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

Patricia P. Faivre-Rampant, Giusi Zaina, Véronique V. Jorge, Stefania Giacomello, Vincent Segura, Simone Scalabrin, Vanina Guérin, Emanuele de Paoli, Christelle Aluome, Maud Viger, et al.

▶ To cite this version:

HAL Id: hal-02636258
https://hal.inrae.fr/hal-02636258v1

Submitted on 4 Oct 2024

Università degli studi di Udine

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

MOLECULAR ECOLOGY RESOURCES

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

1 **New resources for genetic studies in *Populus nigra*: genome wide SNP**

2 **discovery and development of a 12k Infinium array**

3

4 P. Faivre-Rampant[*1], G. Zaina[*2], V. Jorge[3], S. Giacomello[4], V. Segura[3], S. Scalabrin[4], V.

5 Guérin[3], E. De Paoli[4], C. Aluome[1,3], M. Viger[5], F. Cattonaro[4], A. Payne[5], P.

6 PaulStephenRaj[1], MC. Le Paslier[1], A. Berard[1], M.R. Allwright[5], M. Villar[3], G. Taylor[5], C.

7 Bastien[3], M. Morgante[2,4]

8

9 [1] INRA, US1279 EPGV, CEA-IG/CNG, F-91057 Evry, France

10 [2] Dipartimento di Scienze Agrarie e Ambientali, University of Udine, via delle Scienze 208,

11 33100 Udine, Italy

12 [3] INRA, UR 0588 AGPF, Centre INRA Val de Loire, 2163 avenue de la Pomme de Pin, CS

13 40001 – Ardon 45075 Orléans, France

14 [4] IGA, Parco Scientifico e Tecnologico Luigi Danieli, via Jacopo Linussio 51, 33100 Udine,

15 Italy

16 [5] Centre For Biological Sciences, University of Southampton, Life Sciences, SO17 1BJ

17 Southampton, UK

18

19 Author for correspondence:

20 *Patricia Faivre Rampant*

21 *Tel: +33 1 60 87 39 32*

22 *Email: faivre@versailles.inra.fr*

23

24   * These authors contributed equally to this work.

25

26   **Key words**: *Populus nigra*, large scale SNP discovery, HT genotyping design, Population

27   genetics.

28

29   **Running tittle**: *Populus nigra*'s SNP: Discovery and validation

30

31   **Abstract**

32   Whole genome resequencing of 51 *Populus nigra* (L.) individuals from across Western

33   Europe was performed on Illumina platforms. A total number of 1,878,727 SNPs distributed

34   along a *P. nigra* reference sequence were identified. The SNP calling accuracy was validated

35   by comparison with Sanger sequencing data. SNPs were selected within 14 previously

36   identified QTL regions, 2916 expressional candidate genes related to rust resistance, wood

37   properties, water-use efficiency and bud phenology, and 1732 genes randomly spread across

38   the genome. Over 10,000 SNPs were filtered for the construction of a 12k Infinium

39   BeadChip array dedicated to association mapping. The SNPs genotyping assay was

40   performed with 888 *P. nigra* individuals. The genotyping success rate was 91%. Our high

41   success rate was due to the discovery panel design and the stringent parameters applied for

42   SNP calling and selection. In the same set of *P. nigra* genotypes, linkage disequilibrium

43   throughout the genome decayed on average within 5 to 7 kb to half of its maximum value.

44   As application test, ADMIXTURE analysis was performed with a selection of 600 SNPs

45   spread out on the genome and 706 individuals collected along 12 river basins. The admixture

46   pattern was consistent with genetic diversity revealed by neutral markers and geographical

47   distribution of the populations.

48   These newly developed SNP resources and genotyping array provide a valuable tool for

49   population genetic studies and identification of QTLs through natural-population based

50   genetic association in *P. nigra*.

51

52

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 and Kuzovkina, 2013). As a pioneer species, *P. nigra* plays an important role in the establishment of riparian ecosystems (Imbert and Lefèvre, 2003), where it can be found as isolated trees and in pure or mixed stands. Considered as threatened throughout its natural range by anthropogenic disturbances of the river bank and gene introgression from cultivars *(P. deltoides* x *P. nigra*) and from the worldwide spread out fastigiated form *P. nigra* var *italica*, (Cagelli and Lefèvre 1997; Vanden Broeck *et al.,* 2005), 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 coppice 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 parental 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 was also reported in other poplar species (Ingvarsson *et al.,* 2006, Keller *et al.,* 2010, Viger *et al.*, 2013) and also in other temperate widespread forest trees (Savolainen *et al.,* 2007). Past adaptation processes have most likely

77   generated wide reservoirs of standing genetic variation for many other adaptive traits in

78   black poplar.

79   One main challenge is to identify loci/genes that underlie this phenotypic variation.

80   Such information can then be used to access and manage genetic diversity and develop

81   adapted marker-assisted selection schemes (Harfouche *et al.,* 2012). Association genetics is

82   a promising method of achieving this goal in woody species with a long life cycle, late

83   expression of important traits and considerable population genetic diversity (Neale and

84   Savolainen, 2004; Neale and Kremer, 2011). The development of High Throughput (HT)

85   genotyping tools is undoubtedly a prerequisite for such an approach. Single nucleotide

86   polymorphisms (SNPs) are a suitable and very attractive genetic marker for this purpose. It

87   is now well established that HT DNA sequencing technologies are powerful tools enabling

88   the rapid discovery of large numbers of SNPs. Different options have been deployed in plants

89   including tree species, including RNA sequencing *i.e.* HT-sequencing at the transcriptome

90   level (Parchman *et al.,* 2010; Geraldes *et al.,* 2011; Howe *et al.,* 2013; Mantello *et al.,* 2014),

91   and targeted sequencing, *i.e.* HT-sequencing of particular (captured) genomic regions such

92   as the gene-enriched portion (Zhou and Holiday, 2012) and restricted genomic DNA

93   (Grattapaglia *et al.,* 2011, Schilling *et al.,* 2014). For species with a relatively small genome,

94   like *Populus* sp. (500Mb), whole genome HT-sequencing can be sensibly achieved (Slavov

95   *et al.,* 2012; Evans *et al.,* 2014). Recently, studies have demonstrated the usefulness of both

96   HT-sequencing and SNP arrays to assess candidate gene association genetics in natural

97   populations of *P. trichocarpa* (Porth *et al.,* 2014; Mc Krown *et al.,* 2014). The success of

98   association studies mainly depends on the availability of SNPs, the extent of linkage

99   disequilibrium (LD), the extent of phenotypic variation of interest and the genetic structure

100   in the association population. In *P. nigra*, these determinants are poorly documented. Indeed,

101   studies were limited to relatively few SNPs identified within 2 to 39 genes, and LD was

102 reported to decay within 300 to 1000 bp (Chu *et al.,* 2009; Marroni *et al.,* 2012; Guerra *et*

103 *al.,* 2013; Chu *et al.,* 2014).

104   In order to perform association studies in *P. nigra*, our aims were to identify SNP at

105 whole-genome scale and to develop a SNP bead chip array. Due to the expected rapid decay

106 of LD in most undomesticated tree species, we opted for a candidate-genomic-region

107 approach that focused for leaf rust resistance, bud phenology, water-use efficiency and wood

108 chemistry on both QTL intervals identified in *P.nigra* mapping pedigrees (Rohde *et al.*,

109 2011, Fabbrini *et al.*, 2012, Elmalki, 2013, Guet *et al.*, 2015) and candidate genes underlying

110 QTLs in other poplar species (Novaes *et al.,* 2009; Rajan *et al.,* 2010; Rae *et al.,* 2008;

111 Monclus *et al.,* 2012; Viger *et al*., 2013). SNP outside the candidates were also selected to

112 provide genomic control tools to characterize neutral diversity and detect population

113 structure. To reach this objective, we first created a *P. nigra* reference sequence using the *P.*

114 *trichocarpa* genome sequence as a template (Tuskan *et al*. 2006) and identified a large set

115 of SNPs at the whole genome scale by HT-resequencing of 51 *P. nigra* genomes. Second,

116 we defined a SNP selection strategy in order to design a useful SNP array for candidate-

117 based association studies in natural populations. Third, the usefulness of the array was

118 evaluated by genotyping 888 *P. nigra* individuals. Data analysis focused on LD decay with

119 distance and on the genetic structure of a large *P. nigra* association population sampled in

120 12 river basins over Western Europe.

121

122 **Material and methods**

123 *SNP discovery and selection*

124   *Discovery panel and whole genome re-sequencing*

125   A SNP discovery panel of 51 individuals selected as representative of the genetic

126 diversity of an association population covering the range of the black poplar in Western

127 Europe was used for HT-genome sequencing (Table S1).

128   Nuclear DNA was isolated from young leaves as described by Zhang *et al*. (1995)

129 and Chalhoub *et al*. (2004). Whole-genome re-sequencing was performed at the Institute of

130 Applied Genomics (IGA, Udine, Italy) and the INRA-EPGV/CEA-IG/CNG (Institut

131 National de la recherche Agronomique-Etude du Polymorphisme des Génomes

132 Végétaux/Commissariat à l'Energie Atomique-Institut de Génomique/Centre National de

133 Génotypage, Evry, France) facilities using either a GAII analyzer or Hiseq 2000 Illumina

134 platforms (Inc. San Diego, CA, USA). Paired-end sequencing libraries were prepared

135 following the "Illumina Paired-End Sample Preparation" protocol, using an insert size

136 spanning from 300 to 600 bp. Paired-end runs were performed for 75, 100, 110 or 114 cycles

137 following Illumina instructions (Table 1). Illumina sequencer analyzer provided a quality

138 score (Qscore) for each base, an average Qscore value was assigned to each read. Reads with

139 Qscore values >30 were considered as good sequences.

140   Four individuals covering the wide Western latitudinal range of *P. nigra*, Poli (South-

141 Italy), BEN3 (Spain), Blanc de Garonne (BDG) (South-West-France) and 71077-308 (East-

142 France) were sequenced at coverage >25x (Tables 1, S1). Our objective was twofold; to

143 maximize the genetic variation among individuals and to identify reliable SNPs. Forty-seven

144 individuals covering the European latitudinal range were selected and sequenced at lower

145 coverage (Tables 1, S1) in order to maximize the discovery of informative SNPs.

146

147   *P. nigra reference sequence*

148   To avoid confusion between interspecific polymorphisms between *P. trichocarpa*

149 and *P. nigra* species and prompt the detection of intraspecific polymorphisms within *P. nigra*

150   (Isabel *et al.*, 2013), we created a *P. nigra* reference sequence using short reads of the

151   genotype 71077-308 (27x). This genotype was chosen for its read Qscore > 32. Paired-end

152   reads were aligned onto the *P. trichocarpa* genome V2.0 (Tuskan *et al.*, 2006). Indeed, pilot

153   analyses on Sanger-sequenced BAC inserts showed the feasibility of using the *P.*

154   *trichocarpa* genome sequence as a template for *P. nigra* (Zaina, unpublished data). The

155   mapping of raw short reads was performed with the CLC Genomics Workbench v.4 (CLC

156   Bio, Aarhus, Denmark). Mapping parameters were given in figure 1. Only paired-end reads

157   that aligned to a unique location of the genome were considered. Duplications and repetitions

158   were identified with RepeatScout using default parameters (Price *et al.*, 2005). Due to

159   computing constraints, only the first 40 scaffolds were extracted as part of the *P. nigra*

160   reference sequence to be used in the SNP calling.

161

162   ***Strategy of SNP detection for designing the array***

163   A multi-step strategy was designed to recover variants for the Illumina Infinium

164   iSelect HD Custom BeadChip technology. The paired-end raw sequences of the 4 genotypes

165   >25x were mapped separately onto the *P. nigra* reference sequence using the same procedure

166   adopted above to create the *P. nigra* reference itself, with the exception of similarity set to

167   0.95. Reads for the 47 remaining accessions were aligned similarly but as a joint set. SNP

168   detection was then performed on each of the 5 alignments, with the parameters detailed in

169   figure 1. To evaluate the accuracy of the SNPs calling a comparison with the SNPs detected

170   using ABI3730 Sanger sequencing was performed (Table S2, Methods S1 and S2).

171   Deletion-Insertion Polymorphisms (DIPs) were also detected to optimize SNP

172   selection for the array design. DIPs were detected using the CLC software v.4 (Fig. 1).

173   Given the objective of the SNP array, candidate genomic regions (14) were

174   considered on the basis of QTL for rust resistance, bud phenology in *P. nigra* and water-use

175   efficiency, wood properties in other *Populus* species (Fig. 3, Table S3). Candidate genes

176   (2916) for the same traits were also considered on the basis of transcriptome studies and the

177   literature (Fig. 3, Table S3). SNPs belonging to those candidate regions or genes were

178   considered for the subsequent selection. Additional SNPs were retained within gene models

179   (1732) spread across the poplar genome.

180   A pipeline written in Bash and Perl was set up to extract useful SNPs with 60-bp

181   flanking sequences. The pipeline rescued only loci whose flanking sequences did not contain

182   any SNP and/or DIP. If this was not possible, the pipeline was set to select SNPs with no

183   SNPs and/or DIPs within ± 10 bp of the target SNP. The pipeline also discarded the SNPs

184   within duplicated or repetitive regions.

185   A collection of SNPs detected by Sanger re-sequencing of full-length genes and gene

186   fragments obtained previously by University of Udine and INRA teams in the framework of

187   Popyomics and National projects were also considered (Method S1, Table S2).

188   The whole set of extracted SNPs was subjected to the Assay Design Tool by Illumina

189   (https://iCom.illumina.com) in order to score and validate the SNPs in terms of the bead-

190   chip performance. Final selection was performed to reach the desired 11,999 beads. This

191   final selection was based on the SNP location in the genome (Table S3): i. 80 SNPs/Mb were

192   retrieved from the QTL area showing a considerable effect (the phenotypic variance

193   explained by the QTL was set at > 10%) ii. 20 SNPs/Mb were retrieved from the QTL area

194   showing a low or moderate effect, iii. 5 SNPs/Mb were retrieved from non-QTL regions.

195   SNPs requiring a single bead type (Infinium II) were also preferred to maximize the number

196   of loci on the chip. In a few regions, the final target could not be reached with the current

197   criteria, which were thus gradually relaxed to meet the targets. Moreover, for functional

198   candidate genes for rust resistance and bud phenology, more than one SNP were selected per

199   gene with the same criteria.

200

201 ***Genotyping assay***

202 ***Plant material***

203   A set of 888 individuals comprising 838 native *P. nigra* individuals (originating from

204 12 river basins and collected in the western part of Europe (Tables 2 and S1), of which most

205 belonged to the Europop (Cotterell *et al.,* 2004) and the French National collections, and 50

206 full sib progenies were used in this study (Table S1). Among the 838 native *P. nigra*, 814

207 were part of the European association population established in the framework of the EU

208 projects Popyomics, Evoltree, NovelTree and EnergyPoplar, and had already been

209 genotyped with SSR markers (Storme *et al.,* 2004, DeWoody *et al*., 2015; Jorge unpublished

210 data). Within the total set, 11 individuals were used as parents in 9 different crosses and 2 to

211 6 progenies per cross were genotyped to facilitate and validate SNP clustering.

212

213 ***SNP genotyping***

214   One sample (BDG) was repeated 14 times and used for technical control. DNA

215 samples from 24 individuals were included twice to assess the repeatability of allele calls.

216 SNP genotyping was conducted on the Illumina Platform at CEA-IG/CNG by INRA-EPGV

217 according to the standard protocol of Illumina. Genotypes were recovered with Genotyping

218 module v 1.9.4 (Genome Studio software v 2011.1, Illumina Inc.). Clusters were generated

219 using a GenCall score cut-off of 0.15 as recommended by Illumina. The GenCall score,

220 estimated for each data point (SNP $\times$ individual sample), implemented by the Genome

221 Studio software reflected the position of the data point within the genotype cluster.

222 Genotypes with lower GenCall scores are located further from the center of the genotype

223 cluster and had lower reliability. Only those individuals with $> 95\%$ call rates were selected

224 (*i.e.* the proportion of individual samples successfully genotyped in a locus). SNP clusters

225   were automatically generated and then the quality of the 3 expected clusters of each SNP

226   was inspected visually. Subsequent adjustment of the cluster calling was performed if

227   needed.

228

229   *Linkage disequilibrium and population structure*

230   To estimate LD decay and analyze population structure on neutral genetic diversity,

231   SNPs and individuals were filtered according to several criteria. First, SNPs and individuals

232   with missing data above 10% were discarded. Then, segregation and linkage conformity was

233   checked within a 3x3 factorial mating design (Method S3). Finally, SNPs showing a

234   significant departure from the Hardy-Weinberg equilibrium within more than 6 populations

235   were discarded. LD between all pairs of SNPs was estimated as the square of the allelic

236   correlation in R (R Core Team, 2014).

237   Population structure was investigated using the software ADMIXTURE (Alexander *et*

238   *al.,* 2009), with K ancestral population ranging from 1 to 15. Since we used a candidate-

239   based approach, the selected SNPs were not evenly spread throughout the genome. To

240   account for such variation in SNP density across the genome, we sampled several subsets of

241   SNPs. These subsets were sampled by chromosome, taking into account physical

242   chromosome length and the desired final number of SNPs using different approaches:

243-   2000-LD: 2000 SNPs minimizing the LD between SNPs by applying the Kennard and Stone

244   algorithm (Kennard and Stone, 1969) to the LD matrix by chromosome,

245-   600-LD: same as above but with a total target of 600 SNPs,

246-   600-dist: 600 SNPs well scattered by applying the Kennard and Stone algorithm to the

247   physical distance matrix by chromosome,

248-   600-random: 600 SNPs randomly sampled by chromosome.

249   These 4 subsets were compared together and to the total set of high-quality SNPs to evaluate

250   population structure by cross-validation in ADMIXTURE. The set that minimized the cross-

251   validation error was selected to analyze population structure. The optimal number of groups

252   was also determined by cross-validation for this set. The optimal set of SNPs according to

253   the cross-validation in ADMIXTURE was used to carry out Principal Component Analysis

254   (PCA) in R (R Core Team, 2014) as a complementary analysis of population structure. We

255   used the optimal set of SNPs to estimate a measure of LD corrected for the bias attributed to

256   population structure and cryptic relatedness as proposed by Mangin *et al.,* (2012). Briefly,

257   we used the optimal set of SNPs to compute a genomic relationship matrix between

258   individuals (Van Raden, 2008), and used this matrix to estimate a corrected measure of LD

259   defined as the squared partial allelic correlation between SNPs (Lin *et al.,* 2012). The

260   relationship between LD and physical distance was assessed following the model of Hill and

261   Weir (1988) in order to determine the distance where LD decays to half its maximum value.

262

263   **Results**

264   Illumina next generation DNA sequencing technology was used to re-sequence 4 *P.*

265   *nigra* genotypes (71077-308, BDG, BEN3 and Poli) at coverage >25x and 47 other

266   genotypes at lower coverage. The read data and relative raw coverage obtained for each

267   genotype are reported in Table 1.

268

269   *SNP detection*

270   *P. nigra reference sequence*

271   The sequence data obtained from the clone 71077-308 were selected due to their

272   good quality to produce a reference sequence for *P. nigra* species, exploiting a mapping

273   approach *versus* the *P. trichocarpa* genome sequence v2.0. We previously proved the

274   feasibility of this approach by mapping the short reads of another *P. nigra* genotype (the

275   Spanish clone BEN3) *versus* two *P. nigra* BAC-clone sequences and *versus* the *P.*

276   *trichocarpa* sequence portions corresponding to the BAC inserts. In the intraspecific

277   alignment, the BAC sequences were covered for 98% of their length, as expected, and in the

278   interspecific alignment, 75% of the corresponding *P. trichocarpa* regions were covered

279   (Zaina, unpublished data). In the present work, the 71077-308's short reads covered 79% of

280   the *P. trichocarpa* genome sequence V2.0. After mapping, we considered only the *consensus*

281   specific to the first forty scaffolds, which resulted in a sequence 388,572,533 bp long (gaps

282   included), representing the sequence used hereafter as the *P. nigra* reference sequence.

283

284   *SNP calling*

285   We used the *P. nigra* reference sequence obtained to map the paired-end reads of

286   71077-308, BDG, BEN3 and Poli (>25x). Approximately 60% input reads of 71077-308,

287   BDG and Poli were mapped to a unique position in the reference sequence. The exception

288   of BEN3 with a lower amount of mapped reads (42%) was explained by the lower quality

289   score (reads average Qscore < 26) of its reads compared to the others (Table S4). In addition

290   to the four alignments produced above, the reads derived from the re-sequencing of the 47

291   individuals (<25x) were mapped as a whole against the *P. nigra* reference sequence to obtain

292   a fifth alignment.

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

294   the procedure summarized in figure 1. The total number of SNPs detected in each alignment

295   along the *P. nigra* reference sequence is shown in Table 3, and referred to as input SNPs.

296   The figure 2 shows the distribution of the input SNPs detected through the 5 alignments

297   across the main 19 chromosomes of the reference *P. nigra*. Out of 388,572,533 bp of the *P.*

298 *nigra* reference sequence 110,098,472 bp were covered by the 4 genotypes and provided a

299 total of 1,878,727 SNPs. The SNP frequency resulted to be 1 polymorphism every 58.6 bp.

300   To estimate SNP calling accuracy, we compared the SNPs identified within the 18

301 candidate genes for light signaling pathway (Table S2) resulting from the re-sequencing,

302 using both Sanger and Illumina methods. A total of 96,164 sites were analyzed, including

303 1186 polymorphic sites from the Sanger SNP detection. The Illumina SNP detection resulted

304 in 92.9% Sensitivity, 99.8% Specificity and 99.7% Accuracy, and provided 141 false

305 positives (*i.e.* SNPs identified in Illumina data but not in Sanger data), corresponding to a

306 10.6% False Discovery rate (Method S2).

307

308 ***Development of the 12k Infinium BeadChip array***

309   A total of 296,964 SNPs were retrieved from the 47 genotypes in the candidate

310 regions while the other 4 genotypes provided 344,709 (Poli), 112,262 (BEN3), 174,035

311 (BDG) and 155,846 (71077-308) SNPs within the same regions (Table 3). The differences

312 in the number of loci between the 5 alignments were consistent with the depth-coverage and

313 read quality of the different genotypes. A map was created by using the IUPAC codes to

314 group all the SNPs belonging to the different genotypes within the candidate loci. The map

315 was integrated with the DIPs identified in the same five alignments (data not shown), to

316 improve the further selection of SNPs for an efficient bead-chip array design (*i.e.* no

317 polymorphisms within the SNP flanking sequences). Eventually, 189,616 SNPs, which

318 correspond to 1 SNP every 1159 bp in the candidate regions and genes, were retained. This

319 last set of 189,616 SNPs was subjected to the Illumina Assay Design Tool (ADT) to test for

320 suitability with the bead-chip design. 133,821 SNPs passed the test, showing an ADT score

321 $\geq$ 0.6 (*i.e.* the score threshold recommended by Illumina) (Table S5). A set of 669 SNP

322 distributed onto the non-candidate regions were selected with the same criteria (Table S5).

323   In addition to the SNPs identified by the Illumina HTre-sequencing, 4691 SNPs from

324   the Sanger re-sequencing of candidate genes in *P. nigra* were considered (Fig. 1, Table S2).

325   After filtering selection detailed in figure 1, 2690 Sanger SNPs were available. Thus, the

326   very last pool of SNPs consisted of 137,180 loci. To get the desired number of 11,999 beads

327   required for the Illumina bead-chip array, the SNPs were reduced to 10,331 loci according

328   to the stringent criteria detailed in Material and Methods (Tables S6, S7). Among them,

329   6311were located in QTL intervals.

330

331   *Infinium BeadChip array performance*

332   Of the 10,331 SNPs, 9127 included in the bead pool (88%) remained in the array

333   after Illumina technical dropout. Eight samples were excluded for technical errors and 19

334   were excluded due to low call rate. The selection finally revealed 861 genotypes with a call

335   rate $\geq$ 0.95. Each cluster was then inspected manually. SNPs were classified into different

336   classes: polymorphic, monomorphic and failed (Table S8). Our validation showed 8322 well

337   clustered SNPs leading to a chip success rate estimated at 91%; 8259 of them were

338   polymorphic (90%). The reproducibility rate was 100% when we compared the 12 inter-

339   plate controls. The same rate was obtained from the comparison of i. biological replicates of

340   BDG and 1 inter-plate control, ii. duplicates of 24 genotypes. Heritability-based SNP

341   validation was estimated to assess SNP assay quality. This was defined as the number of

342   offspring genotypes that agreed with the expected inheritance over the total number of

343   possible genotype calls. In 9 families, there were 608 Mendelian transmission

344   inconsistencies out of the 411,877 allelic transmissions assayed, *i.e.* a genotyping miscall

345   rate of 0.15% (ranging from 0.08% to 0.21%). We observed that 1.65% of SNPs had

346   segregating errors.

347   A set of 259 SNPs from Sanger data was used to validate the efficiency of SNP genotyping

348   in 10 individuals for which both Infinium and Sanger sequence data were available. We

349   observed a very high rate of concordance (96%-99%) (Table S9). For 71077-308, BDG,

350   BEN3 and Poli, we then compared genotype calls from NGS re-sequencing data to genotype

351   calls from the chip. The concordance observed varied between 80% and 100% (Table S10).

352   Of the 8259 SNPs, 7186 were located within 4903 genes; and 1132 genes harbored more

353   than 2 SNPs (Table S11).

354

355   ***Application of the array***

356   ***Identification of clonal duplication***

357   Polymorphic sites (8259) were used to compute pair-wise similarity between all pairs

358   of individuals. This analysis identified 35 duplets, 9 triplets, 4 quadruplets, 2 septuplets, and

359   one duodeciduplet (Table S12). With the exception of 5 groups (3 duplets, one triplet and

360   one quadruplet), all the individuals belonging to the same group came from the same

361   population. Genotyping work performed with SSR markers was used to trace the origin of

362   these results (Method S3, Table S12). Redundant individuals were removed from the

363   individual data set for further analyses.

364

365   ***Population structure***

366   We applied additional filters on SNPs and individuals for genetic analyses. Data

367   Filtering on missing data (> 10%) resulted in discarding 13 SNPs and 26 individuals.

368   Additional SNPs were discarded: 216 SNPs due to segregation problems (missing or not-

369   expected genotyping class, segregation distortion and non-expected linkage, Fig S1) in

370   factorial mating design (data not shown) and 98 SNPs due to significant deviations from

371   Hardy-Weinberg equilibrium within at least 6 populations In the resulting set of individuals,

372   36 SNPs were monomorphic and were thus discarded from further genetic analyses. The

373   final data matrix included 7896 high-quality polymorphic SNPs genotyped in 706

374   individuals. Due to our biased sampling of SNPs within candidate regions (Fig. 3, Table

375   S13), we further selected several subsets of 600 and 2000 SNPs as being potentially better

376   distributed throughout the genome. The optimal number of ancestral clusters K=7,

377   corresponding to the lowest cross-validation error, was obtained with the set of 600 SNPs

378   selected (Fig. 4a). The corresponding admixture results are shown in Figures 4b. Basento

379   and Paglia populations from South and middle Italy emerged as distinct groups. For the other

380   populations a clear admixture pattern was revealed, although individuals from the same

381   populations still tended to cluster together. A principal component analysis on the same

382   optimal set of 600 SNPs confirmed the results from ADMIXTURE. Indeed a relatively clear

383   clustering of individuals according to their geographical origin was observed (Fig. S1).

384

385   *linkage disequilibrium*

386   As expected by the MAF (Minimum Allele Frequency) threshold (>0.2) applied to

387   select SNPs in our discovery panel, the MAF of 92% of the high-quality genotyped SNPs is

388   higher than 0.2 in the 7 admixture clusters. The frequency distribution of SNPs was more or

389   less even across different MAF classes and across ADMIXTURE clusters with the exception

390   of Italian clusters (Fig. S3). We calculated both LD and LD corrected for population structure

391   confounding between all pairs of SNPs. The relationship between LD and physical distances

392   was plotted and modeled (Fig. 5). As expected, the corrected LD decayed slightly faster than

393   the uncorrected LD with physical distance: the r^2 and corrected r^2 dropped to half their

394   maximum value within 5 and 7 kb, respectively.

395

396   **Discussion**

397   We reported the development of a high-quality SNP array in *P. nigra*. To our

398   knowledge, this is the first significant SNP resource that has been reported for black poplar.

399   As poplar has a relatively small genome (500 Mb), we decided to re-sequence the whole

400   genome instead of using a genome reduction procedure developed by Stölsting *et al.,* 2013.

401   In poplar, SNPs are mostly species specific (Isabel *et al.,* 2013), thus the available genome

402   of *P. trichocarpa* could not be used directly as a reference to detect SNPs. Nevertheless, we

403   were able to use it as a template to map the short reads of *P. nigra* to obtain a reference

404   sequence of the black poplar genome. Indeed, the alignment of paired-end reads allowed us

405   to obtain 389 x10^6 bp of *P. nigra* specific sequences (approximately 79% of the *P.*

406   *trichocarpa* genome). The excluded regions generally corresponded to variations between

407   the genomes of *P. trichocarpa* and *P. nigra*, which we expected to be mostly repetitive

408   regions as observed by Ma *et al.,* (2013), between the genomes of *P. euphratica* and *P.*

409   *trichocarpa*, or large insertion/deletions due to transposable elements as observed by Zaina

410   and Morgante, (unpublished results) among BAC insert sequences belonging to *P. nigra*, *P.*

411   *deltoides* and *P. trichocarpa*.

412   The comparison between the *P. nigra* reference sequence and 71077-308, BDG,

413   BEN3 and Poli genotypes provided the first *P. nigra* whole genome SNP collection. The

414   Italian genotype, Poli, contained more SNPs than the French and Spanish genotypes. This

415   result was consistent with their genetic distances to the 71077-308 used to build the *P. nigra*

416   genome reference (Jorge and Villar, unpublished results). Our procedure used to identify

417   SNPs from resequencing of 4 genotypes >25x and 47 genotypes <25x proved to be reliable,

418   reducing false discovery rate.

419   During our SNP selection process, most of the SNPs were lost during the final step,

420   *i.e.* the selection of SNPs with no polymorphisms in their 60-bp flanking sequences. This

421   can be explained by the high level of SNP frequency and heterozygosity in *P. nigra*. Hence

422 a huge collection of SNPs originating from complete genome coverage and a large SNP

423 discovery panel was required to reach our final target of 12k beads. According to Groenen

424 *et al.,* (2011), the number of SNPs should be at least 10 times higher than the number targeted

425 for the final chip. The good genotyping results demonstrated that the strategy developed to

426 detect and select SNPs was very effective, despite the lack of reference sequence for *P. nigra*.

427 The high rate of concordant data between genotyping and SNP calling from Sanger

428 sequencing and NGS genome sequencing, revealed the robustness of our selection criteria.

429 Our genotyping success rate (91%) exceeded those recorded for other plant species with the

430 same Infinium technology and in the same genotyping throughput range (6k-10k) (Chagné

431 *et al.,* 2012; Verde *et al.,* 2012; Bachlava *et al.*, 2012; Peace *et al.*, 2012; Sim *et al.,* 2012;

432 Delourme *et al.,* 2013; Li *et al.,* 2014; Dalton-Morgan, 2014; Lepoittevin *et al.*, 2015;

433 Livingstone *et al.*, 2015). The success of the SNP array was due to the composition of the

434 SNP discovery panel reflecting the genetic diversity of the populations under study. The

435 choice of a high MAF threshold contributed to the high reliability of our genotyping work

436 (Chen *et al.*, 2014); However, the resulting genotypic data are biased toward intermediate

437 frequencies and we may therefore have missed rare alleles potentially affecting some

438 phenotypes of interest, as has previously been reported for wood composition in *P. nigra*

439 (Vanholme *et al.*, 2013).

440 As a first application of the array, in the present work we performed the largest study

441 undertaken to characterize the genetic structure of the Western range of *P. nigra*. We found

442 unexpected replicated genotypes, most replications were found within German populations

443 and could be explained by duplication in nature due to vegetative propagation. The results

444 are comparable to the earlier published data (Storme *et al.,* 2004; Smulders *et al.,* 2008;

445 Chenault *et al.,* 2011), suggesting that in nature *P. nigra* is highly clonal along long tracts of

446 riparian river basins that may stretch for several kilometers. As for other temperate riparian

447   species (*Populus* spp., *Salix* ssp., *Ulmus* ssp; Stuefer *et al.,* 2002; Santos del Blanco *et al.,*

448   2013; Lin *et al.,* 2009; Fuentes-Utrilla *et al.,* 2014), the rate of clonality observed could

449   enable persistence of local populations under unfavorable conditions (Storme *et al.,* 2004;

450   Smulders *et al.,* 2008; Chenault *et al.,* 2011). ADMIXTURE analysis agreed with the PCA

451   results indicating high level of admixture and low level of genetic differentiation between

452   populations. This finding was supported by the low Jost's D values. Important gene flow

453   usually observed in riparian populations such as poplars could explain our results (Imbert

454   and Lefevre, 2003). Individuals belonging to the same river basin clustered together and

455   cluster proximity reflected the close geographical proximity of the river basins within the

456   same drainage system. This general structure is in accordance with previous *P. nigra*

457   population genetic studies, although the sets of populations used only partially overlapped

458   and marker systems were different (Storme *et al.,* 2004; Smulders *et al.,* 2008; DeWoody *et*

459   *al.,* 2015). Besides a high level of admixture, a clear pattern of genetic differentiation

460   remains between populations belonging to different drainage systems. This structure could

461   also be explained by major geographical barriers limiting gene flow. The Alps are a strong

462   factor which separates Italian populations from the rest of Northern Europe populations. In

463   France, this structure is governed by the major watersheds, namely the Rhine, Rhône and

464   Loire/Allier, although some admixture exists between them. The most original data concerns

465   the Dranse population located along a mountain stream of the Alps, which appears admixed

466   mainly from Rhine F and Ticino populations. The Italian populations are also structured

467   along a latitudinal gradient and, by contrast with Northern European and French populations,

468   present a low level of admixture. The Apennines, the contrasted environments of such

469   Mediterranean gradient (max and min temperature, duration of daylight, global radiation)

470   and longer geographical distances act as strong barriers to gene flow between Northern and

471   Southern Italian populations.

472   In the 7 ancestral clusters identified using ADMIXTURE, the purple one is clearly

473   admixed in all predefined populations, and do not follow a particular geographical pattern

474   although the admixture appears more important in French populations (Fig. 4). Admixture

475   could be due to introgression from cultivated poplars (Vanden Broek *et al.*, 2012) i. *P. nigra*

476   and cultivated stands occupy the same habitat; ii. cultivated clones potentially can hybridize

477   with *P. nigra* as most of them are *P. x canadensis* interspecific hybrids involving different

478   *P. nigra* European genetic pools and iii. these clones are very few, highly related and widely

479   deployed in whole Europe. This last reason probably could explain the strong differentiation

480   of the 7th ancestral cluster.

481   Due to the high level of admixture, the 12 populations could be considered together

482   to increase significantly the detection power of association tests, thanks to a large association

483   population size and appropriate association methods which explicitly take into account its

484   specific structure. The extent of LD revealed in this study is probably overestimated due to

485   the selection of SNPs showing a moderate to high MAF, but it was in the same range as that

486   found in *P. trichocarpa* (Slavov *et al.,* 2012). This information is important to develop whole

487   genome association in *P. nigra*. The number of SNPs required to tag the entire *Populus*

488   genome was estimated between 67K and 134K (Slavov *et al.,* 2012; Geraldes *et al.,* 2013).

489   Based on the size of the genome used for these calculations (403 Mb), this means that we

490   need densities between 166 SNPs/Mb and 332 SNPs/Mb. The presence and distribution of

491   polymorphisms seems to be not a limiting factor in the black poplar genome, given the high

492   values of SNP frequency (1 SNP/ 58.6 b). The SNP frequency from this study resulted to be

493   higher than those found in previous studies (Marroni *et al.,* 2012; Chu *et al.,* 2014) since the

494   analysis was targeted to the whole genome, including intergenic regions and pseudogenes.

495   Today either GBS or HT-genotyping array technologies can be proposed to perform

496   Genome-wide association studies (GWAS) in poplar. GBS is a cost-effective method but the

497   high level of missing data and the lack of reproducibility can result on a huge loss of data

498   (Elshire *et al*., 2011). In case of GWAS performing with large populations, the HT-

499   genotyping array techniques could be more efficient if an international consortium designs

500   an optimal SNP array for all the poplar species.

501   In conclusion, we have described the first genome-wide re-sequencing study in an

502   extensive collection of the European native black poplar, *P. nigra* (L.), providing significant

503   new genomic resources for this tree species of conservation and breeding significance

504   throughout Europe and Eurasia. Our analysis has quantified LD decay and population

505   structure providing essential keys to further population genetics in *P. nigra*.

506   We now have the resources in place to refine location of already known QTLs in *P.*

507   *nigra* through multi-pedigrees genetic mapping (Giraud *et al.*, 2014), or association studies

508   based on these natural populations for which phenotypes are available (Rohde *et al.*, 2011,

509   DeWoody *et al.*, 2015, Guet *et al.*, 2015). We demonstrated that the bead-chip could be used

510   for characterization of genetic diversity present in native populations of *P. nigra* or exploited

511   in interspecific breeding pools, enabling development of landscape-scale and genomic-based

512   conservation strategies in the face of climate change.

513

529   **References**
530

531   Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
532   unrelated individuals. *Genome Research* 19: 1655-1664.
533

534   Bachlava E, Taylor CA, Tang S, Bowers JE, Mandel JR, Burke JM, Knapp SJ. 2012. SNP
535   discovery and development of a high-density genotyping array for sunflower. *PLoS ONE* 7:
536   e29814.
537

538   Benetka V, Novotná K, Štochlová P. 2012. Wild populations as a source of germplasm for
539   black poplar (*Populus nigra* L.) breeding programmes. *Tree Genetics and Genomes* 8: 1073-
540   1084.
541

542   Cagelli L, Lefèvre F. 1997. The conservation of *Populus nigra* L. and gene flow within
543   cultivated poplars in Europe (updated). *Bocconea* 7:63-75.
544

545   Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S,
546   Hellens P, Kumar S, Castaro A *et al*. 2012. Genome-Wide SNP Detection, Validation, and
547   Development of an 8K SNP Array for Apple. *PLoS ONE* 7: e31745.
548

549   Chalhoub B, Belcram H, Caboche M. 2004. Efficient cloning of plant genomes into bacterial
550   artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant
551   Biotechnol J*. 2:181–188.
552

553   Chamaillard S, Fichot R, Vincent-Barbaroux C, Bastien C, Depierreux C, Dreyer E, Villar
554   M, Brignolas F. 2011. Variations in bulk leaf carbon isotope discrimination, growth and
555   related leaf traits among three *Populus nigra* L. populations. *Tree Physiology* 31: 1076-1087.
556

557   Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Z W et al., 2014. A high-density
558   SNP genotyping array for rice biology and molecular breeding. *Molecular Plant* 7:541-553.
559

560   Chenault Nicolas C, Arnaud-Haond SA, Juteau MJ, Valade R. Almeida JL, Villar M, Bastien
561   C, Dowkiw A. 2011. SSR-based analysis of clonality, spatial genetic structure and
562   introgression from the Lombardy poplar into a natural population of *Populus nigra* L. along
563   the Loire River. *Tree Genetics and Genomes* 7: 1249-1262.
564

565   Chu Y, Huang Q, Zhang B, Ding C, Su X. 2014. Expression and Molecular Evolution of
566   Two *DREB1* Genes in Black Poplar (*Populus nigra*). *PloS ONE* 9: e98334.
567

568   Chu Y, Su X, Huang Q, Zhang X. 2009. Patterns of DNA sequence variation at candidate
569   gene loci in black poplar (*Populus nigra* L.) as revealed by single nucleotide polymorphisms.
570   *Genetica* 137: 141-150.
571

572   Dalton-Morgan J, Hayward A, Alamery S, Tollenaere R, Mason AS, Campbell E, Patel D,
573   Lorenc MT, Yi B, Long Y *et al*. 2014. A high-throughput SNP array in the amphidiploid
574   species *Brassica napus* shows diversity in resistance genes. *Funct. Integr. Genomics* 14:
575   643-55.
576

577 Delourme R, Falentin C , Fomeju BF, Boillot M, Lassalle G, André I, Duarte J, Gauthier V,
578 Lucante N. 2013. High-density SNP-based genetic map development and linkage
579 disequilibrium assessment in *Brassica napus* L. *BMC Genomics* 14: 120.
580
581 DeWoody JD, Trewin HT, Taylor GT. 2015. Genetic and morphological differentiation in
582 *Populus nigra* L.: Isolation by colonization or isolation by adaptation? *Mol. Ecol.* doi:
583 10.1111/mec.13192.
584
585 Dickmann DI, Kuzovkina J. 2013. Poplars and willow of the world, with emphasis on
586 silviculturally important species (Chapter 2) In Poplars and Willows in the World: meeting
587 the needs of society and the environment. Eds. J.G. Isebrands and J. Richardson, 135 p,
588 FAO/IPC (Food and Agricultural Organization of the United States / International Poplar
589 Commission). Rome, Italy.
590
591
592 El-Maki R. 2013. Architecture génétique des caractères cibles pour la culture du peuplier en
593 taillis à courte rotation, pH D thesis, University of Orléans, 242p.
594
595 Evans L M, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM,
596 Schackwitz W, Gunter L, Chen JG *et al*. 2014. Population genomics of *Populus trichocarpa*
597 identifies signatures of selection and adaptive trait associations. *Nature Genetics* 46/ 1089–
598 1096.
599
600 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K,Buckler ES, Mitchell SE. 2011. A
601 Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.
602 *PLoS ONE* 6(5): e19379.
603
604 Fabbrini F, Gaudet M, Bastien C, Zaina G, Harfouche A, Beritognolo I, Marron N, Morgante
605 M, Scarascia-Mugnozza G, Sabatti M. 2012. Phenotypic plasticity, QTL mapping and
606 genomic characterization of bud set in black poplar. *BMC Plant Biol*. 12:47.
607
608 Fuentes-Utrilla P, Valbuena-Carabaña M, Ennos R, Gil L. 2014. Population clustering and
609 clonal structure evidence the relict state of *Ulmus minor* Mill. in the Balearic Islands glacial
610 history shape the genetic structure of Iberian poplars. *Mol. Ecol.* 21: 3593–3609.
611
612 Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann
613 M, Birol I *et al*. 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by
614 population transcriptome resequencing. *Mol. Ecol. Resour.* 11: 81-92.
615
616 Geraldes A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE,
617 Wymore AM, Grassa CJ, Farzaneh N *et al*. 2013. A 34K SNP genotyping array for *Populus*
618 *trichocarpa*: design, application to the study of natural populations and transferability to
619 other *Populus* species. *Mol. Ecol. Resour.* 13: 306–323.
620
621 Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, Bauland C, Camisan C, Campo L,
622 Meyer N, Ranc N *et al*. 2014. Linkage Disequilibrium with Linkage Analysis of Multiline
623 Crosses Reveals Different Multiallelic QTL for Hybrid Performance in the Flint and Dent
624 Heterotic Groups of Maize. *Genetics* 198: *1717-1734*
625

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, Warren WC, Hillier LW, Crooijmans RP, VereijkenA, Okimoto R, Muir WM, Cheng HH. 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(8) 850-863.

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Scarascia Mugnozza G. 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. *Theor. Popul. Biol.* 33: 54–78.

Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dlan P, Lorenz W, Dean JFD. 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. 2013. *Tree Genetics and Genomes* 9: 621-626.

Jost, L. 2008. GST and its relatives do not measure differentiation. *Mol. Ecol.* 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*. *Mol. Ecol.* 19: 1212–1226.

Kennard RW, Stone LA. 1969. Computer Aided Design of Experiments. *Technometrics* 11: 137-148.

674 Lefèvre F, Barsoum N, Heinze B, Kajba D, Rotach P, De Vries SMG, Turok J. 2001. *In situ*
675 conservation of *Populus nigra*. Ed, International Plant Genetic Resources Institute, Rome,
676 Italy. 58.
677
678 Lepoittevin C, Bodénés C, Chancerel, E, Villate L, Lang T, Lesur I, Boury C, Ehrenmann
679 F, Zelenica D, Boland A *et al*. 2015. Single-nucleotide polymorphism Discovery and
680 validation in high-density SNP array for genetic analysis in European White Oaks. Mol.
681 Ecol. Res. doi: 10.1111/1755-0998.12407
682
683 Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, Saski C, Schnell R *et*
684 *al*. 2015. Making a chocolate chip: development and evaluation of a 6K SNP array
685 for *Theobroma cacao*. *DNA Res* 22: 279-29
686
687 Li X, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, Monteros MJ, Brummer C.2014.
688 Development of an Alfalfa SNP Array and Its Use to Evaluate Patterns of Population
689 Structure and Linkage Disequilibrium. *PLoS ONE* 9: e84329.
690
691 Lin J, Gibbs JP, Smart LB. 2009. Population genetic structure of native versus naturalized
692 sympatric shrub willows (Salix; Salicaceae). *Am. J. Bot.* 96: 771-85.
693
694 Lin CY, Xing G, Xing C. 2012. Measuring linkage disequilibrium by the partial correlation
695 coefficient. *Heredity* 109: 401–402.
696
697 Ma T, Wang J, Zhou G, Yue Z, Hu Q, Chen Y, Liu B, Qiu Q, Wang Z, Zhang J *et al*. 2013.
698 Genomic insights into salt adaptation in a desert poplar. *Nature Communications* 4:
699 doi:10.1038/ncomms 3797.
700
701 Macaya-Sanz D, Heuertz M, Lopez de Heredia U, De Lucas AI, Hidalgo E, Maestro C, Prada
702 A, Alia R, González-Martínez SG. 2012. The Atlantic-Mediterranean watershed, river
703 basins and glacial history shape the genetic structure of Iberian poplars. *Mol.*
704 *Ecol.* 21: 3593–3609.
705
706 Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. 2012. Novel
707 measures of linkage disequilibrium that correct the bias due to population structure and
708 relatedness. *Heredity* 108: 285–291.
709
710 Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi EJ, de Souza
711 Gonçalves P, Vicentini R, Pereira de Souza A. 2014. *De Novo* Assembly and Transcriptome
712 Analysis of the Rubber Tree (*Hevea brasiliensis*) and SNP Markers Development for Rubber
713 Biosynthesis Pathways. *PLoS ONE* 9: e102665.
714
715 Marroni F, Pinosio S, Morgante M. 2012. The quest for rare variants: pooled multiplexed
716 next generation sequencing in plants. *Front. Plant Sci*. 3: 133.
717
718 McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M,
719 Muchero W, Tuskan GA, Ehlting J *et al*. 2014. Geographical and environmental gradients
720 shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New*
721 *Phytologist* 201: 1263–1276.
722

723 Monclus R, Leplé JC, Catherine Bastien C, Bert PF, Villar M, Marron N, Brignolas F, Jorge
724 V. 2012. Integrating genome annotation and QTL position to identify candidate genes for
725 productivity, architecture and water-use efficiency in *Populus* spp. *BMC Plant Biol*. 12: 173.
726

727 Neale DB, Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends*
728 *Plant Sci.* 9: 325-30.
729

730 Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications
731 *Nature Reviews Genetics* 12: 111-122.
732

733 Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C, Dervinis
734 C, Yu Q, Sykes R, Davis M, Martin TA *et al*. 2009. Quantitative genetic analysis of biomass
735 and wood chemistry of *Populus* under different nitrogen levels. *New Phytologist* 182: 878–
736 890.
737

738 Parchman T L, Geist KS, Grahnen JA, Benkman CW, Buerkle CA. 2010. Transcriptome
739 sequencing in an ecologically important tree species: assembly, annotation, and marker
740 discovery. *BMC Genomics* 11: 180.
741

742 Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley
743 C, Mockler TC *et al*. 2012. Development and Evaluation of a Genome-Wide 6K SNP Array
744 for Diploid Sweet Cherry and Tetraploid Sour Cherry. *PLoS ONE* 7: e48305.
745

746 Porth I, Klapšte J, Skyba O, Hannemann J, McKown AD, Guy R D, DiFazio, SP, Muchero
747 W, Ranjan P, Tuskan GA *et al*. 2013. Genome-wide association mapping for wood
748 characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms.
749 *New Phytologist* 200: 710–726.
750

751 Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large
752 genomes. In Proceedings of the13 Annual International conference on Intelligent Systems
753 for Molecular Biology (ISMB-05). Detroit, Michigan.
754

755 Rae AM, Pinel MPC, Bastien C, Sabatti M, Street NR, Tucker J, Dixon C, Marron N, Dillen
756 SY, Taylor G. 2008. QTL for yield in bioenergy Populus: identifying GxE interactions from
757 growth at three contrasting sites. *Tree Genetics and Genomes* 4: 97–112.
758

759 Ranjan P, Yin T, Zhang X, Kalluri UC, Yang X, Jawdy S, Tuskan GA. 2010. Bioinformatics-
760 Based Identification of Candidate Genes from QTLs Associated with Cell Wall Traits in
761 *Populus Bioenergy Research* 3: 172–182.
762

763 Rohde A, Storme V, Jorge V, Gaudet M, Vitacolonna N, Fabbrini F, Ruttink T, Zaina G,
764 Marron N, Dillen S *et al*. 2011. Bud set in poplar - genetic dissection of a complex trait in
765 natural and hybrid populations. *New Phytologist* 189: 106-121.
766

767 Santos-del-Blanco L, de Lucas AI, González-Martínez SG, Sierra-de-Grado R, Hidalgo E
768 2013. Extensive Clonal Assemblies in *Populus alba* and *Populus xcanescens* from the
769 Iberian Peninsula. *Tree Genetics and Genomes* 9: 499 –510.
770

771 Savolainen, O, Pyhäjärvi T, Knürr T. 2007. Gene flow and local adaptation in trees. Annu.
772 *Rev. Ecol. Evol. Syst.* 38, 595–619.

773

774 Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, Mocke KE. 2014.
775 Genotyping-by-Sequencing for *Populus* Population Genomics: An Assessment of Genome
776 Sampling Patterns and Filtering Approaches. *PLoS ONE* 9: e95292.

777

778 Sim SC, Van Deynze A, Stoffel K, Douches DS, Zarka D, Ganal MW, Chetelat R, Hutton
779 SF, Scott JW, Gardner RG, 2012. High-density SNP genotyping of tomato (*Solanum
780 lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE* 7: e45520.

781

782 Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E,
783 Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA *et al.* 2012. Genome
784 resequencing reveals multiscale geographic structure and extensive linkage disequilibrium
785 in the forest tree *Populus trichocarpa*. *New Phytologist* 196: 713-725.

786

787 Smulders MJM, Cottrell JE, Lefèvre F, van der Schoot J, Arens P, Vosman B, Tabbener HE,
788 Grassi F, Fossati T, Castiglione S *et al.* 2008. Structure of the genetic diversity in black
789 poplar (*Populus nigra* L) populations across European river systems: consequences for
790 conservation and restoration. *For. Ecol. Manage* 255: 1388–1399.

791

792 Stanton BJ, Serapiglia MJ, Smart LB. 2013. The domestication and conservation of *Populus*
793 and *Salix* genetic resources. In Poplars and willows: Trees for society and the environment,
794 Eds J.G. Isebrands and J. Richardson, chapter 4.

795

796 Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C. 2013.
797 Genomic scan for single nucleotide poplymorphisms reveals patterns of divergence and gene
798 flow between ecologically divergent species. *Mol. Ecol.* 22: 842-855

799

800 Storme V, Vanden Broeck A, Ivens B, Halfmaerten D, Van Slycken J, Castiglione S, Grassi
801 F, Fossati T, Cottrell JE, Tabbener HE *et al*. 2004. Ex-situ conservation of black poplar in
802 Europe: genetic diversity in nine gene bank collections and their value for nature
803 development. *Theor. Appl. Genet*. 108: 969–981.

804

805 Stueffer IF, Ershamber B, Huber H, Suzuki I. 2002. The ecology and evolutionary biology
806 of clonal plants: an introduction to the proceedings of Clone-2000. *Evolutionary Ecology*
807 15: 223-230.

808

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

812

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

816

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

820

821 Vanholme B, Cesarino I, Goeminne G, Kim H, Marroni F, Van Acker R, Vanholme R,
822 Morreel K, Ivens B, Pinosio S *et al*. 2013. Breeding with rare defective alleles (BRDA): a
823 natural *Populus nigra* HCT mutant with modified lignin as a case study. *New Phytologist*
824 198: 765–776.

825

826 VanRaden PM. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91:
827 4414-23.

828

829 Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara
830 UR, Cattonaro F, Vendramin E *et al*. 2012. Development and Evaluation of a 9K SNP
831 Array for Peach by Internationally Coordinated SNP Detection and Validation in Breeding
832 Germplasm. *PLoS ONE* 7: e35668.

833

834 Viger M, Rodrigues-Acosta M, Rae AM, Morison JIL, Taylor G. 2013. Towards improved
835 drought tolerance in bioenergy crops: QTL for carbon isotope composition and stomatal
836 conductance in *Populus*. *Food and Energy Security*, DOI: 10.1002/fes3.39.

837

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

840

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

843

844   **Data accessibility**

845   Collections of SNPs within the candidate regions and genes and outside are given in

846   supporting information. Primer of Sanger Sequencing project are listed in supporting

847   information

848   The *P. nigra* reference and the raw sequencing data will be available at

849   http://services.appliedgenomics.org/gbrowse/populus/ hosted by Applied Genomic Institute

850   in Udine (Italy).

851   The genotyping data will be available at https://urgi.versailles.inra.fr/Tools/GnpIS and

852   http://www.evoltree.eu/index.php/e-recources/portals.

853

854   **Author contributions**

855   PFR, GZ, VJ, SG, VS, VG, AB -Sanger re-sequencing and SNP identification

856   PFR, GZ, VJ, SG, VS, AB, MM -NGS re-sequencing and SNP identification

857   PFR, VJ, VS, GZ, MV, AP, GT -Design of the SNP array

858   MVil -Collecting of *P. nigra* samples

859   CB, GT -Design of the population sampling

860   PFR, MCL, FC, MM -Coordination of NGS re-sequencing work

861   PFR, MCL -Coordination of the genotyping work

862   CA, SS, EDP -Bioinformatics, data basing

863   PFR, PP, VG -Analysis of genotypic data

864   VJ, VS, CB -Population genetics analysis

865   PFR, GZ, VJ, VS, CB -Writing of the manuscript

866   MVil, GT, MRA -Revision of the manuscript

867

868

869 Table 1: Raw sequence data used for SNP detection. *Vert de Garonne and Cazebonne 25
870 were subsequently found identical genotypes after HT genotyping. (A) Adour.
871

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

872

873   Table 2: Summary of the number of *P. nigra* genotypes per river basin in the European
874   *P.nigra* association populations.
875

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

876
877

878   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	937,79	282,299	491,85	460,047
Whithin 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 5 accessions			278,330		
Supported by at least one >25x genotype clone			189,616		

879

880   **Figure legends**
881
882   Figure 1: Workflow of SNP detection and selection.
883
884   Figure 2: Genomic distribution of SNPs detected for the development of the 12k bead-chip
885   array. Around the plot colored bars represent the 19 *Populus* chromosomes (unit used is 2
886   Mb). Within the plot the traces represent the SNP distribution (calculated in windows of 100
887   kb) of BDG (red) BEN3 (light-blue) Poli (light-green) 71077-308 (yellow) 47 genotypes
888   (violet). The grey ovals tag the putative centromeric regions. The grey arrows tag the putative
889   centromeric regions. The red arrows highlight homozygous regions for the 71077-308 clone,
890   they represent homozygous genomic regions. Such homozygous areas have already been
891   observed in previous studies based on genetic mapping in *P. nigra* (El-Malki, 2013). The
892   plot was computed using the Circos software (Krzywinski *et al.* 2009).
893
894   Figure 3: Chromosomal distribution of SNP densities and summary of QTL locations for
895   wood composition, bud phenology, water-use efficiency and rust resistance in the poplar
896   genome. Numbers of SNP were calculated for all 500kb windows across all 19
897   chromosomes. Black vertical bars indicated low priority QTL intervals -1: bud phenology -
898   4: rust resistance -6: bud phenology, wood composition and wood density -8: bud phenology
899   and wood composition -10: bud phenology, wood composition and water-use efficiency -
900   11: rust resistance -12: rust resistance -13: bud phenology. Red vertical bars indicated high
901   priority QTL intervals -2: wood composition -3: rust resistance and bud phenology -5: wood
902   composition, wood density and bud phenology -7: wood composition and bud phenology -
903   12: bud phenology and water-use efficiency -13: wood composition -14: rust resistance.
904   Details on QTL position and references are given in table S3.
905
906   Figure 4: Population structure analysis estimated for 600 SNP distributed throughout the *P.
907   nigra* genome in validated genotypes – 4a: Estimation of the best value of K determined by
908   the cross validation error implemented in ADMIXTURE software. K was tested for different
909   sets of SNP detailed in the Material and Methods section.
910   – 4b: Admixture results from 706 individuals and 600 SNP K=6, K=7, K=8. Each color
911   represents a different ancestral cluster. Each individual was represented as a thin vertical bar
912   which was divided into color segments that were proportional to its memberships in the
913   ancestral clusters. At K=8, individuals collected along the Rhône river basin were divided
914   into 2 subpopulations, one is located on the upper part and the other one on the lower part of
915   the river. – 4c: Geographical distribution of the populations and the genetic structure
916   revealed by ADMIXTURE
917
918
919   Figure 5: Linkage disequilibrium *vs* physical distances. -5a: The decay of LD was
920   investigated by plotting all pairwise r^2 values against physical distance windows of 100kb. -
921   5b: r^2 values were corrected according the populations structure. -5c: The decay of LD was
922   investigated by plotting 600 pairwise r^2 values against physical distance windows of 100kb.
923

924   **List of supplemental data**
925
926   Methods S1: DNA extraction and Sanger sequencing of gene amplicons.
927
928   Methods S2: Calculation of Illumina sequencing accuracy.
929
930   Methods S3: Validation and Origin of replicates data with SSR genotyping.
931
932   Figure S1: Test of SNP segregation conformity within 8 progenies belonging to a 3x3
933   factorial mating design.
934   We genotyped the 6 parents and 290 progenies belonging to 8 families including in a 3 x 3
935   factorial mating design. The segregating markers were classified in 5 groups according to
936   the expected segregation pattern deduced from genotype of the parents: BC1 (AB x AA), F1
937   (AA x BB), F2 (AB x AB), Mono. (AA x AA) and Miss. (missing data in at least one parent).
938   Numbers in black are the total number of markers in each class. Conformity of the
939   segregation pattern with the parental genotype has been checked in each family (numbers in
940   red, numbers with * are number of marker for which a F2 segregating class is missing).
941   Approximately 98 % of the markers analyzed in the progeny fit the expected Mendelian
942   segregation ratios in each family. $\chi2$ tests for segregation distortion were performed pooling
943   half-sib families (lines and columns from the factorial mating design) at thresholds of
944   P = 0.01. Among the SNP, 216 showed segregation distortion.
945
946   Figure S2: Principal component analysis: The first second and third axes explain 2.39%,
947   1.89%, 1.71% of the total variance respectively. Each dot represents one individual.
948   Individuals used in the SNP discovery panel are indicated by black dots. The first axis
949   differentiates South France populations from the East France populations and Northern Italy
950   population. The second axis, revealed the separation of the Italian populations. The
951   distribution of the discovery panel along the axes reflects the variation of the populations
952   studied.
953
954   Figure S3: Distribution of Minor Allele Frequencies (MAF) for 7.896 SNPs in 7 clusters and
955   the association population (706 individuals). Clusters are constituted based on Admixture
956   analyses with 600 SNPs (see Fig. 4b).
957
958   Table S1: SNP-panel discovery and list of genotyped *P. nigra* individuals.
959   [1-9] progenies derived from controlled crosses between [1] SRZ and VGN [2] 71077-308 and
960   VGN [3] SRZ and BDG [4] 71041-302 and BDG [5] 71072-501 and BDG [6] 71072_501 and SRZ
961   [7] 71072-501 and SRZ [8] 71077-308 and L150-089 (*P. deltoides*) [9] 58-861 and Poli.
962
963   Table S2: Primer pairs developed within genes for Sanger re-sequencing and SNP
964   collections. -Collection 1: Light signaling pathway -Collection 2: Rust resistance, wood
965   properties, drought stress, randomly distributed along the genome
966
967   Table S3: List of candidate regions and candidate genes based on location of QTL hot spots
968   for rust resistance drought stress, bud phenology, wood composition and transcriptome
969   studies. Number in brackets were the QTL numbering in figure 3, QTL region and traits
970   written in italic were inherited *P. deltoides* or *P. trichocarpa* species.
971

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

975 Table S5: List of SNPs extracted from HT-sequencing data. The SNP are denoted by
976 SNP_IGA followed by the chromosome or scaffold number (V2.0) and the base position
977 within the scaffold.

979 Table S6: Origin and number of SNPs included in the 12 000 BeadChip array.

981 Table S7: List of SNP included in the 12 000 BeadChip array.

983 Table S8: Performance of the BeadChip array.

985 Table S9: Comparison of genotyping data and Sanger data.

987 Table S10: Comparison of genotyping data and NGS data.

989 Table S11: Genomic position and gene assignation of the 8259 useful SNP.

991 Table S12: List and origin of unexpected replicates.

993 Table S13: Chromosomal distribution of SNP numbers, SNP distances and SNP densities.
994 As expected from our selection strategy, the number of high quality SNPs per chromosome
995 was highly variable (from 72 on chromosome 9 to 1870 on chromosome 6) (Table 4).
996 Chromosome 6 had the highest density of SNPs (67 SNPs/Mb), and chromosome 18 the
997 lowest density (4.3 SNPs/Mb).The largest physical region with no SNP was found on
998 chromosome 17.

P. nigra reference sequence

71077-308 PE reads
Mapping of PE reads vs *P.trichocarpa* v2.0
CLC Genomics Workbench v.4

- Length fraction : 0.9
- Similarity : 0.9
- Min PE distance 250 b
- Max PE distance 800 b
- Unique matches retained

Masked for duplications and repetitions

- RepeatScout, default parameters

P. nigra variant calling

71077-308, BEN3, BDG, Poli, pool of 47 individuals PE reads
Mapping of PE reads vs *P. nigra* reference sequence
CLC Genomics Workbench v.4

- Similarity : 0.95
- Min coverage : 0.1 to 0.5 the average coverage
- Max coverage : 1.5 the average coverage
- Min variant frequency

SNP | 0,35 for 71077-308, BEN3, BDG, Poli
      | 0,15 for the pool of 47 individuals

DIP 0,1

- Second allele frequency
>0,1 for 71077-308, BEN3, BDG, Poli
>0,05 for the pool of 47 individuals

Extraction of SNPs for candidate regions and genes

60 b flanking sequences with no SNPs/DIPs
Remove duplicated / repetitive 121 b sequences

Final SNPs included in the chip

189 616 SNPs   +   *4 691 Sanger SNPs*

- ADT score	$\geq 0,85$	>0,6
- BLASTn identity	> 0,97	>0,9
- Second allele frequency	$\geq 0,2$	>0,05

9443 SNPs   **888 SNPs**

Figure 2

Page 41 of 43

Molecular Ecology Resources

Figure 3

Figure 4

Page 43 of 43

Molecular Ecology Resources

A

B

C

K=6

K=7

K=8

Ancestry

NL   Rhine D   Rhine F   Dranse   Durance   Drôme   Ebro   Loire   Allier   Ticino   Paglia   Basento

Figure 5

Molecular Ecology Resources

Page 44 of 43

a

LD Decay over 100 kb
(uncorrected R²)

b

LD Decay over 100 kb
(corrected R²)

c

LD Decay over 100 kb
(corrected R², using 600 SNPs)