High-density SNP assay development for genetic analysis in maritime pine (Pinus pinaster)

Christophe Plomion, Jérôme Bartholomé, Isabelle Lesur Kupin Lesur, Christophe C. Boury, Isabel Rodríguez-Quilón, Hélène Lagraulet, François Ehrenmann, Laurent Bouffier, Jean-Marc J.-M. Gion, D. Grivet, et al.

MOLECULAR ECOLOGY
RESOURCES

High-density SNP assay development for genetic analysis in maritime pine (Pinus pinaster)

Journal:	*Molecular Ecology Resources*
Manuscript ID:	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Plomion, Christophe; INRA, Bartholomé, Jérôme; INRA, Lesur, Isabelle; INRA, Boury, Christophe; INRA, Rodríguez-Quilón, Isabelle; INIA, Lagraulet, Hélène; INRA, Ehrenmann, François; INRA, Bouffier, Laurent; INRA, Gion, Jean-Marc; CIRAD, Grivet, Delphine; INIA, de Miguel, Marina; INRA, de Maria, Nuria; INIA, Cervera, María Teresa; Instituto Nacional de Investigción y Tecnología Agraria y Alimentaria (INIA). Centro de Investigación Forestal (CIFOR), Ecología y Genética Forestal Bagnoli, Francesca; CNR, Isik, Fikret; NCSU, Vendramin, Giovanni; Institute of Biosciences and Bioresources, CNR, González-Martínez, Santiago; CIFOR-INIA, Forest Ecology and Genetics
Keywords:	Adaptation, Bioinfomatics/Phyloinfomatics, Gymnosperms, Molecular Evolution, Natural Selection and Contemporary Evolution, Population Genetics - Empirical

Molecular Ecology Resources – Permanent Genetic Resources

High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*)

Plomion C[1,2†], Bartholomé J[1,2], Lesur I[1,3], Boury C[1,2], Rodríguez-Quilón I[4], Lagraulet H[1,2], Ehrenmann F[1,2], Bouffier L[1,2], Gion JM[1,5], Grivet D[4], de Miguel M[1,2], de María N[4], Cervera MT[4], Bagnoli F[6], Isik F[7], Vendramin GG[6], González-Martínez SC[4]

[1]INRA, UMR1202 BIOGECO, F-33610 Cestas, France

[2]Univ. Bordeaux, BIOGECO, UMR 1202, F-33400 Talence, France

[3]HelixVenture, F-33700 Mérignac, France

[4]INIA, Forest Research Centre, E-28040 Madrid, Spain

[5]CIRAD, UMR AGAP, F-33612 Cestas, France

[6]Institute of Biosciences and Bioresources, National Research Council, Sesto Fiorentino (FI), Italy

[7]North Carolina State University, Department of Forestry and Environmental Resources, Raleigh, NC, USA

Keywords: Maritime pine, SNP, linkage mapping, population genetics

[†]Corresponding author: Christophe PLOMION

INRA, UMR1202 BIOGECO, F-33610 Cestas, France

e-mail: plomion@pierroton.inra.fr

Running title: SNP discovery and validation in maritime pine

Abstract

Maritime pine provides essential ecosystem services in the south-western Mediterranean basin, where it covers around 4 million ha. Its scattered distribution over a range of environmental conditions makes it an ideal forest tree species for studies of local adaptation and adaptive responses to climatic change. Highly-multiplexed single-nucleotide polymorphism (SNP) genotyping arrays are increasingly used to study genetic variation in living organisms and for practical applications in plant and animal breeding and genetic resource conservation. We developed a 9k Illumina Infinium SNP array and genotyped maritime pine trees from i) a three-generation inbred (F2) pedigree, ii) the French breeding population, and iii) natural populations from Portugal and the French Atlantic coast. A large proportion of the exploitable SNPs (2,052 / 8,410, i.e. 24.4%) segregated in the mapping population and could be mapped, providing the densest ever gene-based linkage map for this species. Population genetics and structure analyses based on 3,981 SNP markers common to the Portuguese and French gene pools revealed high levels of differentiation, leading to the identification of a set of highly differentiated SNPs that could be used for seed provenance certification. Finally, we discuss how the validated SNPs could facilitate the identification of ecologically and economically relevant genetic factors in this species, improving our understanding of the demography and selective forces shaping its natural genetic diversity, and providing support for new breeding strategies.

Introduction

Maritime pine (*Pinus pinaster* Aiton, Pinaceae) is a long-lived wind-pollinated forest tree species native to the western part of the Mediterranean Basin. Its natural range extends from northern Morocco in the south to French Brittany in the north, and from Portugal in the west to Italy in the east (http://www.euforgen.org/distribution-maps/). It is found in various ecological situations, from sea level to an altitude of 2,100 m in the High Atlas (Morocco), from regions characterised by heavy annual rainfall in a maritime climate to dry regions in the semi-arid Mediterranean climate, and from calcareous to acidic soils (Alía and Martín, 2003). Its scattered distribution has resulted in local adaptations and high levels of genetic differentiation for adaptive traits across its natural distribution range (González-Martínez *et al*. 2002, Santos-del-Blanco *et al.* 2012, Lamy *et al.* 2011, 2014). Considerable genetic differentiation between ecotypes has been reported for various neutral molecular markers, providing clear evidence for a geographic structure of genetic diversity (Bucci *et al*. 2007, Santos-del-Blanco *et al*. 2012, Jaramillo-Correa *et al*. 2010, 2015).

This fast-growing tree has been widely planted in systematic reforestation programs since the 19th century, to secure coastal (along the Atlantic) and inland (Castilian plateau) sand dune areas, to drain marshes and to create new forests for resin production. Over the last 50 years, this species has been commercially exploited as a timber resource for the forestry industry (sawmills, wood panels, pulp and paper). In recent years, it has also been used as a source of chemicals for the bio-industry (Rohdewald 2002, Touriño *et al*. 2005, Jorge *et al*. 2002), as bioactive phenolic compounds can be extracted from its bark. There are now 4.2 million ha under maritime pine within its natural range and 200,000 ha outside this range (mostly in Australia) (Bouffier *et al*. 2013). The breeding of maritime pine began in the 1960s in

southwestern France, after several species and provenance trials had shown that the local ecotype was the best adapted and fastest growing tree in the Aquitaine soil and climatic conditions (Illy 1966, Harfouche 1995). This programme has now reached its third generation, and is one of the most advanced conifer breeding programmes in the world (Mullin *et al.* 2011).

The maritime pine has also been adopted by the forest tree genetics research community as a key model species for investigations of the genetic variation or linked mutations underlying phenotypic variability, particularly those selected by the environment and involved in local adaptation (reviewed by González-Martínez *et al.* 2011). It is now hoped that the discovery of polymorphisms causing changes in gene expression and/or amino-acid sequences will lead to innovations in genetic resource management, for both breeding (Isik 2014) and conservation strategies (Ouborg *et al.* 2010). Such discoveries should also lead to changes in silviculture practices to take into account the evolutionary processes inferred from neutral and selected markers (Lefèvre *et al*. 2014). Major efforts have been devoted to the sequencing and assembly of the maritime pine transcriptome (Canales *et al*. 2014), for studies of the molecular basis of the phenotypic response to biotic and abiotic constraints (e.g. Le Provost *et al*. 2013). Moreover, since the pioneering work of Lepoittevin *et al*. (2010), describing the design of the first multiplex single-nucleotide polymorphism (SNP) genotyping assay in maritime pine, medium-scale SNP-arrays have been developed (Supplemental Table 1). These assays have made it possible to characterise hundreds of trees, at hundreds of loci, for various applications: nucleotide diversity analysis (Plomion *et al*. 2014), QTL detection (de Miguel *et al*. 2014), association mapping (Lepoittevin *et al.* 2012, Budde *et al.* 2014), analyses of the correlations between genotype frequencies and environmental gradients (Jaramillo-Correa *et al.* 2015) and linkage map construction (Chancerel *et al.* 2013).

Given the high throughput and reliability of the Infinium platform from Illumina (e.g. Bartholomé *et al*. 2015a for eucalyptus, Pavy *et al.* 2013 for spruce), we used this platform to design a customised genotyping array for maritime pine, including about 5,000 SNPs from the studies cited above and about 4,000 SNPs newly identified from RNA-seq data and new amplicon resequencing. The resulting 9k SNP array is the largest genotyping chip ever produced for this species. We assessed the suitability of this array for genetic linkage mapping and genetic differentiation analysis between two gene pools (Portugal and French Atlantic coast). Finally, we discuss the potential utility of this SNP array for exploring genetic diversity and its contribution to phenotypic variation, genetic inferences about historical demographic events, the past action of natural selection and adaptive evolution, and the implementation of novel tree breeding strategies.

Materials and methods

Design of an Illumina Infinium array for maritime pine

We designed a 9k Illumina Infinium SNP array for maritime pine, including the two subsets of SNPs described below.

Previously available SNPs

Over the past five years, several studies have reported the development of SNP markers in maritime pine (see Supplemental Table 1). We selected a total of 4,997 SNPs from assays based on VeraCode, GoldenGate or Infinium Illumina technologies (Illumina, San Diego, CA, USA) and incorporated them into the same Infinium genotyping array. Some SNPs had been used on a number of occasions (270 SNPs were common to at least two studies), whereas

126 others were developed specifically for one study, as shown in Figure 1. SNP redundancy was
127 checked by aligning flanking sequences with the maritime pine UniGene (Canales *et al*. 2014:
128 http://www.scbi.uma.es/sustainpinedb/home_page). Alignments were performed by carrying
129 out BlastN searches in the Blastall V2.2.26 suite (e-value = 10^{-5}). We retained the best aligned
130 contig for each sequence and the position of the SNP was retrieved from that contig. When a
131 redundant SNP was associated with different dbSNP accessions, we retained the accession
132 with the longest flanking sequence. Overall, we retained 4,712 different SNPs in this first
133 subset.
134
135 *Newly developed SNPs*
136 This second subset comprised SNPs obtained by the random screening of EST data or
137 specifically detected in candidate gene sequences.
138
139 Newly developed SNPs from 454 sequence reads
140 A flowchart describing the steps involved in the identification of SNPs from 454 data is
141 shown in Figure 2. Three genotypes involved as progenitors of interprovenance hybrids in the
142 framework of the maritime pine breeding programme (accessions 110-4019-1 from Corsica
143 (C), 0284-2 from Landes (L) and 112-4-1 (M) from Morocco) were used. A composite cDNA
144 library was constructed with the SMART PCR cDNA synthesis kit (Clontech, Laboratories
145 Inc., Mountain View, CA, USA) for each tree. The C and M libraries contained equal
146 proportions of cDNAs from differentiating xylem, swelling buds and young needles, whereas
147 the L library consisted of equal proportions of cDNAs from differentiating xylem, swelling
148 and quiescent buds, and young and mature needles. Pyrosequencing (454 Titanium, Roche,
149 Branford, CT, USA) was performed with the Roche-454 Genome Sequencing platform (FLX
150 Titanium technology). Sequences (available under accession numbers SRX031589,

SRX208012 and SRX031592 from the NCBI short-read archive) were cleaned with the Pyrocleaner tool (Mariette *et al*. 2011), which removes particularly short (<150 bp) and long reads (>600 bp), reads with a percentage of Ns (ambiguous base calls) greater than 2%, low-complexity regions, and duplicated reads. For each library, the SNPs were identified by aligning each set of sequences against the 198,425 contigs of the second maritime pine UniGene established by Chancerel *et al*. (2013). Alignments were performed with the CLC Genomics Workbench Reference Mapping function of CLC Genomics Workbench V6.0 (CLC Bio, Aarhus, Denmark), with the default parameters. SNPs were then detected with the neighbourhood quality standard algorithm (NQS) and the following parameters: minimum coverage = 6, minimum central base quality = 20, minimum neighbourhood quality over a window length of 11 nucleotides = 15, maximum gap and mismatch count = 2 and minimum allele frequency = 20%.

Newly developed SNPs from candidate genes

Grivet *et al*. (2015) obtained 105 *in vitro* SNPs from two full-length candidate genes with functions relating to phenology (*col1* and *gia*) and 64 amplicons sequenced as part of the CRIEC (Comparative Re-sequencing in European Conifers) project, an EvolTree (http://www.evoltree.eu) initiative. An additional set of *in silico* SNPs was then obtained from 281 candidate genes with functions relating to abiotic and biotic (plant defence) stress responses in forest trees selected from the following sources: i) 66 genes from published (Perdiguero *et al.* 2013) and unpublished maritime pine sequences available from GenBank, ii) 149 genes from sequencing studies in other conifers (Wachowiak *et al*. 2009, Kujala and Savolainen 2012), iii) 53 transcripts displaying differential expression in the presence and absence of pine wood nematode infection (Santos *et al.* 2013), iv) 10 genes associated with adventitious shoot induction and plant development in pines (Alonso *et al*. 2007, Ordás *et al*.

unpublished for *knox* genes), and v) 3 genes potentially involved in cavitation resistance in beech (Lalagüe *et al.* 2014). These sequences were blasted (BlastN, e-value = 10^{-5}) against the maritime pine UniGene to retrieve the best matching contig and the sequences surrounding the SNP from a catalogue of 55,607 available SNPs (Figure 3). Redundant SNPs were removed.

Annotation of synonymous and non-synonymous substitutions

The respective positions of SNPs were defined in the contigs of the maritime pine UniGene from Canales *et al.* (2014). Within the coding sequences (when characterised), non-synonymous and synonymous SNPs were annotated by comparing the amino acids translated from the reference codon to the codon containing the SNP. The functional annotation was retrieved from the study by Canales *et al*. (2014). All these items of information are available in Supplemental Table 3.

The populations studied and genetic analysis

The SNP assays were tested and validated on the basis of Mendelian segregation in mapping pedigrees and genetic diversity analysis in an elite breeding population from France and natural populations of different origins (French Atlantic coast and Portugal).

Mapping population and linkage analysis

The mapping population consisted of a three-generation inbred pedigree (F2) obtained by the self-pollination of an inter-provenance “Landes x Corsica” hybrid (accession H12 resulting

from the control cross between L146 and C10 genotypes). In total, 638 F2 seeds were planted in a nursery in June 1998 and 626 saplings were transplanted into the field in March 1999 (4 m × 2 m, 0.51 ha, Lacanau de Mios, France). After 15 years, 565 F2 plants were still available for genetic analysis. We used 92 F2 plants to test the Mendelian segregation of the SNP markers and to associate them with a particular genetic position. The F2 plants with the most recombinant genotypes were selected with MapPop software (http://www.bio.unc.edu/faculty/vision/lab/mappop/, Vision et al. 2000) and a linkage map was established by Lavigne *et al*. (manuscript in preparation), by genotyping 477 F2 plants for 248 SNPs distributed over 12 linkage groups (LG), the haploid chromosome number in pines.

The *R* package *onemap* v2.0-3 (Margarido *et al.* 2007, Mollinari *et al*. 2009) was used for linkage mapping. Only SNPs and individuals for which there was less than 5% and 1% missing data, respectively, were considered in the analysis. SNPs were clustered into LGs on the basis of a LOD score greater than 10. The LG names were defined on the basis of previously mapped loci (Chancerel *et al.* 2013). The RECORD algorithm (Os *et al*. 2005) was used to order markers within LGs, with the following parameters: LOD=3 and max.rf=0.4. Recombination rates were converted into genetic distances (cM) with the Kosambi mapping function (Kosambi 1943). The goodness-of-fit of SNP segregations to the expected Mendelian segregation ratio (i.e. 1:2:1 for an F2 population) was assessed in chi-squared tests, with adjustment of the significance threshold for simultaneous multiple tests (Benjamini and Yekutieli 2001) within each LG. The same procedure was also applied to a previous SNP dataset genotyped in the same F2 family, but with different genotypes and mapping software (Chancerel *et al*. 2013). The two genetic maps were then combined into a composite linkage map with the *R* package *LPmerge* (Endelman and Plomion, 2014).

226

227 *Populations of unrelated individuals and genetic diversity estimation*

228 The French Atlantic coast gene pool was represented by two subsets of individuals: i) 50 trees
229 from two natural populations, Hourtin and Petrocq, sampled from a clonal collection
230 (CLONAPIN) established directly from the source populations (as described by Rodríguez-
231 Quilón *et al.* 2015). Hourtin (45°11'N, 1°09'E) and Petrocq (44°04'N, 1°18'E) are coastal
232 populations growing at low altitude (<30 m a.s.l.) and under typical maritime climate (annual
233 precipitation of 980-1248 mm and mean annual temperature of 12-13°C); and ii) 50 elite trees
234 from a larger set of about 600 genotypes mass-selected in natural forests of the Aquitaine
235 region in the early 1960s to constitute the first generation of the maritime pine breeding
236 program (Illy 1966). These 50 genotypes constitute the founders of a three-generation
237 pedigreed population used to develop proof-of-concept for genomic selection in maritime
238 pine.

239

240 The Portuguese population was also represented by two subsets of genotypes: i) 19 trees
241 sampled from two provenances in a provenance trial carried out at Mimizan (France:
242 44°20'N, 1°28'W). These provenances were described by Illy (1966): 6 trees were from seeds
243 collected at "Pinhal de Leiria" (an 11 ha coastal forest located at 39°79'N, 8°98'W; 0-50 m
244 a.s.l.; annual precipitation of 700-900 mm) and 13 trees were from seeds collected at "Trás-
245 os-Montes" (a mountain forest located at 41°57'N, 7°50'W; 1150 m a.s.l., annual
246 precipitation of 1200-1400 mm); and ii) 23 additional trees from "Pinhal de Leiria" collected
247 from the CLONAPIN maritime pine collection.

248

249 We first tested departure from Hardy-Weinberg equilibrium, in standard chi-squared tests
250 with a nominal significance threshold of 1.05-1.15 × 10^{-5}, corresponding to an experiment-

wise type I error of 5 % (Bonferroni correction to account for multiple testing). We then estimated three genetic diversity parameters for each SNP: minor allele frequency (MAF), observed (Ho) and expected heterozygosity (He, Nei's index of genetic diversity corrected for sample size, Nei 1987). These three parameters were highly correlated (>0.95). We therefore present data for MAF and *He* only. Genetic differentiation (F_{st}) was assessed for pairs of populations, with a set of 3,981 polymorphic SNPs common to both populations, with GDA software (Lewis and Zaykin 2001). Finally, Bayesian clustering analysis (Structure software, Pritchard *et al*. 2000, Falush *et al*. 2003) was performed to identify different gene pools in maritime pine, as described by Jaramillo-Correa *et al*. (2015).

Folded site frequency spectrum (SFS) for the French and Portuguese groups

Sites with missing data were excluded from the samples, to ensure general consistency across loci in terms of the number of sequences analysed for each site. Furthermore, we retained the same SNPs for the French and Portuguese groups, 3,513 SNPs in total, corresponding to 88% of the original dataset (3,981 SNPs). Finally, two individuals were discarded from the French group as they presented 19% (INIA_PET11) and 10% (INIA_PET12) missing data across the 3,981 SNPs. The estimate of folded SFS was therefore based on 3,513 common SNPs with no missing data, genotyped in 94 French and 42 Portuguese individuals.

SNPs were analysed as unpolarized (i.e. no ancestral state inferred), with the least frequent nucleotide at a given SNP site considered to correspond to the minor allele and thus used for the calculation of minor allele frequency (MAF). MAF was calculated for the French and Portuguese groups separately. The folded SFS was then plotted from the observed MAF and from the expected SFS. We used equation 6 from the paper by Ganapathy and Uyenoyama

(2009) to determine the expected SFS. This equation is widely used to describe the expected SFS for a genomic sample of SNPs, each of which is assumed to correspond to a mutation on an independent gene genealogy.

DNA extraction and genotyping

Developing needles were collected after bud burst for each genotype of the F2, natural and breeding populations. They were stored at -80°C or dried with silica gel before DNA extraction. About 30 - 40 mg of frozen/dried needles were crushed with a mixer mill (Retsch MM300, Haan, Germany). Genomic DNA was isolated with the Invisorb DNA plants 96 kit from Invitek (GmbH, Berlin, Germany), according to the manufacturer's instructions. All concentrations were determined with a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and a fluorescence assay (Quant-IT kit, Invitrogen, Carlsbad, CA, USA). Samples with DNA concentrations greater than 50 ng/µl (based on fluorescence measurements) were used for the Infinium assay. Genotyping was carried out at the Genes Diffusion Facility (Douai, France).

Results

SNP detection

The SNP genotyping array included 8,949 SNPs, 4,712 of which (i.e 52.65%) were already available (166 from Lepoittevin *et al*. 2010, 835 SNPs from Chancerel *et al*. 2011, 3,378 SNPs from Chancerel *et al.* 2013, 434 SNPs from de Miguel *et al.* 2014, and 184 SNPs from

Lavigne *et al.* (manuscript in preparation) (Figure 1). The other 4,237 (i.e 47.35%) corresponded to newly developed SNPs.

Newly developed SNPs

In silico SNPs from 454 sequence reads (Figure 2)

Pyrosequencing of the C (Corsica), M (Morocco) and L (Landes) libraries provided 188,494, 611,673 and 458,882 reads, respectively. After cleaning, we retained 107,089, 484,348 and 225,842 reads for C, L and M, respectively. We identified 2,039 SNPs in the C dataset, 4,937 in the L dataset and 4,752 in the M dataset. We then selected SNPs in a test-cross configuration (1:1 segregation), i.e. heterozygous in one parent and homozygous in the other, because, in a full-sib family with biallelic codominant markers, pairs of markers presenting this configuration provide the best estimate of recombination rates, particularly if compared with pairs having a test-cross and intercross (1:2:1) configuration (Plomion *et al*. 1997). Thus, we selected 2,830, 1,695 and 2,957 SNPs for the C × M, L × C and L × M crosses, respectively. In total 784, 2,046 and 911 SNPs were informative for the C, M and L parents, respectively. This procedure resulted in the retention of 3,741 unique SNPs as particularly suitable for linkage mapping.

In vitro and *in silico* SNPs from candidate genes (Figure 3)

Of the 281 candidate genes, 166 were associated with at least one contig of the maritime pine UniGene. We retrieved 888 SNPs from these 166 contigs. Finally, after removing duplicated sequences and SNPs located less than 100 bp apart, we retained 391 SNPs for SNP array development. In addition, 105 *in vitro* SNPs from 66 resequenced candidate genes were identified and incorporated into the SNP array.

Annotation of synonymous and non-synonymous substitutions

In total, 4,257 (47.6%) SNPs were found in contigs for which a full-length protein was predicted, with 1,557 (36.6%) of these SNPs located in non-coding regions. SNPs in coding sequences (2,700 *i.e.* 63.4%) were characterised in terms of their synonymous (S) or non-synonymous (NS) nature and codon responsible for the NS change (Figure 4A and Supplemental Table 3). We identified 567 (21%) synonymous SNPs, and found that the point mutation affected the third position in the codon in 87.5% of cases. There were 2,133 (79%) non-synonymous SNPs with the mutation affecting the first codon position in 21.5% and the second codon position in 70.6%. For NS mutations, the change in the protein is indicated in Supplemental Table 3. Most of the SNPs were transitions (59.3%). The most common transition was C↔T. The most frequent transversion was G↔T (Figure 4B).

Linkage mapping

Of the 8,410 SNPs used to genotype the 92 trees of the F2 mapping population, 4,634 (55.1%) were monomorphic, 2,052 (from 1,672 different contigs) were polymorphic (24.4%) and the remaining 1,724 (20.5%) corresponding to failed assays. All polymorphic SNPs passed the quality threshold and were therefore used in linkage analysis. The total map length was 1,993 cM, spread over 12 LGs corresponding to the haploid number of chromosomes for the maritime pine genome. LG length ranged from 129.2 (LG11) to 198.2 cM (LG5) and the mean number of SNPs per LG was 171 cM (F2_N in Table 1). The mean SNP density was 0.98, but more than half of the markers were grouped into clusters (i.e. groups of markers displaying no recombination). This resulted in 901 unique positions, separated by a mean

distance of 2.25 cM. Two factors could account for the clustering of SNPs: i) the small size of the mapping population (92 genotypes) and ii) the presence of more than one SNP in 260 of the 1,672 mapped contigs. Distorted SNPs were retained in the linkage analysis and accounted for 3.6% of the mapped markers. They were grouped together in five regions located on three LGs (segregation distortion regions, SDRs): two linked SDRs on LG#3 with 10 and 5 markers, one SDR on LG#6 (26 markers) and three SDRs on LG#10 (14, 12 and 2 markers) (Figure 5). The order of SNPs was similar (Spearman ρ=0.98) to that obtained from the linkage map reconstructed with the dataset of Chancerel *et al*. (2013) based on different genotypes of the same progeny (F2_O in Table 1). Only 56 of the 1,180 markers common to the two genetic maps differed in order between the two maps (Figure 5). These differences in marker order occurred between tightly linked markers (<5 cM), mostly on LG#1 and LG#10, for which 10 and 15 markers were involved, respectively. Finally, based on the highly conserved marker order between the two maps, a composite map was established (F2_C in Table 1) with LPmerge software. It contained 2,353 SNPs (including 1,121 SNPs in the same contigs). This composite map included 1,661 different loci, and is therefore the densest yet gene-based linkage map for this species. The total map length was 1,711.7 cM. There were 955 unique positions, separated by a mean distance of 1.82 cM.

Genetic diversity and population structure analysis

Only a few markers displayed significant departure from Hardy-Weinberg equilibrium in each population: 6 (0.10%) for the French Atlantic-based breeding population, 16 (0.26%) for the natural French Atlantic population and 18 (0.35%) for the Portuguese population. These markers, together with those for which >10% of the data were missing in any of the three populations, were removed from further analyses, resulting in a set of 5,016 loci (including

24% monomorphic loci) successfully scored in all populations. We found no differences in genetic diversity, assessed with this common SNP set, between the natural and breeding populations in the French Atlantic region (*He* of 0.336 vs. 0.332, based on polymorphic loci only; Table 2), suggesting that the mass selection in natural forests for constitution of the base breeding population for this species was broad enough to collect most of the standing genetic variation of this ecotype. This is further supported by the absence of significant genetic differentiation between the natural and mass-selected populations (F_{st} = 0.0005, 95% CI: -0.00002, 0.00107). Genetic diversity was slightly lower for the Portuguese population (*He* of 0.319) than for both the breeding and natural populations in France. High levels of genetic differentiation were observed between the French Atlantic and Portuguese populations (F_{st} = 0.0847, 95% CI: 0.08087, 0.08864), with 263 SNPs having F_{st} values above 0.25. These SNP markers are good candidates to replace the biochemical assay (based on terpene content analysis, Baradat and Marpeau-Bezard, 1988) currently used to determine the putative origin of adult forest stands in Aquitaine before the collection of seeds and their distribution for commercial purposes in France. Indeed, seed-lots from Portugal were introduced into Aquitaine in the 1950s and the stands they formed suffered frost damage after the exceptionally cold winter of 1985 (Ribeiro *et al.* 2002). Therefore, from 1986 onwards, candidate stands for seed collection in Aquitaine had to be certified as of French origin. The Bayesian clustering pattern observed, with the identification of only two gene pools (K=2), with the French natural and breeding populations grouped together into a single gene pool significantly separated from the Portuguese population (Figure 6), is an strong asset in this respect. This result will guide the development of highly informative genotyping multiplexes based on the Sequenom MassARRAY technology that has already been developed for this species (Chancerel *et al.* 2013).

As expected, the SFS of both the French and Portuguese groups showed a deficit of the observed low-frequency variants (Supplemental Figure 1). This deficit is probably due to a combined effect of the approach developed to select most of the SNPs (i.e. *in silico* analysis of ESTs obtained from small numbers of genotypes, precluding the capture of rare variants of the SFS), and the sampling scheme (a small sample size for each group), hindering the detection of rare variants even if present on the SNP array.

Discussion

In this section, we discuss how the substantial number of polymorphisms detected and validated in this study will help researchers working on maritime pine to address scientific questions in population genomics and quantitative genetics with large samples of individuals from natural and pedigreed populations that can now be surveyed for genetic variation at a few thousand expressed genes across the genome.

Linkage and QTL mapping

High-throughput SNP genotyping has been successfully implemented in several conifers (Pavy *et al*. 2008; Eckert *et al*. 2009a; Chancerel *et al.* 2011, 2013). This approach has generated thousands of markers for genetic mapping, significantly improving our understanding of large, complex conifer genomes (Jermstad *et al.* 2011, Pavy *et al*. 2012, Martínez-García *et al*. 2013, Neves *et al.* 2014). We show here that a combination of highly multiplexed SNP genotyping and a selective mapping strategy is useful for the mapping of hundreds of genes at low cost. The genetic map obtained with this strategy was highly reliable and accurate, as revealed by comparison with a previously developed map based on different

425 genotypes from the same progeny (Chancerel *et al.* 2013). The use of markers common to the
426 two sets of offspring made it possible to construct a composite genetic linkage map
427 confirming the positions of most of the previously mapped markers. Highly saturated gene-
428 based maps are very useful for the identification of orthologous genes, facilitating the study of
429 genome structure and evolution in conifers (Mackay *et al*. 2012, Ritland *et al*. 2011). Conifers
430 display a high degree of genome macrostructure conservation (Chagné *et al.* 2003, Pavy *et al.*
431 2012, Krutovsky *et al*. 2004, Jermstad *et al.* 2011). Comparative mapping is therefore a useful
432 tool to transfer information between species. SNP-based genetic maps have been used to
433 anchor large genome scaffolds to support conifer sequencing projects (Martínez-García *et al.*
434 2013, Neale *et al.* 2014). In addition, high-density gene-based maps are essential for the
435 identification of genes controlling targeted traits through their colocalization with QTLs
436 (Monclus *et al*. 2012, de Miguel *et al.* 2014, Lind *et al*. 2014, Li *et al.* 2014, Bartholomé *et al*.
437 2015b). Thus, the development and application of high-throughput genotyping technologies to
438 non-model species without a reference genome sequence, such as maritime pine, represents a
439 major advance towards the identification of genes controlling traits of interests.
440
441 *Association mapping*
442
443 In long-lived outbreeding organisms, such as forest trees, association mapping is an attractive
444 alternative to QTL mapping, as it can be used to identify molecular variation underlying
445 phenotypes from multiple genetic backgrounds without the need to produce segregating
446 progenies (Neale and Savolainen 2004). Under certain circumstances (e.g. moderate-high
447 heritability), association mapping can also be applied to natural populations. This approach is
448 useful for studies of adaptive traits in the precise environments in which they evolved (see
449 Parchman *et al*. 2012; Budde *et al*. 2014 for forest trees). Recent years have seen an explosion

in the number of association mapping studies in model organisms, such as humans and *Arabidopsis* (see reviews in Stranger *et al.* 2011; Weigel 2012). However, only a few studies, mostly focusing on technological characters associated with wood quality, have been carried out on forest trees (e.g. González-Martínez *et al*. 2011 and references therein; Cappa *et al*. 2013; Guerra *et al.* 2013). Progress has been particularly slow in maritime pine, partly due to the lack of a reliable and cost-efficient genotyping platform capable of dealing with the large sample sizes required to detect small- and medium-sized allelic effects. However, promising results were obtained in the first genetic association studies in maritime pine, based on only a few hundred markers. Lepoittevin *et al*. (2012) found two polymorphisms significantly associated with growth (in a HD-Zip III transcription factor) and wood cellulose content (in a fasciclin-like arabinogalactan protein), respectively. Cabezas *et al.* (2015) discovered a strong association between several (linked) polymorphisms in *korrigan,* a gene encoding a membrane-bound endo-1,4-β-d-glucanase involved in cellulose biosynthesis, and various growth traits, including height and polycyclism. Our SNP array includes all SNPs reported to display significant genotype:phenotype associations in previous studies. It is also enriched in SNPs from candidate genes, based on all the available information (published or unpublished) concerning genes displaying signatures of natural selection (e.g. Eveno *et al*. 2008; Grivet *et al.* 2011), involved in environmental associations (e.g. Jaramillo-Correa *et al*. 2015) or displaying differential expression (e.g. for pine nematode resistance, Santos *et al*. 2012; or drought response, Perdiguero *et al*. 2013) in this species. We therefore expected the use of this SNP genotyping array to generate highly informative data, increasing our understanding of the ecological drivers of adaptation and of use for breeding applications, including genomic selection (see below). For example, Westbrook *et al*. (2013) found that a small number of significantly associated SNPs (~20-30) had the same predictive power as the full dataset (4,854 SNPs) in SNP-based models for oleoresin flow in *Pinus taeda* L. Finally, this SNP

array will allow unprecedented explorations of the molecular basis of polygenic quantitative traits, through the implementation of multilocus association models (e.g. piMASS, Guan and Stephen 2011), for which high-density genotyping is required. Most of the traits relevant for adaptation or breeding are polygenic. The development of high-density SNP genotyping assays for forest trees thus constitutes a first step towards the use of molecular information in operational forestry.

Genomic selection

High-throughput genotyping platforms, such as the new SNP array developed here, can guide breeding and selection decisions (Eggen 2011). Genomic selection (GS) is a paradigm shift first introduced into animal breeding in 2008 for the selection of superior individuals in many countries (Goddard 2009). The major difference between GS and marker-assisted selection (MAS) is the number of markers used. GS makes use of a much larger number of markers to trace all the QTLs with small or large effect (Hayes and Goddard 2010). SNP-based genotyping platforms are reliable and repeatable for the genotyping of large numbers of individuals. There are few missing genotypes, and these genotypes can be handled (dropped or imputed) without markedly decreasing data quality. Thus, SNP arrays have become the choice genotyping platform for animal breeding and human genetics studies, despite advances in the efficiency of DNA sequencing technologies. Whole-genome sequences have been obtained for several major forest trees subjected to breeding programmes in recent years, and draft assemblies of whole genomes are now available even for conifers, which have particularly large genomes (Birol *et al.* 2013, Nystedt *et al.* 2013, Neale *et al*. 2014). These resources should facilitate the development of efficient genotyping by sequencing platforms in the near future. GS is expected to revolutionise forest tree breeding, by decreasing the need

for expensive and time-consuming progeny-testing practices. If successful, GS could halve the long breeding cycles (>15 years) of forest trees and double the genetic gain per unit time (Isik 2014). One probable application of GS application in forest trees would involve the use of data from different sources, such as progeny tests and genotyping centres, in a single-step approach to predicting the genetic merit of individuals (Legarra *et al.* 2009). The predicted model for one cycle can then be refined as new data become available. As breeding cycles progress and genotyping/sequencing costs fall, progeny testing will have a lesser effect on selection decisions. Nevertheless, the lack of a reliable and cost-efficient genotyping platform remains the major bottleneck for the routine application of GS in forest trees. The new high-density SNP array presented here constitutes a first step towards the implementation of GS in maritime pine.

Population genomics

The high-density SNP array developed in this study provides a powerful tool for the genome-wide genotyping of a large number of populations across the full distribution range of the species. The genotyping of hundreds of individuals across the entire range of maritime pine, with SNPs located both in coding and noncoding regions, would increase our understanding of the role of the evolutionary, demographic and adaptive mechanisms acting on natural populations. The available molecular markers have shown that maritime pine populations are spatially structured into regional gene pools connected by gene flow, particularly in the Iberian Peninsula (Burban and Petit 2003, Bucci *et al*. 2007, Jaramillo-Correa *et al.* 2015). However, the timing of the historical events leading to this spatial separation, and the degree of connectivity between gene pools, remain to be determined. In addition, maritime pine grows in diverse environmental conditions, resulting in the local adaptation of populations

over the range of this species (González-Martínez *et al*. 2002, Jaramillo-Correa *et al*. 2015, Serra-Varela *et al*. 2015), but the molecular mechanisms underlying this adaptation are poorly understood. Most ecologically and economically important traits are quantitative in nature and controlled by many loci (Rockman 2012). Extensive genome coverage is therefore required to disentangle selection from other factors that have shaped genomic variation across populations (e.g. genetic drift). This SNP array provides a means of identifying functional variation and the molecular bases of adaptation, through various methods based on the differentiation of allele frequencies between populations (F_{ST}-based methods; see e.g. Prunier *et al*. 2011, Chen *et al*. 2012), site-frequency spectrum statistics (e.g., Eckert *et al.* 2009b), correlations with environmental variables (e.g., Eckert *et al*. 2010, Jaramillo-Correa *et al.* 2015), or combinations of these approaches. 'Reverse' ecology approaches connecting genomic data with environmental parameters have also proved useful for identifying the major ecological drivers of adaptation (Levy and Borenstein 2012). This SNP array is therefore highly promising for studies aiming to interpret the signatures left in the genome by different evolutionary forces, making inferences about complex population processes, or investigating geographic patterns of genome-wide genetic variation.

Conclusion

In this study we report the discovery of new SNP markers in maritime pine, from RNA-seq and amplicon resequencing data, and the establishment of an Infinium genotyping array including SNPs that have already been validated. The SNP array presented here represents a major step forward for population and conservation genetics and for breeding in maritime pine. However, this array is subject to a number of limitations. Ascertainment biases, resulting from the small panel of individuals from which SNPs were obtained (ascertainment width)

and from the stringent criteria in terms of minimum allele frequency and read coverage used to retain the SNPs (ascertainment depth), must be taken into account when interpreting the results of future studies of natural populations. A systematic bias would be expected for estimates of nucleotide diversity, inferences about population structure, evolutionary processes based on natural selection and/or historical demographic models, particularly if based on the site frequency spectrum (reviewed by Helyar *et al*. 2011). Furthermore, most of the SNPs were obtained from independent contigs of the UniGene, precluding the use of linkage disequilibrium and haplotype diversity as metrics for population genetic inferences. In the near future, we will develop genotyping-by-sequencing approaches, which should make it possible to decrease the problem of ascertainment bias, because the individuals of interest are sequenced directly. Considering allele drop out of genotyping-by-sequencing approaches which reduces genome complexity through restriction digest (e.g. restriction-site-associated DNA sequencing, Baird et al. 2008) as a major drawbacks for genetic inferences in highly heterozygous species such as maritime pine (Davey et al. 2013, Gautier et al. 2013, Puritz et al. 2014, Mastretta-Yanes et al. 2015), future development of marker technology in this species will be based on sequence capture and direct sequencing (Gnirke et al. 2009).

Acknowledgements

This study was carried out with financial support from the European Union Sixth and Seventh Framework Programmes: Evoltree (no. 016322), Procogen (no. 289841), Noveltree (no. 211868), Forger (no. KBBE-289119), TipTree (BiodivERsA-ERANET); the French National Research Agency: ANR-FLAG (ANR-12ADAP-007-01); the Italian Ministry of Education, University and Scientific Research: Biodiversitalia (RBAP10A2T4); the Spanish Ministry of Science and Innovation: AdapCon (CGL2011-30182-C02-01). We thank C García-Barriga for

DNA extraction. D Grivet was supported by a “Ramón y Cajal” Fellowship from the Spanish Ministry of Science and Innovation, I Lesur by ANR-FLAG and C Boury by Procogen. H Lagraulet received a PhD fellowship from INRA (EFPA division and ACCAF metaprogramme) and Région Aquitaine (no. 20111203004). J Bartholomé was supported by a postdoctoral fellowship from "Conseil Général des Landes". I Rodríguez-Quilón acknowledges a PhD scholarship (FPI-INIA) from the Spanish National Institute for Agricultural and Food Research and Technology and the Spanish Ministry of Science and Innovation.

Data accessibility

Roche 454 raw data have been deposited in the short-read archive of NCBI (accessions: SRX031589 for genotype 110-4019-1, SRX208012 for genotype 0284-2 and SRX031592 for genotype 112-4-1). The assembled contigs of the second maritime pine UniGene are available from:

http://genotoul-contigbrowser.toulouse.inra.fr:9092/Pinus_pinaster2/index.html.

Authors' contributions

IL and HL: *in silico* SNP detection in mapping populations; IL: SNP array design and SNP annotation; CB, LB, JMG, FB, GGV, DG, IRQ, and SCGM: logistical aspects of genotyping; CB, JB IRQ, and DG: SNP scoring; CP, IL, MdM, MTC, NdM and FH: database compilation from published and unpublished data; JB: linkage mapping; SCGM and DG: population genetics analysis; CP: design of the study and overall coordination. All the authors participated in the writing of the paper, and have read and approved the manuscript submitted.

600

601 **References**

602

603 Alía R, Martin S (2003) EUFORGEN Technical Guidelines for Genetic Conservation and Use
604 for Maritime Pine (*Pinus pinaster*). IPGRI: Rome. Italy, 6p.

605

606 Alonso P, Cortizo M, Cantón FR, *et al.* (2007). Identification of genes differentially expressed
607 during adventitious shoot induction in *Pinus pinea* cotyledons by subtractive hybridization
608 and quantitative PCR. *Tree Physiology,* **27**, 1721-1730.

609

610 Baird NA, Etter PD, Atwood TS, *et al. (*2008*)* Rapid SNP discovery and genetic mapping
611 using sequenced RAD markers. *PLoS ONE,* **3**, e3376.

612

613 Baradat P, Marpeau-Bezard A (1988) Le pin maritime, *Pinus pinaster* Ait., biologie et
614 génétique des terpènes pour la connaissance et l'amélioration de l'espèce, Ph.D. Thesis,
615 University of Bordeaux I.

616

617 Bartholomé J, Mandrou E, Mabiala A, *et al.* (2015a) High-resolution genetic linkage maps of
618 *Eucalyptus* improve BRASUZ1 reference genome assembly. *New Phytologist,* doi:
619 10.1111/nph.13150.

620

621 Bartholomé J, Mabiala A, Savelli B, Bert D, *et al.* (2015b) Genetic architecture of carbon
622 isotope composition and growth in *Eucalyptus* across multiple environments. *New*
623 *Phytologist,* doi: 10.1111/nph.13301.

624

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics,* **29**, 1165-1188.

Birol I, Raymond A, Jackman SD, *et al*. (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**,1492-1497.

Bouffier L, Annie Raffin A, Alía R (2013) Maritime pine – *Pinus pinaster* Ait. In: *Best Practice for Tree Breeding in Europe (eds Mullin TJ & Lee SJ).* pp. 65-75. Skogforsk, Sweden, Uppsala.

Bucci G, González-Martínez SC, Le Provost G, *et al.* (2007) Range-wide phylogeography and gene zones in *Pinus pinaster* Ait. revealed by chloroplast microsatellite markers. *Molecular Ecology,* **16**, 2137–2153.

Budde KB, Heuertz M, Hernández-Serrano A, *et al.* (2014) *In situ* genetic association for serotiny, a fire-related trait, in Mediterranean maritime pine (*Pinus pinaster* Aiton). *New Phytologist*, **201**, 230-241.

Burban C, Petit RJ (2003) Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. *Molecular Ecology,* **12**, 1487–1495.

Cabezas JA, González-Martínez SC, Collada C, *et al.* (2015) Nucleotide polymorphisms in a pine ortholog of the *Arabidopsis* degrading enzyme cellulase KORRIGAN are associated with early growth performance in *Pinus pinaster.* (in press)

650 Canales J, Bautista R, Label P, *et al.* (2014) *De novo* assembly of maritime pine
651 transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnology*
652 *Journal,* **12**, 286–299.
653
654 Cappa EP, El-Kassaby YA, Garcia MN, *et al.* (2013) Impacts of population structure and
655 analytical models in genome-wide association studies of complex traits in forest trees: A case
656 study in *Eucalyptus globulus*. *PLoS ONE,* **8**, e81267.
657
658 Chagné D, Brown G, Lalanne C, *et al.* (2003) Comparative genome and QTL mapping
659 between maritime and loblolly pines. *Molecular Breeding,* **12**, 185-195.
660
661 Chancerel E, Lamy JB, Lesur I, *et al.* (2013) High density linkage mapping in a pine tree
662 reveals a genomic region associated with inbreeding depression and provides clues to the
663 extent and distribution of meiotic recombination. *BMC Biology,* **11**:50.
664
665 Chancerel E, Lepoittevin C, Le Provost G, *et al.* C (2011) Development and implementation
666 of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative
667 mapping with loblolly pine. *BMC Genomics,* **12**: 368.
668
669 Chen J, Källman T, Ma X, *et al.* (2012) Disentangling the roles of history and local selection
670 in shaping clinal variation of allelee frequencies and gene expression in Norway spruce (*Picea*
671 *abies*). *Genetics,* **191**: 865–881.
672
673 Davey JW, Cezard T, Fuentes-Utrilla P, *et al.* (2013) Special features of RAD sequencing
674 data: implications for genotyping. *Molecular Ecology,* **22**, 3151–3164.

de Miguel M, Cabezas JA, de María N, *et al.* (2014) Genetic control of functional traits related to photosynthesis and water use efficiency in *Pinus pinaster* Ait. drought response: integration of genome annotation, allele association and QTL detection for candidate gene identification. *BMC Genomics,* **15**, 464.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, *et al.* (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics,* **185**, 969–982.

Eckert, AJ, Pande B, Ersoz ES, *et al.* (2009a) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes,* **5**, 225–34.

Eckert AJ, Wegrzyn JL, Pande B, *et al.* (2009b) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. menziesii). *Genetics,* **183**: 289–298.

Eggen A (2012) The development and application of genomic selection as a new breeding paradigm. *Animal Frontier,* **2**, 10–15.

Endelman JB, Plomion C (2014) LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics,* **30**, 1623-1624.

699 Eveno E, Collada C, Guevara MA, *et al.* (2008) Contrasting patterns of selection at *Pinus*
700 *pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses.
701 *Molecular Biology and Evolution,* **25**, 417-437.
702
703 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus
704 genotype data: linked loci and correlated allele frequencies. *Genetics,* **164**, 1567–1587.
705
706 Ganapathy G, Uyenoyama MK (2009) Site frequency spectra from genomic SNP surveys.
707 *Theoretical Population Biology*, **75**, 346–354.
708
709 Gautier M, Gharbi K, Cezard T, *et al.* (2013) The effect of RAD allele dropout on the
710 estimation of genetic variation within and between populations. *Molecular Ecology,* **21**,
711 3165–3178.
712
713 Gnirke A, Melnikov A, Maguire J, *et al.* (2009) Solution hybrid selection with ultra-long
714 oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology,* **27**, 182-
715 189.
716
717 Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term
718 response. *Genetica,* **136**, 245–257.
719
720 Gonzalez-Martinez SC, Dillon S, Garnier-Géré P, *et al.* (2011). Patterns of nucleotide
721 diversity and association mapping. In: *Genetics, Genomics and Breeding of Conifer Trees*
722 *(eds. Plomion C, Bousquet J, & Kole C).* pp. 239–275. Edenbridge Science Publishers and
723 CRC Press, USA, New-York.

González-Martínez SC, Alía R, Gil L (2002) Population genetic structure in a Mediterranean pine (*Pinus pinaster* Ait.): a comparison of allozyme markers and quantitative traits. *Heredity,* **89**, 199–206.

Grivet D, Sebastiani F, Alía R, *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular Biology and Evolution,* **28**, 101-116.

Grivet D, Avia K, Vaattovaara A, *et al.* (2015) Contrasting patterns of nucleotide diversity and adaptive evolution in two widespread European pines. In review.

Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Annals of Applied Statistics*, **5**, 1780–1815.

Guerra FP, Wegrzyn JL, Sykes R, *et al.* (2013) Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist,* **197**, 162-76.

Harfourche A (1995) Variabilité géographique et hybridation interraciale chez le pin maritime (*Pinus pinaster* Ait.). Thèse de Doctorat, Université Henri Poincaré Nancy I, France, 153p.

Hartl DL, Clark AG (1998) Principles of population genetics. 3rd edition. Sinauer Associates, USA, Sunderland, MA.

748 Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal
749 breeding. *Genome,* **53**, 876–883.
750
751 Helyar SJ, Hemmer-Hansen J, Bekkevold D, *et al.* (2011) Application of SNPs for population
752 genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology*
753 *Resources,* **11**, 123–136.
754
755 Illy G (1966) Recherches sur l'amélioration génétique du pin maritime. *Annal des Sciences*
756 *Forestières,* **23**, 769-948.
757
758 Isik F (2014) Genomic selection in forest tree breeding: The concept and an outlook to the
759 future. *New Forests,* **45**, 379-401.
760
761 Jaramillo-Correa JP, Grivet D, Lepoittevin C, *et al.* (2015) Molecular proxies of climate
762 maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). *Genetics,* doi:
763 10.1534/genetics.114.173252.
764
765 Jaramillo-Correa JP, Grivet D, Terrab A, et al. (2010) The Strait of Gibraltar as a major
766 biogeographic barrier in Mediterranean conifers: a comparative phylogeographic survey.
767 *Molecular Ecology,* **19**, 5452-68.
768
769 Jermstad KD, Eckert AJ, Wegrzyn JL, *et al.* (2011) Comparative mapping in *Pinus*: Sugar
770 pine (*Pinus lambertiana* Dougl.) and loblolly pine (*Pinus taeda* L.). *Tree Genetics and*
771 *Genomes,* **7**, 457–68.
772

773 Jorge FC, Pascoal Neto C, Irle C, *et al.* (2002) Wood adhesives derived from alkaline extracts
774 of maritime pine bark: preparation, physical characteristics and bonding efficacy. *European*
775 *Journal of Wood and Wood Products,* **60**, 303-310.
776
777 Kosambi DD (1943) The estimation of map distances from recombination values. *Annals of*
778 *Human Genetics,* **12**, 172-175.
779
780 Krutovsky KV, Troggio M, Brown GR, *et al.* (2004) Comparative mapping in the Pinaceae.
781 *Genetics,* **168**, 447–61.
782
783 Kujala S, Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots
784 pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genetics and Genomes,* **8**, 1451-1467.
785
786 Lalagüe H, Csilléry K, Oddou-Muratorio S, *et al.* (2014) Nucleotide diversity and linkage
787 disequilibrium at 58 stress response and phenology candidate genes in a European beech
788 (*Fagus sylvatica* L.) population from southeastern France. *Tree Genetics and Genomes,* **10**,
789 15-26.
790
791 Lamy JB, Bouffier L, Burlett R, Plomion C, Cochard H, Delzon S (2011) Uniform selection
792 as the primary evolutionary force of cavitation resistance across a species range. *PLoS ONE,*
793 **6**, e23476.
794
795 Lamy JB, Delzon S, Bouche P, *et al.* (2014) Limited genetic variability and phenotypic
796 plasticity for cavitation resistance in a Mediterranean pine. *New Phytologist*, **201**, 874-86.
797

Lefèvre F, Boivin T, Bontemps A, *et al.* (2014) Considering evolutionary processes in adaptive forestry. *Annals of Forest Science,* **71**, 723-739

Le Provost G, Domergue F, Lalanne C, *et al.* (2013) Cuticular wax: an essential component of fast-growing maritime pine saplings to cope with water deficit. *BMC Plant Biology,* **13**, 95.

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science,* **92**, 4656–4663.

Lepoittevin C, Frigerio JM, Garnier-Géré P, *et al.* (2010) *In vitro vs. in silico* detected SNPs for the development of a genotyping array: What can we learn from a non-model species? *PLoS ONE,* **5**, e11034.

Lepoittevin C, Harvengt L, Plomion C, Garnier-Géré P (2012) Association mapping for growth, straightness and wood-chemistry traits in the *Pinus pinaster* Aquitaine breeding population. *Tree Genetics and Genomes,* **8**, 113–126.

Levy R, Borenstein E (2012) Reverse ecology: from systems to environments and back. *Advances in Experimental Medicine and Biology,* **751**, 329-345.

Lewis PO, Zaykin D (2001) Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html.

Li Z, Hällingback HR, Abrahamsson S, *et al.* (2014) Functional multi-locus QTL mapping of temporal trends in Scots pine wood traits. *Genes Genomes Genetics,* **4**, 2365-79.

Lind M, Källman T, Chen J, *et al.* (2014) A *Picea abies* linkage map based on SNP markers identifies QTLs for four aspects of resistance to *Heterobasidion parviporum* infection. *PLoS ONE,* **9**, e101049.

Mackay J, Dean JFD, Plomion C, *et al.* (2012) Towards decoding the conifer giga-genome. *Plant Molecular Biology,* **80**, 555–69.

Mariette J, Noirot C, Klopp C (2011) Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Research Notes,* **4**, 149.

Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas,* **144**, 78-79.

Martínez-García PJ, Stevens KA, Wegrzyn JL, *et al.* (2013) Combination of multipoint maximum likelihood (MML) and regression mapping algorithms to construct a high-density genetic linkage map for loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes,* **9**, 1529–35.

Mastretta-Yanes A, Arrigo N, Alvarez N, *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources,* **15**, 28-41.

Mollinari M, Margarido GRA, Vencovsky R, Garcia AAF (2009) Evaluation of algorithms used to order markers on genetic maps. *Heredity*, **103**, 494-502.

Monclus R, Leplé JC, 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.

Mullin TJ, Andersson B, Bastien JC, et al. (2011) Economic importance, breeding objectives and achievements. In: *Genetics, Genomics and Breeding of Conifer Trees (eds. Plomion C, Bousquet J, & Kole C).* pp. 40-127. Edenbridge Science Publishers and CRC Press, USA, New-York.

Neale DB, Wegrzyn JL, Stevens KA, *et al.* (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology,* **15**, R59

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science,* **9**, 325–30.

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, USA, New York.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2014) A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. *Genes, Genomes, Genetics,* **4**, 29–37.

Nystedt B, Street NR, Wetterbom A, *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature,* **497**, 579-584.

Os H, Stam P, Visser RF, Eck H (2005) RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics,* **112**, 30-40.

Ouborg NJ, Pertoldi C, Loeschcke V, *et al.* (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics,* **26**, 177–187.

Parchman TL, Gompert Z, Mudge J, *et al.* (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology,* **21**, 2991–3005.

Pavy N, Gagnon F, Rigault P, et al. (2013) Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources,* **13**, 324-336

Pavy N, Pelgas B, Laroche J, *et al.* (2012) A Spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biology,* **10**, 84.

Pavy N, Pelgas B, Beauseigle S, *et al.* (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics,* **9**, 21.

Perdiguero P, Barbero M del C, Cervera MT, et al. (2013) Molecular response to water stress in two contrasting Mediterranean pines (*Pinus pinaster* and *Pinus pinea*). *Plant Physiology and Biochemistry,* **67**, 199-208.

Plomion C, Chancerel E, Endelman JB, *et al.* (2014) Genome-wide distribution of genetic diversity and linkage disequilibrium in a mass-selected population of maritime pine. *BMC Genomics,* **15**, 171.

Plomion C, Costa P, Bahrman N (1997) Genetic analysis of needle protein in maritime pine. 1. Mapping dominant and codominant protein markers assayed on diploid tissue, in a haploid-based genetic map. *Silvae Genetica,* **46**, 161-165.

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

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology,* **20**, 1702–1716.

Puritz JB, Matz MV, Toonen RJ, *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology* **23**, 5937–5942.

Ribeiro MM, Le Provost G, Gerber S, *et al.* (2002) Origin identification of maritime pine stands in France using chloroplast single-sequence repeats. *Annals of Forest Science,* **59**, 53-62.

Ritland K, Krutovsky KV, Tsumura Y, *et al.* (2011) Genetic Mapping in Conifers. In: *Genetics, Genomics and Breeding of Conifer Trees (eds. Plomion C, Bousquet J, & Kole C).* pp. 196–238. Edenbridge Science Publishers and CRC Press, USA, New-York.

Rockman MV (2012) The Qtn Program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution,* **66**, 1-17.

Rodríguez-Quilón I, Santos-del-Blanco L, Grivet D, *et al.* (2015) Local effects drive heterozygosity-fitness correlations in an outcrossing long-lived tree. Submitted.

Rohdewald P (2002) A review of the French maritime pine bark extract (Pycnogenol®), an herbal medication with a diverse clinical pharmacology. *International Journal of Clinical Pharmacology and Therapeutics,* **40**, 158-168.

Santos CS, Pinheiro M, Silva AI, *et al.* (2012) Searching for resistance genes to *Bursaphelenchus xylophilus* using high-throughput screening. *BMC Genomics,* **13**, 599.

Santos-del-Blanco L, Climent J, González-Martínez SC, Pannell JR (2012) Genetic differentiation for size at first reproduction through male *versus* female functions in the widespread Mediterranean tree *Pinus pinaster*. *Annals of Botany,* **110**, 1449-1460.

Serra-Varela MJ, Grivet D, Vincenot L, *et al.* (2015) Does phylogeographic structure relate to climatic niche divergence? A test using maritime pine (*Pinus pinaster* Ait.). in review

945 Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association
946 studies for human complex trait genetics. *Genetics,* **187**, 367-383.
947
948 Touriño S, Selga A, Jiménez A, *et al.* (2005) Procyanidin fractions from pine (*Pinus pinaster*)
949 bark: radical scavenging power in solution, antioxidant activity in emulsion, and
950 antiproliferative effect in melanoma cells. *Journal of Agricultural and Food Chemistry,* **53**,
951 4728-4735.
952
953 Vision TJ, Brown DG, Shmoys DB, *et al.* (2000) Selective mapping: A strategy for
954 optimizing the construction of high-density linkage maps. *Genetics,* **155**, 407-420.
955
956 Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of
957 local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus*
958 *sylvestris* L.). *Tree Genetics and Genomes,* **5**, 117–132.
959
960 Weigel D (2012) Natural variation in *Arabidopsis*: from molecular genetics to ecological
961 genomics. *Plant Physiology*, **158**, 2–22.
962
963 Westbrook JW, Resende MFR, Munoz P, *et al.* (2013). Association genetics of oleoresin flow
964 in loblolly pine: discovering genes and predicting phenotype for improved resistance to bark
965 beetles and bioenergy potential. *New Phytologist,* **199**, 89–100.
966
967 **Figure and table legends**
968

969 **Figure 1** Number of SNPs developed over the last five years for maritime pine and
970 interdependence of studies (for more details see Supplemental Table 2). Studies from de
971 Miguel *et al*. (2014) in orange, Chancerel *et al.* (2013) in blue, Chancerel *et al.* (2011) in
972 green, Lepoittevin *et al*. (2010) in red, and Lavigne *et al*. (manuscript in preparation) in black.
973
974 **Figure 2** Strategy for developing SNPs segregating in three full-sib progenies (CxM, LxC,
975 LxM), from high-throughput sequencing of the parental lines (C, L and M).
976
977 **Figure 3** Strategy for developing new SNPs in candidate genes.
978
979 **Figure 4 Synonymous and non-synonymous substitutions**. (A) Synonymous (S) and non-
980 synonymous (NS) mutations at the three codon positions. (B) Proportion of transitions and
981 transversions.
982
983 **Figure 5** Comparison between the two F2 genetic linkage maps (F2_N in blue and F2_O in
984 white). Segregation distortion regions are indicated in orange.
985
986 **Figure 6** Bayesian clustering analysis for *K*=2, showing the French Atlantic (red) and
987 Portuguese (green) maritime pine gene pools. The red and green portions of each bar
988 (individual tree) indicate the probability of genetic ancestry (y-axis) for each cluster.
989
990 **Table 1** Characteristics of the F2 genetic linkage maps. F2_N: map obtained from the 9 k-
991 array (this study), F2_O: map constructed with data from Chancerel *et al*. (2013) and F2_C:
992 composite map generated from the F2_N and F2_O maps.
993

Table 2 Genetic diversity estimates for French Atlantic and Portuguese maritime pine populations.

Supporting Information

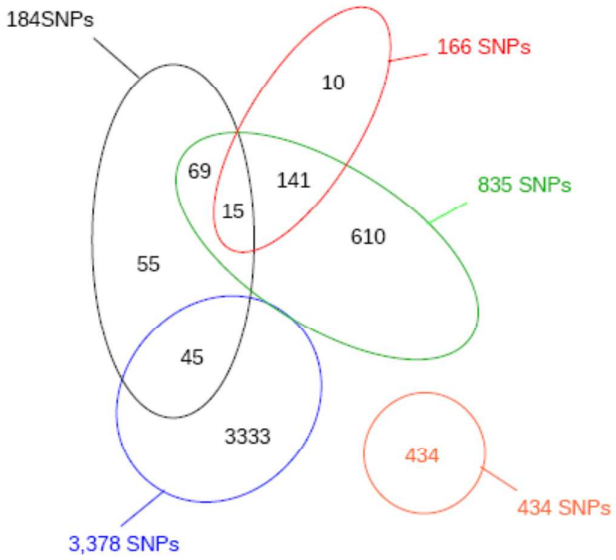
Supplemental Table 1 List of SNP arrays already developed for maritime pine and associated research questions.

Supplemental Table 2 List of SNP markers already developed for maritime pine

Supplemental Table 3 List of SNP markers with sequence, dbSNP accession, type of SNP, corresponding contig ID in PineV3 and functional annotation, and linkage group assignment on the F2, CL, ML, CM linkage maps.

Supplemental Figure 1 Distribution of minor allele frequency (MAF from France [A] and Portugal [B]) and folded site frequency spectrum (SFS from France [C] and Portugal [D]). The expected distribution of minor allele variants is skewed towards an excess of rare variants. According to Hartl and Clark (1998), more than one third of all polymorphisms would be expected to have a frequency below 5%, reflecting purifying selection.

Figure 1



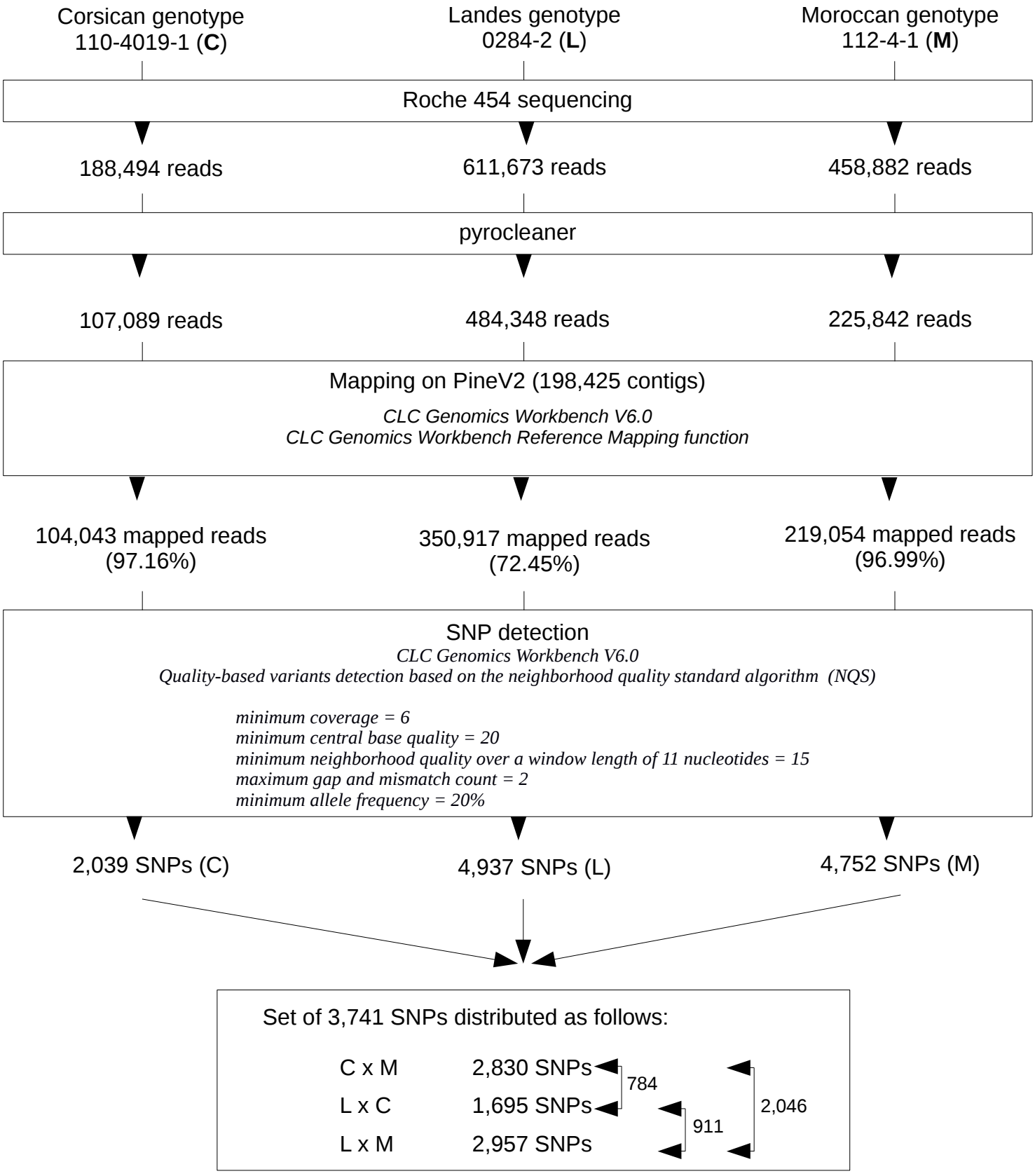
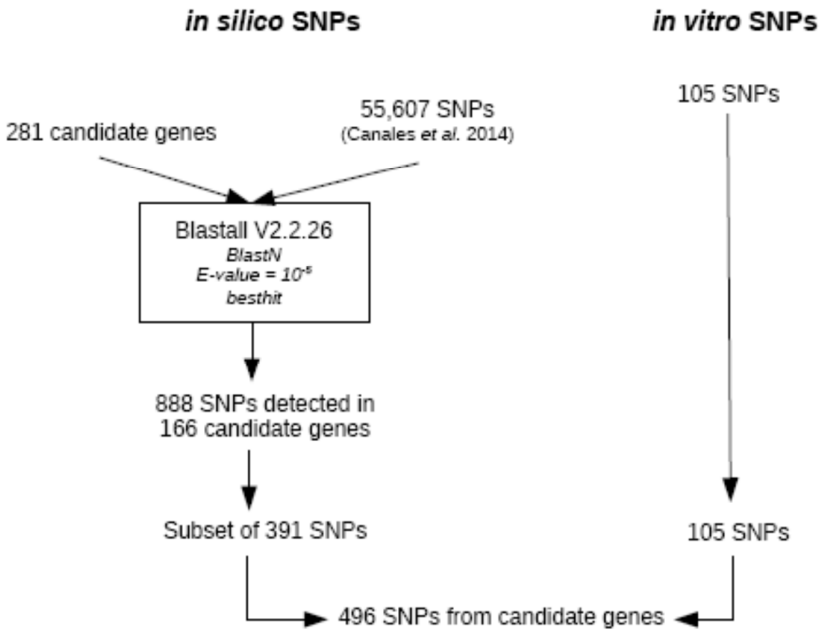
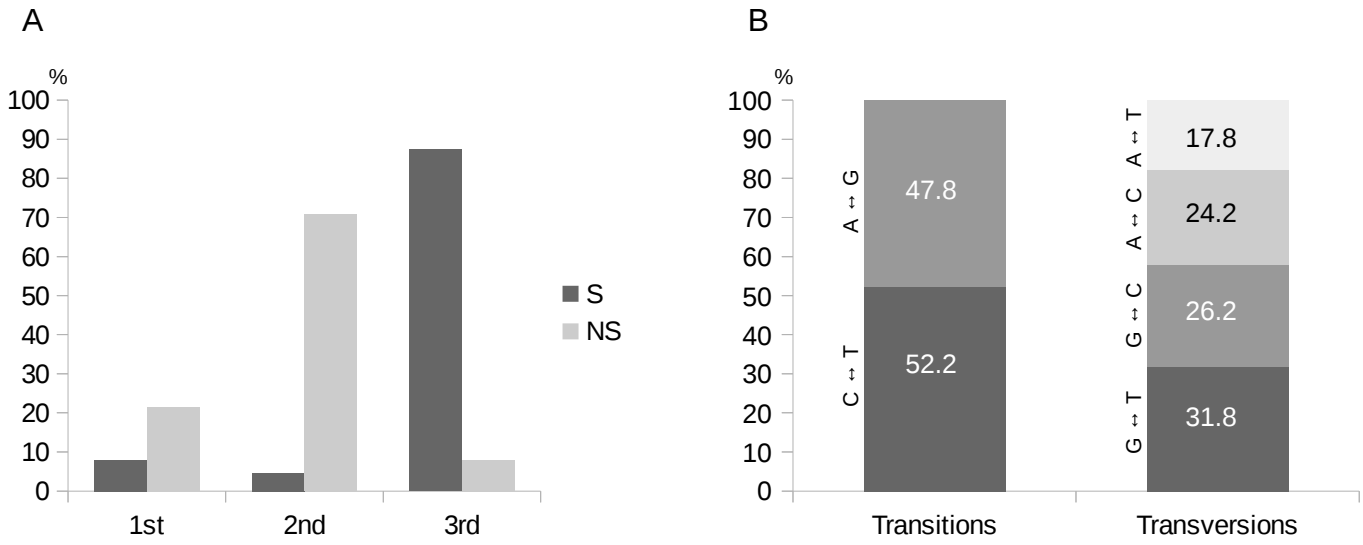

Corsican genotype
110-4019-1 (C)
Landes genotype
0284-2 (L)
Moroccan genotype
112-4-1 (M)
Roche 454 sequencing
188,494 reads
611,673 reads
458,882 reads
pyrocleaner
107,089 reads
484,348 reads
225,842 reads
Mapping on PineV2 (198,425 contigs)
CLC Genomics Workbench V6.0
CLC Genomics Workbench Reference Mapping function
104,043 mapped reads
(97.16%)
350,917 mapped reads
(72.45%)
219,054 mapped reads
(96.99%)
SNP detection
CLC Genomics Workbench V6.0
Quality-based variants detection based on the neighborhood quality standard algorithm (NQS)
minimum coverage = 6
minimum central base quality = 20
minimum neighborhood quality over a window length of 11 nucleotides = 15
maximum gap and mismatch count = 2
minimum allele frequency = 20%
2,039 SNPs (C)
4,937 SNPs (L)
4,752 SNPs (M)
Set of 3,741 SNPs distributed as follows:
C x M 2,830 SNPs
L x C 1,695 SNPs
L x M 2,957 SNPs
784
911
2,046

Figure 3





Synonymous and non-synonymous substitutions in the polymorphic SNPs.
(A) Synonymous (S) and non-synonymous (NS) mutations at the three codon positions.
(B) Proportion of transitions and transversions.

Figure 5

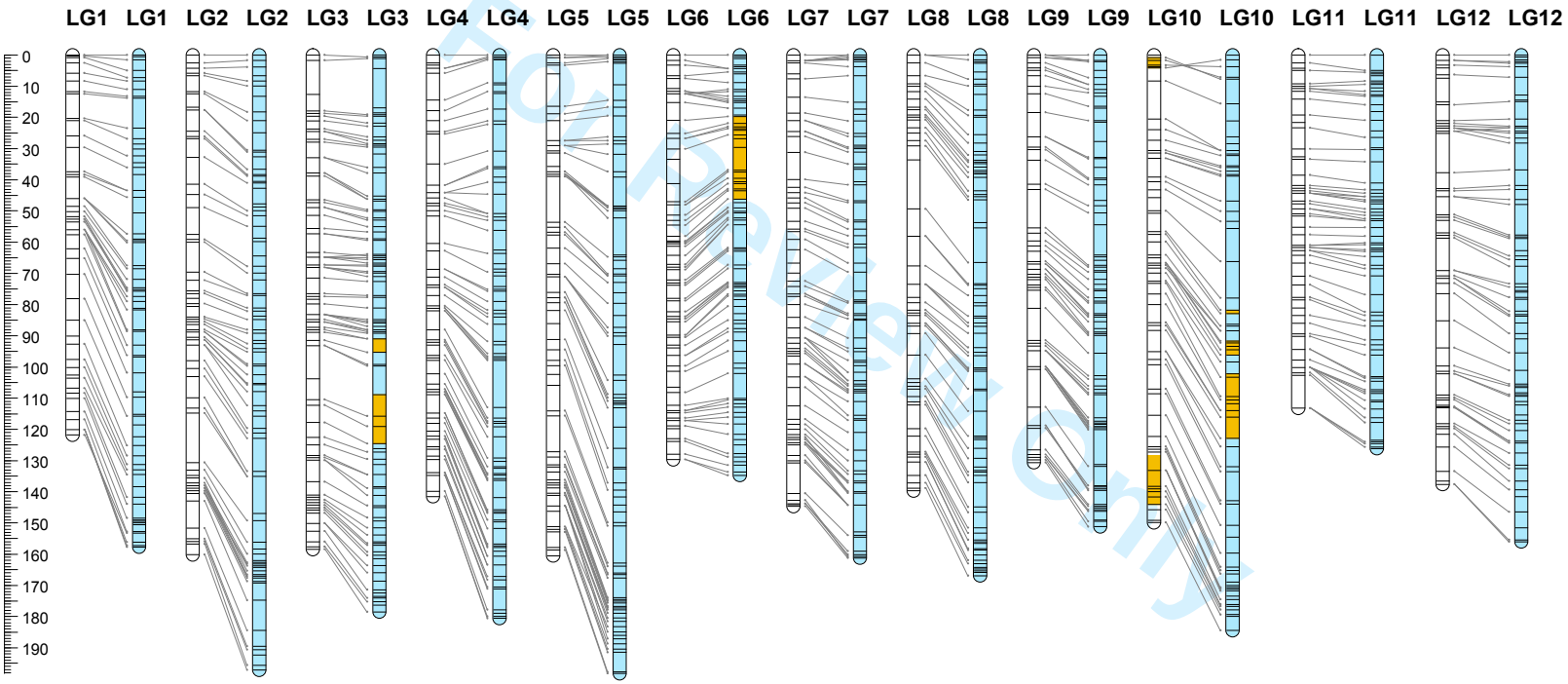


Figure 6

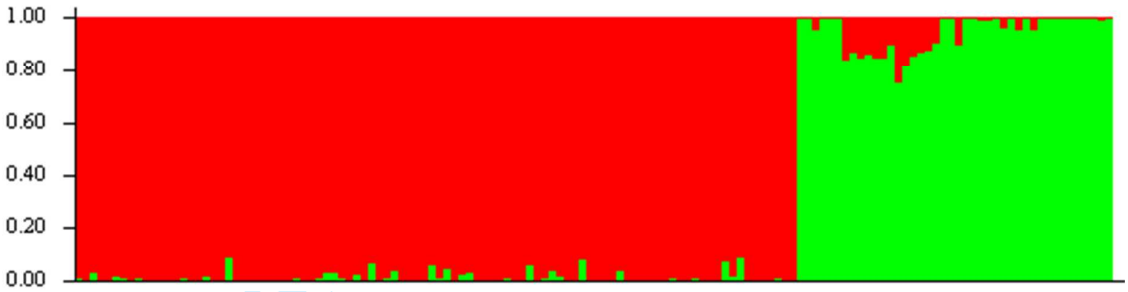


Table 1

	LG	1	2	3	4	5	6	7	8	9	10	11	12	Total
F2_N	Size (cM)	157.7	197.2	178.5	180.6	198.2	134.8	161.2	166.9	151.17	184.4	126.2	156	1,992.9
	Number of SNPs	157	152	190	176	199	163	181	155	184	148	156	191	2,052
	Number of unique positions	65	81	88	77	80	80	76	79	74	66	68	67	901
	Distance between SNPs (cM)	1.01	1.31	0.94	1.03	1	0.83	0.9	1.08	0.83	1.25	0.81	0.82	0.98
	Distance between unique positions (cM)	2.46	2.47	2.05	2.38	2.51	1.71	2.15	2.14	2.07	2.84	1.88	2.36	2.25
	Distorted SNPs (%)	0	0	7.9	0	0	16	0	0	0	18.9	0	0	3.6
F2_O	Size (cM)	121.8	160.1	158.4	141.5	160.5	129.7	144.6	139.7	130.6	149.9	113.2	137.6	1,687.6
	Number of SNPs	106	117	147	138	126	113	124	93	134	107	130	146	1,481
	Number of unique positions	41	57	60	54	56	58	58	47	45	54	49	50	629
	Distance between SNPs (cM)	1.16	1.38	1.09	1.03	1.28	1.16	1.18	1.5	0.98	1.41	0.88	0.95	1.17
	Distance between unique positions (cM)	3.04	2.86	2.69	2.67	2.92	2.28	2.54	3.04	2.97	2.83	2.36	2.81	2.75
	Distorted SNPs (%)	0	0	0	0	0	0	0	0	0	33	0	0	2.75
F2_C	Size (cM)	124.1	160.1	159.3	142.2	160.5	131.3	144.6	139.7	132.2	155.4	123.7	138.8	1,711.7
	Number of SNPs	179	177	221	191	219	188	201	174	207	173	194	229	2,353
	Number of unique positions	64	81	95	75	88	88	83	81	75	76	74	75	955
	Distance between SNPs (cM)	0.7	0.91	0.72	0.75	0.74	0.7	0.72	0.81	0.64	0.9	0.64	0.61	0.74
	Distance between unique positions (cM)	1.97	2	1.69	1.92	1.84	1.51	1.76	1.75	1.79	2.07	1.69	1.88	1.82

Table 2

Population		N	% mono	All loci			Common polymorphic loci		
				Loci #	MAF	*He*	Loci #	MAF	*He*
French Atlantic	Breeding population	46	0.242	5,016	0.201	0.266	3,981	0.254	0.336
	Natural population	50	0.241	5,016	0.198	0.263	3,981	0.250	0.332
Portugal		42	0.227	5,016	0.187	0.253	3,981	0.236	0.319