection

Genotype	Origin	River basin	Read length (b)	Total bp produced	Raw coverage (X)
Poli	Italy	Sinni River	100	34 031 232 782	81.6
BEN3	Spain	Ebro	100	21 882 737 550	52.5
71077-308	France	Rhône	76, 114	11 614 046 643	27.8
Blanc_de_Garonne	France	Garonne	100	10 499 784 562	25.1
92538	France	Creuse (Loire)	100	8 874 612 395	21.3
72145-7	France	Gard (Rhône)	100	8 279 967 553	19.8
6-A06	France	Drôme (Rhône)	100	8 124 691 652	19.5
1-A10	France	Drôme (Rhône)	100	7 616 642 138	18.3
92525-25	France	Loire	100	7 379 085 905	17.7
92520-6	France	Loire	100	7 100 652 141	17
92510-3	France	Loire	100	6 599 547 430	15.8
Sarrazin	France	Garonne	100	6 545 172 797	15.7
Vert_de_Garonne*	France	Garonne	100	5 865 971 615	14
6-A23	France	Drôme (Rhône)	100	5 733 143 633	13.7
NVHOF2/19	Germany	Rhine-D (Rhine)	100	5 638 954 091	13.5
6-A31	France	Drôme (Rhône)	100	4 957 635 050	11.9
99582-1	France	Loire	100	4 749 535 204	11.4
Cazebonne_25*	France	Garonne	100	3 885 764 113	9.3
PG-22	Italy	Paglia (Tibre)	100	3 542 852 254	8.5
SN-21	Italy	Ticino (Pô)	100	3 183 780 277	7.6
Ginsheim3	Germany	Rhine-D (Rhine)	100	3 114 417 000	7.5
NL-1238	Netherlands	Rhine_Ijssel	100	3 095 875 836	7.4
98568-1	France	Rhine F (Rhine)	100	2 811 019 907	6.7
SN-11	Italy	Ticino (Pô)	100	2 791 982 335	6.7
NL-1217	Netherlands	Rhine_Ijssel	100	2 543 452 219	6.1
NVHOF3/17	Germany	Rhine-D (Rhine)	100	2 475 035 580	5.9
FTNY19	Hungary	Tisa	100	2 419 647 905	5.8
Ginsheim1	Germany	Rhine-D (Rhine)	100	2 351 224 600	5.6
C2	Spain	Ebro	100	2 160 560 966	5.2
SN-26	Italy	Ticino (Pô)	100	2 174 897 241	5.2
C1	Spain	Ebro	100	2 116 880 335	5
NL-1329	Netherlands	Rhine_Ijssel	100	2 067 806 626	5
NL-1682	Netherlands	Rhine_Waal/Maas	100	2 046 322 170	4.9
PG-05	Italy	Paglia (Tibre)	100	2 055 865 151	4.9
cart5	Spain	Ebro	100	1 936 051 399	4.6
NL-2051	Netherlands	Individual clone	100	1 826 967 332	4.4
73193-25	France	Gave_de_Pau (A)	100	1 647 799 444	4
N-11	Italy	Ticino (Pô)	100	1 676 606 505	4
PG-13	Italy	Paglia (Tibre)	100	1 665 449 401	4
N-38	Italy	Ticino (Pô)	100	1 540 547 636	3.7
C6	Spain	Ebro	100	1 460 806 904	3.5
58-861	Italy	Cenischia (Pô)	100	1 425 822 523	3.4
FTNY18	Hungary	Tisa	100	1 336 413 883	3.2
BDX-06	France	Gave_de_Pau (A)	100	1 199 931 013	2.9
RIN4	Spain	Ebro	100	1 224 325 600	2.9
SN-40	Italy	Ticino (Pô)	100	1 195 698 229	2.9
C12	Spain	Ebro	100	1 026 605 990	2.5
71072-501	France	Rhône	100	1 020 158 073	2.4
NL-1797	Netherlands	Rhine_Waal/Maas	100	910 082 000	2.2
NVHOF3/5	Germany	Rhine-D (Rhine)	100	878 908 000	2.1
N-47	Italy	Ticino (Pô)	100	691 873 200	1.7

*Vert de Garonne and Cazebonne 25 were subsequently found to be identical genotypes after HT-genotyping. (A) Adour.

Four individuals covering the wide Western latitudinal range of *Populus nigra*, Poli (South-Italy), BEN3 (Spain), Blanc de Garonne (BDG) (South-West-France) and 71077–308 (East-France) were sequenced at a coverage >25× (Tables 1 and S1, Supporting information). Our objective was twofold; to maximize the genetic

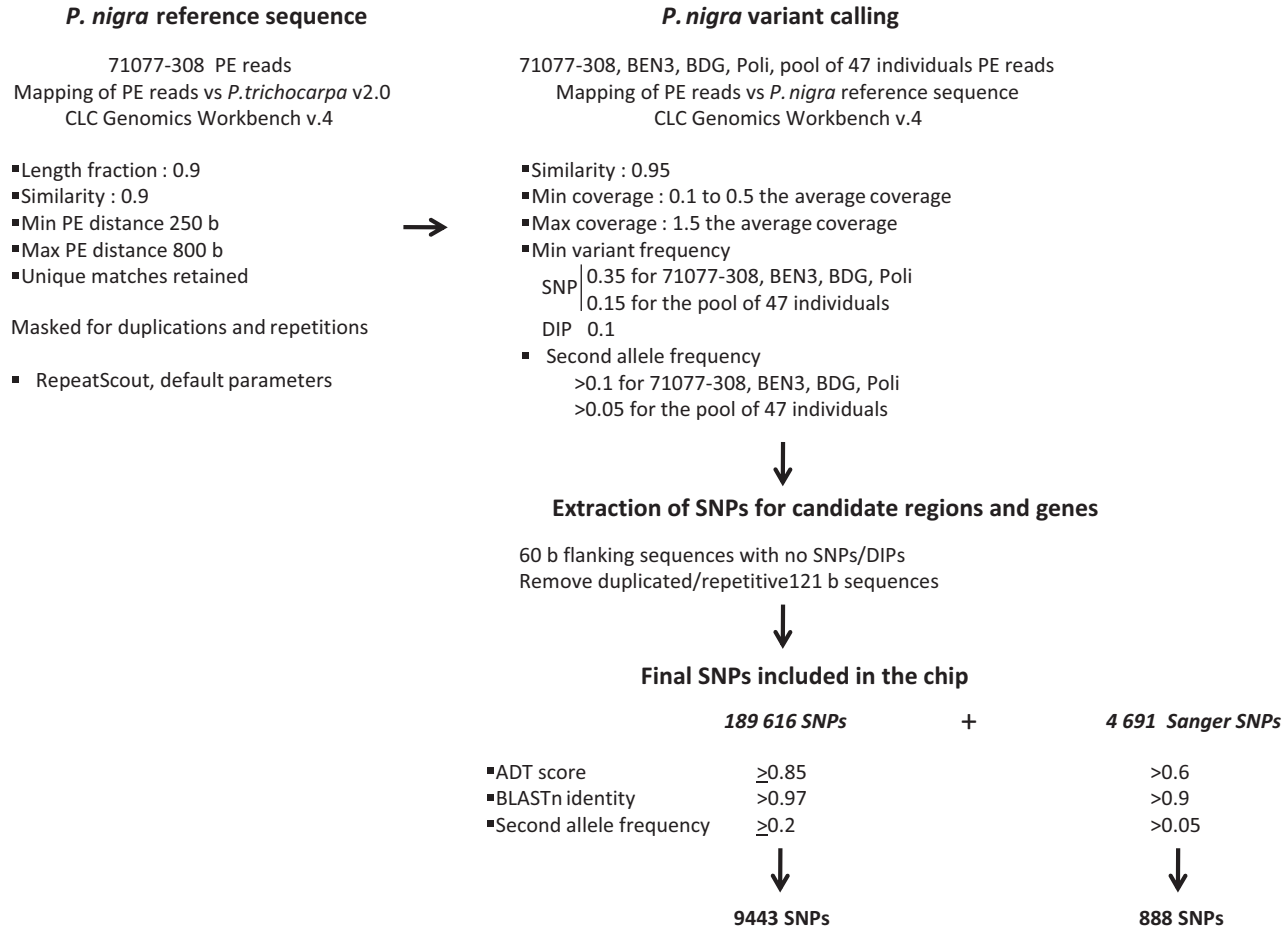

Fig. 1 Workflow of SNP detection and selection.

variation among individuals and to identify reliable SNPs. Forty-seven individuals covering the European latitudinal range were selected and sequenced at lower coverage (Tables 1 and S1, Supporting information) in order to maximize the discovery of informative SNPs.

Populus nigra reference sequence. To avoid confusion resulting from interspecific polymorphisms between *Populus trichocarpa* and *P. nigra* and prompt the detection of intraspecific polymorphisms within *P. nigra* (Isabel *et al.* 2013), we created a *P. nigra* reference sequence using short reads of the genotype 71077-308 (27×). This genotype was chosen for its high read Qscore >32. Paired-end reads were aligned onto the *P. trichocarpa* genome V2.0 (Tuskan *et al.* 2006). Indeed, pilot analyses on Sanger-sequenced BAC inserts showed the feasibility of using the *P. trichocarpa* genome sequence as a template for *P. nigra* (Zaina, unpublished data). The mapping of raw short reads was performed with the CLC GENOMICS WORKBENCH v.4 (CLC Bio, Aarhus, Denmark). Mapping parameters are shown in Fig. 1. Only paired-end reads that aligned to a unique location of the genome were considered. Duplications and repetitions were identified with RepeatScout using default parameters (Price *et al.* 2005). Due to computing constraints, only the first 40 scaffolds were extracted as part of the *P. nigra* reference sequence to be used in the SNP calling.

Strategy of SNP detection for designing the array. A multistep strategy was designed to recover variants for the Illumina Infinium iSelect HD Custom BeadChip technology. The paired-end raw sequences of the 4 genotypes >25× were mapped separately onto the *P. nigra* reference sequence using the same procedure adopted above to create the *P. nigra* reference itself; with the exception that similarity was set to 0.95. Reads for the 47 remaining accessions were aligned similarly but as a joint set. SNP detection was then performed on each of the five alignments with the parameters detailed in Fig. 1. To evaluate the accuracy of the SNP calling a comparison with the SNPs detected using ABI3730 Sanger sequencing was performed (Table S2, Methods S1 and S2, Supporting information).

Deletion-Insertion Polymorphisms (DIPs) were also detected to optimize SNP selection for the array design. DIPs were detected using the CLC SOFTWARE v.4 (Fig. 1).

Given the objectives of the SNP array; candidate-genomic regions (14) were selected on the basis of QTL for rust resistance and bud phenology in *P. nigra* and water-use efficiency and wood properties in other *Populus* species (Table S3, Supporting information). Candidate genes (2916) for the same traits were also considered on the basis of transcriptome studies and the literature (Table S3, Supporting information). SNPs located within these candidate regions or genes were considered at the subsequent selection step. Additional SNPs were retained within gene models (1732) spread across the poplar genome.

A pipeline written in Bash and Perl was set up to extract useful SNPs with 60-bp flanking sequences. The pipeline rescued only loci whose flanking sequences did not contain any SNPs and/or DIPs. If this was not possible, the pipeline was set to select SNPs with no SNPs and/or DIPs within ± 10 bp of the target SNP. The pipeline also discarded SNPs within duplicated or repetitive regions.

A collection of SNPs detected by Sanger resequencing of full-length genes and gene fragments obtained previously by the University of Udine and INRA teams within the framework of Popyomics and National projects were also considered (Method S1, Table S2, Supporting information).

The whole set of extracted SNPs was subjected to the Assay Design Tool by Illumina (https://iCom.illumina.com) in order to score and validate the SNPs in terms of the bead-chip's performance. A final round of selection was performed to reach the desired 11 999 beads. This final selection was based on the SNPs' location in the genome (Table S3, Supporting information): (i) 80 SNPs/Mb were retrieved from QTL regions showing a considerable effect (the phenotypic variance explained by the QTL was >10%) (ii) 20 SNPs/Mb were retrieved from QTL regions showing a low or moderate effect, (iii) 5 SNPs/Mb were retrieved from non-QTL regions. SNPs requiring a single bead type (Infinium II) were also preferred to maximize the number of loci on the chip. In a few regions, the final target could not be reached with the described criteria, which were gradually relaxed in these instances to obtain the target number. Moreover, for functional candidate genes for rust resistance and bud phenology, more than one SNP was selected per gene according to the same criteria.

Genotyping assay

Plant material. A collection of 888 individuals, comprising 838 native *P. nigra* individuals originating from 12 western European river basins (Tables 2 and S1, Supporting information) (of which most belonged to the Europop (Storme *et al.*, 2004) and the French National collections) and 50 full sib progenies, were used in this study (Table S1, Supporting information). Among the 838 native *P. nigra*, 814 were part of the European association population established in the framework of the EU projects Popyomics, Evoltree, NovelTree and EnergyPoplar, and had already been genotyped with SSR markers (Storme *et al.* 2004; DeWoody *et al.* 2015; Jorge unpublished data). Within the entire collection, 11 individuals were used as parents in nine different crosses and two to six progenies per cross were genotyped to facilitate and validate SNP clustering.

SNP genotyping. One sample (BDG) was replicated 14 times and used for technical control. DNA samples from 24 individuals were included twice to assess the repeatability of allele calls. SNP genotyping was conducted on the Illumina Platform at CEA-IG/CNG by INRA-EPGV according to the standard Illumina protocol. Genotypes were recovered with Genotyping Module v 1.9.4 (Genome Studio software v 2011.1, Illumina Inc.). Clusters were generated using a GenCall score cut-off of 0.15 as recommended by Illumina. The GenCall score, estimated for each data point (SNP x individual sample) and implemented by the Genome Studio software, reflected the position of the data point within the genotype cluster. Genotypes with lower GenCall scores are located further from the centre of the genotype cluster and have lower reliability. Only those individuals with >95% call rates were selected (i.e. the proportion of individual samples successfully genotyped in a locus). SNP clusters were automatically generated and then the quality of the three expected clusters of each SNP was inspected visually. Subsequent adjustment of the cluster calling was performed if needed.

Table 2 Summary of the number of *Populus nigra* genotypes per river basin in the European *P. nigra* association population

River Basins	Country	No. individuals genotyped
Dranse (Rhône)	France	40
Durance (Rhône)	France	13
Drôme (Rhône)	France	155
Loire	France	180
Rhine F	France	62
Allier	France	113
Basento	Italy	14
Paglia	Italy	22
Ticino	Italy	103
Rhine-D	Germany	54
Netherlands NL	Netherlands	48
All stands-Ebro	Spain	9

Linkage disequilibrium and population structure

To estimate LD decay and analyse population structure based upon neutral genetic diversity, SNPs and individuals were filtered according to several criteria. First, SNPs and individuals with more than 10% missing data were discarded. The segregation and the linkage conformity were checked within a 3 × 3 factorial mating design (Fig. S1, Supporting information). Finally, SNPs showing a significant departure from Hardy—Weinberg equilibrium within more than sxi populations were discarded. LD between all pairs of SNPs was estimated as the square of the allelic correlation in R (R Core Team, 2014).

Population structure was investigated using the software ADMIXTURE (Alexander *et al.* 2009) with K ancestral population ranging from 1 to 15. Since we used a candidate-based approach the selected SNPs were not evenly spread throughout the genome. To account for such variation in SNP density across the genome we sampled several subsets of SNPs. These subsets were sampled by chromosome; taking into account physical chromosome length and the desired final number of SNPs using different approaches:

1. 2000-LD: 2000 SNPs minimizing the LD between SNPs by applying the Kennard and Stone algorithm (Kennard & Stone 1969) to the LD matrix by chromosome;
2. 600-LD: same as above but with a total target of 600 SNPs;
3. 600-dist: 600 SNPs well scattered by applying the Kennard and Stone algorithm to the physical distance matrix by chromosome;
4. 600-random: 600 SNPs randomly sampled by chromosome.

These four subsets were compared together and to the total set of high-quality SNPs to evaluate population structure by cross-validation in ADMIXTURE. The subset that minimized the cross-validation error was selected to analyse population structure. The optimal number of groups (K) was also determined by cross-validation for this subset. The optimal subset of SNPs according to the cross-validation in ADMIXTURE was also used to carry out Principal Component Analysis (PCA) in R (R Core Team, 2014) as a complementary analysis of population structure. Genetic differentiation between pairs of populations was estimated by calculating Jost's D values (Jost 2008).

We used the optimal subset of SNPs to estimate a measure of LD corrected for the bias attributed to population structure and cryptic relatedness as proposed by Mangin *et al.* (2012). Briefly, we used the optimal subset of SNPs to compute a genomic relationship matrix between individuals (VanRaden 2008) and used this matrix to estimate a corrected measure of LD defined as the squared partial allelic correlation between SNPs (Lin *et al.* 2012). The relationship between LD and physical distance was assessed following the model of Hill & Weir (1988) in order to determine the genomic distance at which LD decays to half its maximum value.

Results

Illumina next generation DNA sequencing technology was used to resequence four *P. nigra* genotypes (71077-308, BDG, BEN3 and Poli) at >25× coverage and 47 other genotypes at lower coverage. The read data and relative raw coverage obtained for each genotype are reported in Table 1.

SNP detection

P. nigra reference sequence. The sequence data obtained from the clone 71077-308 were selected due to their good quality to produce a reference sequence for *P. nigra*; exploiting a mapping approach vs. the *P. trichocarpa* genome sequence v2.0. We previously proved the feasibility of this approach by mapping the short reads of another *P. nigra* genotype (the Spanish clone BEN3) vs. two *P. nigra* BAC-clone sequences and vs. the *P. trichocarpa* sequence portions corresponding to the BAC inserts. In the intraspecific alignment, the BAC sequences were covered for 98% of their length as expected and, in the interspecific alignment, 75% of the corresponding *P. trichocarpa* regions were covered (Zaina, unpublished data). In the present work, 71077-308's short reads covered 79% of the *P. trichocarpa* genome sequence V2.0. After mapping, we considered only the *consensus* specific to the first forty scaffolds, which resulted in a sequence 388 572 533 bp long (gaps included), representing the sequence described hereafter as the *P. nigra* reference sequence.

SNP calling. We used the *P. nigra* reference sequence obtained to map the paired-end reads of 71077-308, BDG, BEN3 and Poli (>25×). Approximately 60% of the input reads from 71077-308, BDG and Poli were mapped to a unique position in the reference sequence. The exception of BEN3, with a lower proportion of mapped reads (42%), was explained by the lower quality score (reads average Qscore <26) of its reads compared to the others (Table S4, Supporting information). In addition to the four alignments produced above, the reads derived from the resequencing of the 47 individuals (<25×) were mapped as a whole against the *P. nigra* reference sequence to obtain a fifth alignment.

These alignments were used for SNP discovery at the whole genome scale following the procedure

Table 3 Numbers of SNPs identified for the development of the bead-chip array

SNPs	47 accessions	Poli	BEN3	BDG	71077-308
Input	758 043	937 790	282 299	491 850	460 047
Within candidate loci	296 964	344 709	112 262	174 035	155 846
After DIP removal	279 813	314 457	105 212	157 061	143 312
Supported by five accessions	278 330				
Supported by at least one >25× genotype clone	189 616				

summarized in Fig. 1. The total number of SNPs detected in each alignment along the *P. nigra* reference sequence is shown in Table 3 and these are referred to as input SNPs. Figure S2 (Supporting information) shows the distribution of the input SNPs detected through the five alignments across the main 19 chromosomes of the *P. nigra* reference. Of the 388 572 533 bp of the *P. nigra* reference sequence 110 098 472 bp were covered by the four genotypes and provided a total of 1 878 727 SNPs. The SNP frequency averaged as one polymorphism every 58.6 bp.

To estimate SNP calling accuracy, we compared the SNPs identified within the 18 candidate genes for the light signalling pathway (Table S2, Supporting information) resulting from the resequencing; using both Sanger and Illumina methods. A total of 96 164 sites were analysed including 1186 polymorphic sites from the Sanger SNP detection. The Illumina SNP detection achieved 92.9% Sensitivity, 99.8% Specificity and 99.7% Accuracy and provided 141 false positives (*i.e.* SNPs identified in Illumina data but not in Sanger data) corresponding to a 10.6% False Discovery rate (Method S2).

Development of the 12k Infinium BeadChip array

A total of 296 964 SNPs were retrieved from the 47 genotypes in the candidate regions while the other four genotypes provided 344 709 (Poli), 112 262 (BEN3), 174 035 (BDG) and 155 846 (71077-308) SNPs within the same regions (Table 3). The differences in the number of loci between the five alignments were consistent with the depth-coverage and read quality of the different genotypes. A map was created using the IUPAC codes to group all the SNPs belonging to different genotypes within the candidate loci. The map was integrated with the DIPs identified in the same five alignments (data not shown) to improve the further selection of SNPs for an efficient bead-chip array design (i.e. SNPs with no polymorphisms within the flanking sequences). Eventually, 189 616 SNPs, corresponding to one SNP every 1159 bp in the candidate regions and genes, were retained. This last set of 189 616 SNPs was subjected to the Illumina Assay Design Tool (ADT) to test their suitability for inclusion within the bead-chip design; 133 821 SNPs passed the test, showing an ADT score ≥0.6 (i.e. the score threshold recommended by Illumina. A set of 669 SNPs distributed within the noncandidate regions were also selected with the same criteria.

In addition to the SNPs identified by the Illumina HT resequencing, 4691 SNPs from the Sanger resequencing of candidate genes in *P. nigra* were considered (Fig. 1, Table S2, Supporting information). After filtering selection as detailed in Fig. 1, 2690 Sanger SNPs were available. Thus, the final pool of SNPs consisted of 137 180 loci. To reach the desired number of 11 999 beads required for the Illumina bead-chip array, the SNPs were reduced to 10 331 loci according to the stringent criteria detailed in the Material and Methods (Tables S5 and S6, Supporting information). Among these, 6311 were located within QTL intervals.

Infinium BeadChip array performance

Of the 10 331 SNPs, 9127 included in the bead pool (88%) remained in the array after Illumina technical dropout. Eight samples were excluded for technical errors and 19 were excluded due to low call rate and thus the final selection consisted of 861 genotypes with a call rate ≥0.95. Each cluster was then inspected manually and SNPs were classified as polymorphic, monomorphic or failed (Table S7, Supporting information). Our validation showed 8322 well clustered SNPs leading to a chip success rate estimated at 91%; 8259 of these were polymorphic (90%). The reproducibility rate was 100% when we compared the 12 inter-plate controls. The same rate was obtained from the comparison of: i. biological replicates of BDG and 1 inter-plate control; ii. the 24 duplicated genotypes. Heritability-based SNP validation was estimated to assess SNP assay quality. This was defined as the number of offspring genotypes that agreed with the expected inheritance over the total number of possible genotype calls. In nine families, there were 608 Mendelian transmission inconsistencies out of the 411 877 allelic transmissions assayed; i.e. a genotyping miscall rate of 0.15% (ranging from 0.08% to 0.21%). We observed that 1.65% of SNPs showed segregating errors.

A set of 259 SNPs from Sanger data was used to validate the efficiency of SNP genotyping in 10 individuals for which both Infinium and Sanger sequence data were available. We observed a very high rate of concordance (96–99%) (Table S8, Supporting

information). For 71077-308, BDG, BEN3 and Poli, we then compared genotype calls from NGS resequencing data to genotype calls from the chip. The concordance observed varied between 80% and 100% (Table S9, Supporting information). Of the 8259 SNPs, 7186 were located within 4903 genes and 1132 genes harboured more than 2 SNPs (Table S10, Supporting information).

Application of the array

Identification of clonal duplication. Polymorphic sites (8259) were used to compute pair-wise similarity between all pairs of individuals. This analysis identified 35 duplets, nine triplets, four quadruplets, two septuplets, and one duodeciduplet (Table S11, Supporting information). With the exception of five groups (three duplets, one triplet and one quadruplet), all the individuals belonging to a single group came from the same river population. Earlier genotyping with SSR markers was used to understand the origin of these results (Method S3, Table S11, Supporting information). Redundant individual genotypes were removed from the data set prior to further analyses.

Population structure. We applied additional filters on SNPs and individuals for genetic analyses. Filtering by missing data (>10%) resulted in discarding 13 SNPs and 26 individuals. Additional SNPs were discarded: 216 SNPs due to segregation problems (missing or unexpected genotyping class, segregation distortion and non-expected linkage, Fig. S1, Supporting information) in the factorial mating design (data not shown) and 98 SNPs due to significant deviations from Hardy—Weinberg equilibrium within at least six populations. In the resulting set of individuals, 36 SNPs were monomorphic and were discarded from further genetic analyses. The final data matrix included 7896 high-quality polymorphic SNPs genotyped in 706 individuals. Due to our biased sampling of SNPs within candidate regions (Fig. S3, Table S12, Supporting information), we further selected several subsets of 600 and 2000 SNPs as being potentially more evenly distributed throughout the genome. The optimal number of ancestral clusters (corresponding to the lowest cross-validation error) was $K = 7$ and was obtained with the optimal 600 SNP subset (Fig. 2a). The corresponding admixture results are shown in Fig. 2b. The Basento and Paglia populations from South and Central Italy emerged as distinct groups. For the other populations, a clear admixture pattern was revealed; although individuals from the same populations still tended to cluster together. A principal component analysis on the same optimal set of 600 SNPs confirmed the

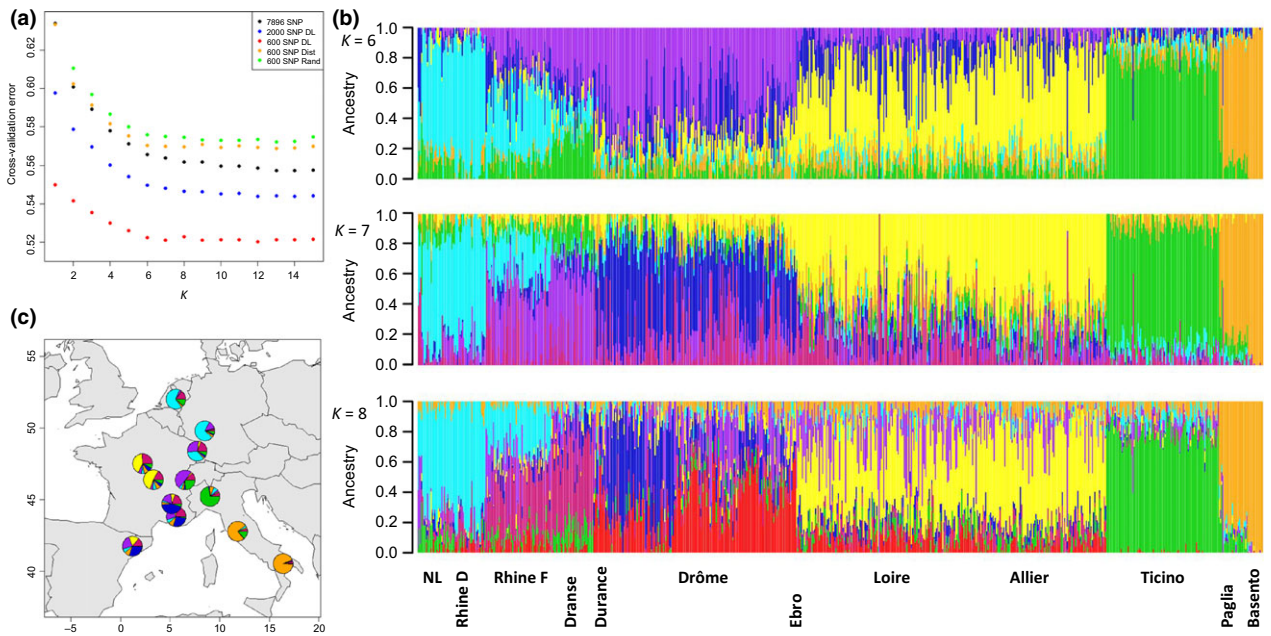

Fig. 2 Population structure analysis estimated from 600 SNPs distributed throughout the *Populus nigra* genome in validated genotypes. (a) Estimation of the best value of K determined by the cross-validation error implemented in ADMIXTURE software. K was tested for different sets of SNP detailed in the Material and Methods section. (b) Admixture results from 706 individuals and 600 SNPs - $K = 6$, $K = 7$, $K = 8$. Each colour represents a different ancestral cluster. Each individual is represented as a thin vertical bar which is divided into colour segments that are proportional to its membership of the ancestral clusters. At $K = 8$ individuals collected along the Rhône river basin were divided into two subpopulations; one is located on the upper part and the other one on the lower part of the river. (c) Geographical distribution of the populations and the genetic structure revealed by ADMIXTURE.

results from ADMIXTURE with a relatively clear clustering of individuals according to their geographical origin observed (Fig. S4, Supporting information).

Linkage disequilibrium

As expected due to the MAF (Minimum Allele Frequency) threshold (>0.2) applied to select SNPs in our discovery panel, the MAF of 92% of the high-quality genotyped SNPs is higher than 0.2 in the seven admixture clusters. The frequency distribution of SNPs was more or less even across different MAF classes and across ADMIXTURE clusters with the exception of the Italian clusters (Fig. S5, Supporting information). We calculated both LD and LD corrected for confounding population structure between all pairs of SNPs. The relationship between LD and physical distance was plotted and modelled (Fig. 3). As expected, the corrected LD decayed slightly faster with physical distance than the uncorrected LD with the r^2 and corrected r^2 dropping to half their maximum value within 7 and 4 kb respectively.

Discussion

We have reported the development of a high-quality SNP array in *Populus nigra*. To our knowledge, this is the first significant SNP resource that has been reported for black poplar. As poplar has a relatively small genome (500 Mb), we decided to resequence the whole genome instead of using the genome reduction procedure developed by Stölting *et al.* (2013). In poplar, SNPs are mostly species-specific (Isabel *et al.* 2013) and thus the already available *Populus trichocarpa* genome sequence could not be used directly as a reference to detect SNPs in *P. nigra*. Nevertheless, we were able to use it as a template to map the short reads from *P. nigra* to obtain a black poplar genome reference sequence. The alignment of paired-end reads allowed us to obtain 389×10^6 bp of *P. nigra* specific sequences (approximately 79% of the *P. trichocarpa* genome). The excluded regions generally corresponded to variations between the genomes of *P. trichocarpa* and *P. nigra* which we expect to be mostly repetitive regions such as those observed by Ma *et al.* (2013) between the genomes of *P. euphratica* and *P. trichocarpa* or large insertion/deletions due to transposable elements as observed by Zaina and Morgante (unpublished results) among BAC insert sequences belonging to *P. nigra*, *P. deltoides* and *P. trichocarpa*.

The comparison between the *P. nigra* reference sequence and 71077-308, BDG, BEN3 and Poli genotypes provided the first *P. nigra* whole genome SNP collection. The Italian genotype, Poli, contained more SNPs than the French and Spanish genotypes. This result was consistent with their respective genetic distances from the French genotype, 71077-308, used to build the *P. nigra* genome reference (Jorge and Villar, unpublished results). The procedure used to identify SNPs from the resequencing of four genotypes >25× and 47 genotypes <25× proved to be reliable; reducing the false discovery rate.

During our SNP selection process, most of the SNPs were lost during the final step; i.e. the selection of SNPs with no polymorphisms in their 60-bp flanking sequences. This can be explained by the high SNP frequency and heterozygosity in *P. nigra*. Hence, a huge collection of SNPs originating from complete genome coverage and a large SNP discovery panel was required to reach our final target of 12k beads. According to Groenen *et al.* (2011), the number of SNPs called should be at least 10 fold greater than the number targeted for the final chip. The positive genotyping results demonstrated that the strategy developed to detect and select SNPs was very effective; despite the lack of a reference sequence for *P. nigra*. The high concordance between genotyping and SNP calling data from Sanger sequencing and NGS genome sequencing revealed the robustness of our selection criteria. Our genotyping success

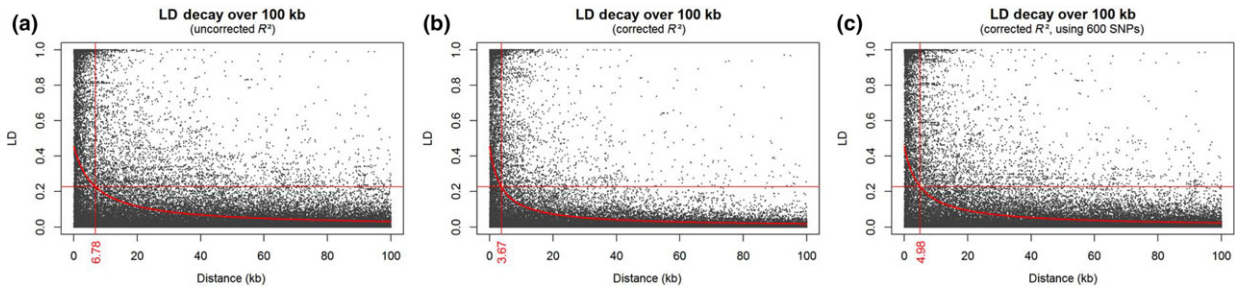

Fig. 3 Linkage disequilibrium vs. physical distances. (a) The decay of LD investigated by plotting all pairwise r^2 values against physical distance windows of 100 kb. (b) r^2 values corrected according to the population structure. (c) The decay of LD investigated by plotting 600 pairwise r^2 values against physical distance windows of 100 kb.

rate (91%) exceeded those recorded for other plant species with the same Infinium technology and in the same genotyping throughput range (6–10k) (Chagné *et al.* 2012; Verde *et al.* 2012; Bachlava *et al.* 2012; Peace *et al.* 2012; Sim *et al.* 2012; Delourme *et al.* 2013; Li *et al.* 2014; Dalton-Morgan *et al.* 2014; Lepoittevin *et al.* 2015; Livingstone *et al.* 2015). The success of the SNP array was due to the composition of the SNP discovery panel reflecting the genetic diversity of the populations under study. The choice of a high MAF threshold contributed to the high reliability of our genotyping work (Chen *et al.* 2014). This means however, that the resulting genotypic data are biased towards intermediate frequencies and we may therefore have missed rare alleles potentially affecting some phenotypes of interest; such as those previously reported for wood composition in *P. nigra* (Vanholme *et al.* 2013).

As a first application of the array, we performed the largest study undertaken to characterize the genetic structure of the Western range of *P. nigra*. We found unexpected genotype replication with most replications found within German populations which are best explained by duplication in nature due to vegetative propagation. These results are comparable to earlier published data (Storme *et al.* 2004; Smulders *et al.* 2008; Chenault *et al.* 2011) suggesting that in nature *P. nigra* is highly clonal along long tracts of riparian river basins that may stretch for several kilometres. As for other temperate, riparian species (*Populus* spp., *Salix* ssp., *Ulmus* ssp; Stueffer *et al.* 2002; Santos-del-Blanco *et al.* 2013; Lin *et al.* 2009; Fuentes-Utrilla *et al.* 2014), the rate of clonality observed could enable persistence of local populations under unfavourable conditions (Storme *et al.* 2004; Smulders *et al.* 2008; Chenault *et al.* 2011). ADMIXTURE analysis agreed with the PCA results indicating a high level of admixture and low level of genetic differentiation between populations. This finding was supported by the low Jost's D (Jost 2008) values (Table S13, Supporting information). Important gene flow usually observed in riparian populations such as poplars could explain our results (Imbert & Lefèvre 2003). Individuals belonging to the same river basin clustered together and cluster proximity reflected the close geographical proximity of the river basins within the same drainage system. This general structure is in accordance with previous *P. nigra* population genetic studies, although the river populations used only partially overlapped and different marker types were employed (Storme *et al.* 2004; Smulders *et al.* 2008; DeWoody *et al.* 2015). Besides a high level of admixture, a clear pattern of genetic differentiation remains between populations belonging to different drainage systems. This structure could also be explained by major geographical barriers limiting gene flow. The Alps are a strong factor which separate Italian populations from other Northern European populations. In France, this structure is governed by the major watersheds, namely the Rhine, Rhône and Loire/Allier, although some admixture exists between them. The most unique data concern the Dranse population located along a mountain stream in the Alps, which appears admixed mainly from the Rhine F and Ticino populations. The Italian populations are also structured along a latitudinal gradient and by contrast with Northern European and French populations present a low level of admixture. The Apennines; the contrasting environments along the Mediterranean gradient (max and min temperature, duration of daylight, global radiation) and longer geographical distances act as strong barriers to gene flow between Northern and Southern Italian populations.

Within the seven ancestral clusters identified using ADMIXTURE, the purple one is clearly admixed in all predefined populations and does not follow a particular geographical pattern; although the admixture does appear more important in French populations (Fig. 2). Admixture could be the result of introgression from cultivated poplars (Vanden Broeck *et al.* 2012) since: i. *P. nigra* and cultivated stands occupy the same habitat; ii. Cultivated clones can potentially hybridize with *P. nigra* as most of them are *P. x canadensis* interspecific hybrids involving different *P. nigra* European genetic pools and iii. These clones are very few in number, highly related and widely deployed throughout Europe. This last factor probably explains the strong differentiation of the 7th ancestral cluster.

Due to the high level of admixture, the 12 populations could be considered together; significantly increasing the statistical power of association studies to detect causal SNPs. This may also be supported by the large association population size and the use of appropriate association methods which explicitly account for population genetic structure. The extent of LD revealed in this study is probably overestimated due to the selection of SNPs showing a moderate to high MAF but it is still in the same range as that found in *P. trichocarpa* (Slavov *et al.* 2012). This information is important to plan whole genome association studies in *P. nigra*. The number of SNPs required to tag the entire *Populus* genome was estimated between 67K and 134K (Slavov *et al.* 2012; Geraldes *et al.* 2013). Based on the size of the genome used for these calculations (403 Mb), it follows that we need marker densities between 166 and 332 SNPs/Mb. The presence and distribution of polymorphisms seems not to be a limiting factor in the black poplar genome; given the high SNP frequency (1 SNP/ 58.6 b). The SNP frequency from this study is higher than those reported in previous studies (Marroni *et al.* 2012; Chu *et al.* 2014) because the analysis was targeted to the whole genome and included intergenic regions and pseudogenes.

Today, either GBS or HT-genotyping array technologies can be utilized to perform genome-wide association studies (GWAS) in poplar. GBS is a cost-effective method but the high level of missing data and the lack of reproducibility can result in a huge loss of data (Elshire *et al.* 2011). In case of GWAS performed in large populations, HT-genotyping array techniques could be more efficient if an international consortium designs an optimal SNP array for all poplar species.

In conclusion, we have described the first genome-wide resequencing study in an extensive collection of the European native black poplar, *P. nigra* (L.) and provided significant new genomic resources for this species of conservation and breeding significance throughout Europe and Eurasia. Our analysis has quantified LD decay and population structure providing essential keys to further population genetics in *P. nigra*.

We now have the resources in place to refine the location of already known QTLs in *P. nigra* through multi-pedigree genetic mapping (Giraud *et al.* 2014) or association studies within natural populations for which phenotypes are available (Rohde *et al.* 2011; DeWoody *et al.* 2015; Guet *et al.* 2015). We have demonstrated that the bead-chip array could be used for the characterization of genetic diversity in native populations of *P. nigra* or exploited in interspecific breeding pools; enabling the development of landscape-scale and genomic-based conservation strategies in the face of climate change.

Acknowledgements

Research was supported by (i) the European Commission through the projects, POPYOMICs (FP5-QLK5-CT-2002-00953), EVOLTREE (FP6-16322), NovelTree (FP7-211868), EnergyPoplar (FP7-211917), WATBIO (FP7-311929), (ii) INRA (AIP Bioressources), BBSRC through a PhD studentship to MRA. The authors acknowledge R. Smulders, C. Maestro and the different owners of black poplar genetic resources gathered in the EVOLTREE collection for allowing access to the referenced material and O. Forestier for the assistance of Guéméné-Penfao/ONF-State-Nursery in the management of the stoolbed. The authors thank M. Sabatti and M. Gaudet for providing Poli, 58-861 and six progenies' DNA and S. Fluch and M. Stierschneider for extracting most of the DNA. We are grateful to the CEA-IG/CNG teams of A. Boland (DNA and Cell Bank service) and MT. Bihoreau (Illumina Sequencing and Infinium genotyping facilities). We thank F. Bitton, R. El-Malki, and R. Bounon for providing Sanger data, A. Chauveau for performing sequencing and genotyping and D. Brunel for her help in designing the SNP detection procedure.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Bachlava E, Taylor CA, Tang S *et al.* (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS ONE*, **7**, e29814.

Benetka V, Novotná K, Štochlová P (2012) Wild populations as a source of germplasm for black poplar (*Populus nigra* L.) breeding programmes. *Tree Genetics and Genomes*, **8**, 1073–1084.

Cagelli L, Lefèvre F (1997) The conservation of *Populus nigra* L. and gene flow within cultivated poplars in Europe (updated). *Bocconea*, **7**, 63–75.

Chagné D, Crowhurst RN, Troggio M *et al.* (2012) Genome-Wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE*, **7**, e31745.

Chalhoub B, Belcram H, Caboche M (2004) Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant Biotechnology Journal*, **2**, 181–188.

Chamaillard S, Fichot R, Vincent-Barbaroux C *et al.* (2011) Variations in bulk leaf carbon isotope discrimination, growth and related leaf traits among three *Populus nigra* L. populations. *Tree Physiology*, **31**, 1076–1087.

Chen H, Xie W, He H *et al.* (2014) A high-density SNP genotyping array for rice biology and molecular breeding. *Molecular Plant*, **7**, 541–553.

Chenault NC, Arnaud-Haond SA, Juteau MJ *et al.* (2011) SSR-based analysis of clonality, spatial genetic structure and introgression from the Lombardy poplar into a natural population of *Populus nigra* L. along the Loire River. *Tree Genetics and Genomes*, **7**, 1249–1262.

Chu Y, Su X, Huang Q, Zhang X (2009) Patterns of DNA sequence variation at candidate gene loci in black poplar (*Populus nigra* L.) as revealed by single nucleotide polymorphisms. *Genetica*, **137**, 141–150.

Chu Y, Huang Q, Zhang B, Ding C, Su X (2014) Expression and molecular evolution of two *DREB1* genes in black poplar (*Populus nigra*). *PLoS ONE*, **9**, e98334.

Dalton-Morgan J, Hayward A, Alamery S *et al.* (2014) A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Functional & Integrative Genomics*, **14**, 643–655.

Delourme R, Falentin C, Fomeju BF *et al.* (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics*, **14**, 120.

DeWoody JD, Trewin HT, Taylor GT (2015) Genetic and morphological differentiation in *Populus nigra* L.: isolation by colonization or isolation by adaptation? *Molecular Ecology*, **24**, 2461–2655.

Dickmann DI, Kuzovkina J (2013) Poplars and willow of the world, with emphasis on silviculturally important species (Chapter 2). In: *Poplars and Willows in the World: Meeting the Needs of Society and the Environment* (eds Isebrands JG, Richardson J), 135 pp. FAO/IPC (Food and Agricultural Organization of the United States / International Poplar Commission), Rome, Italy.

El-Maki R (2013) Architecture génétique des caractères cibles pour la culture du peuplier en taillis à courte rotation, pH D thesis, University of Orléans, 242 p.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Evans LM, Slavov GT, Rodgers-Melnick E *et al.* (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics*, **46**, 1089–1096.

Fabbrini F, Gaudet M, Bastien C *et al.* (2012) Phenotypic plasticity, QTL mapping and genomic characterization of bud set in black poplar. *BMC Plant Biology*, **12**, 47.

Fuentes-Utrilla P, Valbuena-Carabaña M, Ennos R, Gil L (2014) Population clustering and clonal structure evidence the relict state of *Ulmus minor* Mill. in the Balearic Islands glacial history shape the genetic structure of Iberian poplars. *Molecular Ecology*, **21**, 3593–3609.

Geraldes A, Pang J, Thiessen N *et al.* (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources*, **11**, 81–92.

Geraldes A, Difazio SP, Slavov GT *et al.* (2013) A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Molecular Ecology Resources*, **13**, 306–323.

Giraud H, Lehermeier C, Bauer E *et al.* (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics*, **198**, 1717–1734.

Grattapaglia D, Silva Junior OB, Kirst M, Lima BM, de Faria DA, Pappas GJ (2011) High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biology*, **11**, 65.

Groenen MA, Megens HJ, Zare Y *et al.* (2011) The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*, **12**, 274.

Guerra F, Wegrzyn P, Sykes JL, Davis R, Stanton BJ, Neale DB (2013) Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist*, **197**, 162–176.

Guet J, Fabrini F, Fichot R, Sabatti M, Bastien C, Brignolas F (2015) Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). *Tree Physiology*, **35**, 850–863.

Harfouche A, Meilan R, Kirst M *et al.* (2012) Accelerating the domestication of forest trees in a changing world. *Trends in Plant Science*, **17**, 64–72.

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

Howe GT, Yu J, Knaus B *et al.* (2013) SNP resource for Douglas-fir: *de novo* transcriptome assembly and SNP detection and validation. *BMC Genomics*, **14**, 137.

Imbert E, Lefèvre F (2003) Dispersal and gene flow of *Populus nigra* (Salicaceae) along a dynamic river-system. *Journal of Ecology*, **91**, 447–456.

Ingvarsson PK, García MV, Hall D, Luquez V, Jansson S (2006) Clinal variation in *phyB2*, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (*Populus tremula*). *Genetics*, **172**, 1845–1853.

Isabel N, Lamothe M, Thompson SL (2013) A second-generation diagnostic single nucleotide polymorphism (SNP)-based assay, optimized to distinguish among eight poplar (*Populus* L.) species and their early hybrids. *Tree Genetics and Genomes*, **9**, 621–626.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

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.

Kennard RW, Stone LA (1969) Computer Aided Design of Experiments. *Technometrics*, **11**, 137–148.

Lefèvre F, Barsoum N, Heinze B *et al.* (2001) *In Situ Conservation of Populus Nigra*, 58 pp. International Plant Genetic Resources Institute, Rome, Italy.

Lepoittevin C, Bodénés C, Chancerel E *et al.* (2015) Single-nucleotide polymorphism Discovery and validation in high-density SNP array for genetic analysis in European White Oaks. *Molecular Ecology Resources*, **15**, 1446–1459.

Li X, Han Y, Wei Y *et al.* (2014) Development of an Alfalfa SNP Array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS ONE*, **9**, e84329.

Lin J, Gibbs JP, Smart LB (2009) Population genetic structure of native versus naturalized sympatric shrub willows (Salix; Salicaceae). *American Journal of Botany*, **96**, 771–785.

Lin CY, Xing G, Xing C (2012) Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity*, **109**, 401–402.

Livingstone D, Royaert S, Stack C *et al.* (2015) Making a chocolate chip: development and evaluation of a 6K SNP array for Theobroma cacao. *DNA Research*, **22**, 279–29.

Ma T, Wang J, Zhou G *et al.* (2013) Genomic insights into salt adaptation in a desert poplar. *Nature Communications*, **4**, 2797.

Mangin BA, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, **108**, 285–291.

Mantello CC, Cardoso-Silva CB, da Silva CC *et al.* (2014) *De Novo* assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. *PLoS ONE*, **9**, e102665.

Marroni F, Pinosio S, Morgante M (2012) The quest for rare variants: pooled multiplexed next generation sequencing in plants. *Frontiers in Plant Science*, **3**, 133.

McKown AD, Klápště J, Guy RD *et al.* (2014) Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytologist*, **201**, 1263–1276.

Monclus R, Leplé JC, Catherine Bastien C *et al.* (2012) Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in *Populus* spp. *BMC Plant Biology*, **12**, 173.

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

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330.

Novaes E, Osorio L, Drost DR *et al.* (2009) Quantitative genetic analysis of biomass and wood chemistry of *Populus* under different nitrogen levels. *New Phytologist*, **182**, 878–890.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

Peace C, Bassil N, Main D *et al.* (2012) Development and evaluation of a genome-Wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLoS ONE*, **7**, e48305.

Porth I, Klapšte J, Skyba O *et al.* (2013) Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist*, **200**, 710–726.

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. In Proceedings of the13 Annual International conference on Intelligent Systems for Molecular Biology (ISMB-05). Detroit, Michigan.

Rae AM, Pinel MPC, Bastien C *et al.* (2008) QTL for yield in bioenergy Populus: identifying GxE interactions from growth at three contrasting sites. *Tree Genetics and Genomes*, **4**, 97–112.

Ranjan P, Yin T, Zhang X *et al.* (2010) Bioinformatics-Based Identification of Candidate Genes from QTLs Associated with Cell Wall Traits in *Populus*. *BioEnergy Research*, **3**, 172–182.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohde A, Storme V, Jorge V *et al.* (2011) Bud set in poplar - genetic dissection of a complex trait in natural and hybrid populations. *New Phytologist*, **189**, 106–121.

Santos-del-Blanco L, de Lucas AI, González-Martínez SG, Sierra-de-Grado R, Hidalgo E (2013) Extensive Clonal Assemblies in *Populus alba* and *Populus x canescens* from the Iberian Peninsula. *Tree Genetics and Genomes*, **9**, 499–510.

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology Evolution and Systematics*, **38**, 595–619.

Schilling MP, Wolf PG, Duffy AM *et al.* (2014) Genotyping-by-sequencing for *Populus* population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS ONE*, **9**, e95292.

Sim SC, Van DA, Stoffel K *et al.* (2012) High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE*, **7**, e45520.

Slavov GT, DiFazio SP, Martin J *et al.* (2012) Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, **196**, 713–725.

Smulders MJM, Cottrell JE, Lefèvre F *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 Ecology and Management*, **255**, 1388–1399.

Stanton BJ, Serapiglia MJ, Smart LB (2013) The domestication and conservation of *Populus* and *Salix* genetic resources. In: *Poplars and Willows: Trees for Society and the Environment* (eds Isebrands JG, Richardson J) chapter 4, pp. 124–199. The Food and Agriculture Organization of United Nations and CABI, Rome, Italy.

Stölting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.

Storme V, Vanden Broeck A, Ivens B *et al.* (2004) Ex-situ conservation of black poplar in Europe: genetic diversity in nine gene bank collections and their value for nature development. *Theoretical and Applied Genetics*, **108**, 969–981.

Stueffer IF, Ershamber B, Huber H, Suzuki I (2002) The ecology and evolutionary biology of clonal plants: an introduction to the proceedings of Clone-2000. *Evolutionary Ecology*, **15**, 223–230.

Tuskan GA, DiFazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Vanden Broeck A, Villar M, Van Bockstaele E, Van Slycken J (2005) Natural hybridization between cultivated poplars and their wild relatives: evidence and consequences for native poplar populations. *Annals of Forest Science*, **62**, 601–613.

Vanden Broeck A, Cox K, Michiels B, Verschelde P, Villar M (2012) With a little help from my friends: hybrid fertility of exotic *Populus x canadensis* enhanced by related native *Populus nigra*. *Biological Invasions*, **14**, 1683–1696.

Vanholme B, Cesarino I, Goeminne G *et al.* (2013) Breeding with rare defective alleles (BRDA): a natural *Populus nigra* HCT mutant with modified lignin as a case study. *New Phytologist*, **198**, 765–776.

VanRaden PM (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**, 4414–4423.

Verde I, Bassil N, Scalabrin S *et al.* (2012) Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS ONE*, **7**, e35668.

Viger M, Rodrigues-Acosta M, Rae AM, Morison JIL, Taylor G (2013) Towards improved drought tolerance in bioenergy crops: QTL for carbon isotope composition and stomatal conductance in *Populus*. *Food and Energy Security*, **2**, 220–236.

Zhang HB, Zhao X, Ding X, Paterson AH, Wing RA (1995) Preparation of megabase-size DNA from plant nuclei. *The Plant Journal*, **7**, 175–184.

Zhou L, Holiday JA (2012) Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, **13**, 703.

P.F.R., G.Z., V.J., S.G., V.S., V.G., A.B. -Sanger resequencing and SNP identification. P.F.R., G.Z., V.J., S.G., V.S., A.B., M.M. –N.G.S. resequencing and SNP identification. P.F.R., V.J., V.S., G.Z., M.V., A.P., G.T. -Design of the SNP array. M.Vil -Collecting of *P. nigra* samples. C.B., G.T. -Design of the population sampling. P.F.R., M.C.L., F.C., M.M. Coordination of NGS resequencing work. P.F.R., M.C.L. -Coordination of the genotyping work. C.A., S.S., E.D.P.-Bioinformatics, data basing. P.F.R., P.P., V.G. -Analysis of genotypic data. V.J., V.S., C.B.-Population genetics analysis. P.F.R., G.Z., V.J., V.S., C.B.-Writing of the manuscript. M.Vil, G.T., M.R.A. -Revision of the manuscript.

Data accessibility

Collections of SNPs from HT-sequencing are available at http://datadryad.org/review?doi=doi:10.5061/dryad.dk6gj.2

Collections of SNPs from Sanger resequencing are available at https://urgi.versailles.inra.fr/GnpSNP/snp/createSubSnpSelect.do

Primers from the Sanger Sequencing project are listed in the supporting information.

The *P. nigra* reference sequence is available at http://dx.doi.org/10.5061/dryad.dk6gj.2 and the raw sequencing data are deposited in the Sequence Read Archive (SRA) under Study Accession number SRP067688 within BioProject PRJNA303130

The genotyping data and input files for ADMIXTURE and PCA analysis are available at http://dx.doi.org/10.5061/dryad.dk6gjcoorige.2

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 SNP-panel discovery and list of genotyped *Populus nigra* individuals.

Table S2 Primer pairs developed within genes for Sanger resequencing and SNP collections.

Table S3 List of candidate regions and candidate genes based on the location of QTL hot spots for rust resistance, drought stress, bud phenology, wood composition and transcriptome studies.

Table S4 Alignment results of the Poli, BEN3, BDG and 71077-308 short reads onto the *Populus nigra* reference (389 Mb).

Table S5 Origin and number of SNPs included in the 12 000 BeadChip array.

Table S6 List of SNPs included in the 12 000 BeadChip array.

Table S7 Performance of the BeadChip array.

Table S8 Comparison of genotyping data and Sanger data.

Table S9 Comparison of genotyping data and NGS data.

Table S10 Genomic position and gene assignation of the 8259 useful SNPs.

Table S11 List and origin of unexpected replicates.

Table S12 Chromosomal distribution of SNP numbers, SNP distances and SNP densities.

Table S13 Pairwise Jost's D values for *Populus nigra* populations.

Methods S1 DNA extraction and Sanger sequencing of gene amplicons.

Methods S2 Calculation of Illumina sequencing accuracy.

Methods S3 Validation and Origin of replicates data with SR genotyping.

Fig. S1 Test of SNP segregation conformity within eight progenies belonging to a 3x3 factorial mating design.

Fig. S2 Genomic distribution of SNPs detected for the development of the 12k bead-chip array. The coloured bars around the plot represent the 19 *Populus* chromosomes (unit used is 2 Mb).

Fig. S3 Chromosomal distribution of SNP densities and summary of QTL locations for wood composition, bud phenology, water-use efficiency and rust resistance in the poplar genome.

Fig. S4 Principal component analysis: The first, second and third axes explain 2.39%, 1.89% and 1.71% of the total variance respectively.

Fig. S5 Distribution of Minor Allele Frequencies (MAF) for 7896 SNPs in seven clusters and the association population (706 individuals).