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

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

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 574–587 doi: 10.1111/1755-0998.12464

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

C. PLOMION,*† J. BARTHOLOMÉ,*† I. LESUR,*‡ C. BOURY,*† I. RODRÍGUEZ-QUILÓN,§ H. LAGRAULET,*† F. EHRENMANN,*† L. BOUFFIER,*† J. M. GION,*¶ D. GRIVET,§ M. DE MIGUEL,*† N. DE MARÍA,§ M. T. CERVERA,§ F. BAGNOLI,** F. ISIK,†† G. G. VENDRAMIN** and S. C. GONZÁLEZ-MARTÍNEZ§

*BIOGECO, UMR 1202, INRA, F-33610 Cestas, France, †BIOGECO, UMR 1202, University of Bordeaux, F-33400 Talence, France, ‡HelixVenture, F-33700 Mérignac, France, §Forest Research Centre, INIA, E-28040 Madrid, Spain, ¶UMR AGAP, CIRAD, F-33612 Cestas, France, **Institute of Biosciences and Bioresources, National Research Council, Sesto Fiorentino (FI), Italy, ††Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA

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 evolutionary 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 (2052/8410, i.e. 24.4%) segregated in the mapping population and could be mapped, providing the densest ever gene-based linkage map for this species. Based on 5016 SNPs, natural and breeding populations from the French gene pool exhibited similar level of genetic diversity. Population genetics and structure analyses based on 3981 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 genes in this species, improving our understanding of the demography and selective forces shaping its natural genetic diversity, and providing support for new breeding strategies.

Keywords: linkage mapping, maritime pine, population genetics, single nucleotide polymorphism

Received 26 February 2015; revision received 28 August 2015; accepted 3 September 2015

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 2100 m in the High Atlas (Morocco), from regions characterized by heavy annual rainfall in an Atlantic climate to dry regions in the semi-arid Mediterranean climate, and from calcareous to acidic soils (Alía & Martin 2003). Its scattered distribution has resulted in local adaptations and high levels of genetic differentiation for adaptive traits across its natural range (González-Martínez *et al.* 2002; Lamy *et al.* 2011, 2014; Santos-del-Blanco *et al.* 2012). Considerable genetic differentiation between ecotypes has been reported for various neutral molecular markers, providing clear evidence for a geographic structure of genetic diversity in this species (Bucci *et al.* 2007; Jaramillo-Correa *et al.* 2010, 2015; Santos-del-Blanco *et al.* 2012).

This fast-growing tree, which generally reaches its economically optimum size around 40 years old, has been widely planted in systematic reforestation programmes 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

Correspondence: Christophe Plomion, Fax: +33557122881; E-mail: plomion@pierroton.inra.fr

resin production. Over the last 50 years, maritime pine 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 (Jorge *et al.* 2002; Rohdewald 2002; Touriño *et al.* 2005), 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 it (mostly in Australia; Bouffier *et al.* 2013). The breeding of maritime pine began in the 1960s in south-western 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; Harfourche 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).

Maritime pine has also been adopted by the forest tree genetics research community as a key model species to study genetic variation or linked mutations underlying phenotypic variability, particularly those selected by the environment and involved in local adaptation (reviewed by Gonzalez-Martinez *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 (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 (Tables S1 and S2, Supporting information). These assays have made it possible to characterize 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), environmental association (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. Pavy *et al.* 2013, for spruce; Bartholomé *et al.* 2015, for eucalyptus), we used this platform to design a customized genotyping array for maritime pine, including 4712 SNPs from the studies cited above and 4237 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 with 7252 workable SNPs in at least one of the four tested populations. We assessed the suitability of this array for linkage mapping, identification of seed sources (Portugal vs. French Atlantic coast) and comparison between natural and breeding populations (within the French Atlantic gene pool). Finally, we discuss the 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 (8949 SNPs) for maritime pine, including the two subsets of SNPs described below.

Previously available SNPs. Over the past 5 years, several studies have reported the discovery of SNP markers in maritime pine (Tables S1 and S2, Supporting information). We first selected all available SNPs from assays based on VeraCode, GoldenGate or Infinium Illumina technologies (Illumina, San Diego, CA, USA). This number initially amounted to 4997 SNPs. As we knew that some assays had been used on a number of occasions, SNP redundancy was checked using BLASTN by aligning the flanking sequences (ranging from 87 to 419 bp in length) against themselves. We found that 4442 SNPs were used in a single study (94.3%), 255 SNPs (5.4%) in two studies and 15 SNPs (0.3%) in three studies (Fig. S1, Supporting information). Within each of the 270 groups of redundant SNPs, we retained the SNP with the longest flanking sequence. In the case of groups containing strictly identical SNPs (same polymorphic site and same flanking sequences) but with different dbSNP accessions, we arbitrarily selected one of the dbSNP ID. Thus, 285 SNPs were discarded and the final list comprised 4712 SNPs. Then, flanking sequences of each retained SNP were aligned against the most up-to-date maritime pine UniGene (PineV3, Canales *et al.* 2014: http://www.scbi.uma.es/sustainpinedb/home_page) to identify its corresponding contig and position within that contig. Alignments were performed by carrying out BLASTN searches in the BLASTALL version 2.2.26 suite (*E*-value = 10^{-5}). For only two SNPs, no hit was obtained. A single contig was found for 3494 SNPs (74.2%), whereas several contigs equally aligned (same *E*-value, alignment length and %identity) with the flanking sequences of 1218 SNPs (25.8%). Applying a more

stringent *E*-value (10^{-10}, 10^{-30}) did not resulted into a better rate of single hit. Such a high level of multiple assignment was expected considering the relatively high level of redundancy of this resource that comprises 210 513 contigs, while a recent estimate suggests that the pine genome contains about 25 000 genes (J. Wegrzyn & D. Neale, personal communication). For these SNPs, we selected the first PineV3 contig of the BLASTN output, but provided the whole list in Table S3 (Supporting information).

Newly discovered SNPs. This second subset comprised SNPs obtained by the random screening of EST data or specifically detected in candidate gene sequences.

Newly discovered SNPs from 454 sequence reads: A flowchart describing the steps involved in the identification of SNPs from 454 sequence data is shown in Fig. 1. Three genotypes involved as progenitors of interprovenance hybrids in the framework of the maritime pine breeding programme [accessions 110-4019-1 from Corsica (C), 0284-2 from Landes (L) and 112-4-1 (M) from Morocco] were used. A composite cDNA library was constructed with the SMART PCR cDNA synthesis kit (Clontech, Laboratories Inc., Mountain View, CA, USA) for each tree. The C and M libraries contained equal proportions of cDNAs from differentiating xylem, swelling buds and young needles, whereas the L library consisted of equal proportions of cDNAs from differentiating xylem, swelling and quiescent buds, and young and mature needles. Pyrosequencing (454 Titanium; Roche, Branford, CT, USA) was performed with the Roche-454 Genome Sequencing platform (FLX 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 *n*s (ambiguous base calls) >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 (PineV2) established by Chancerel *et al.* (2013). Alignments were performed with the CLC Genomics Workbench Reference Mapping function of

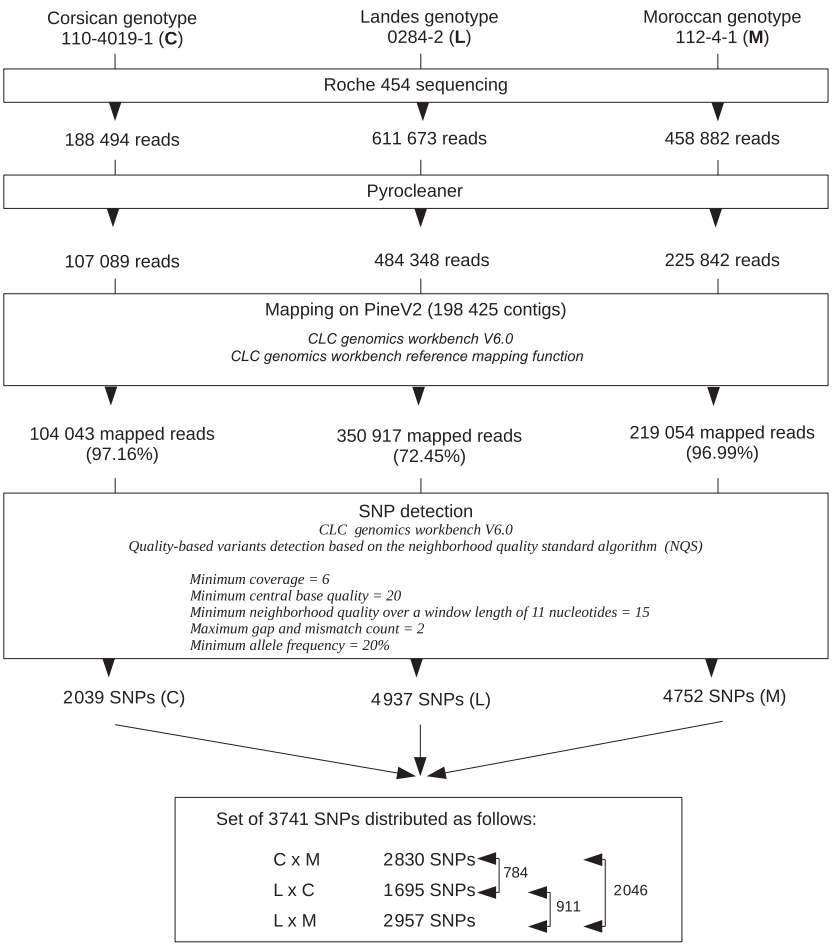

Fig. 1 Strategy for developing SNPs segregating in three full-sib progenies (C × M, L × C, L × M), from high-throughput sequencing of the parental lines (C, L and M).

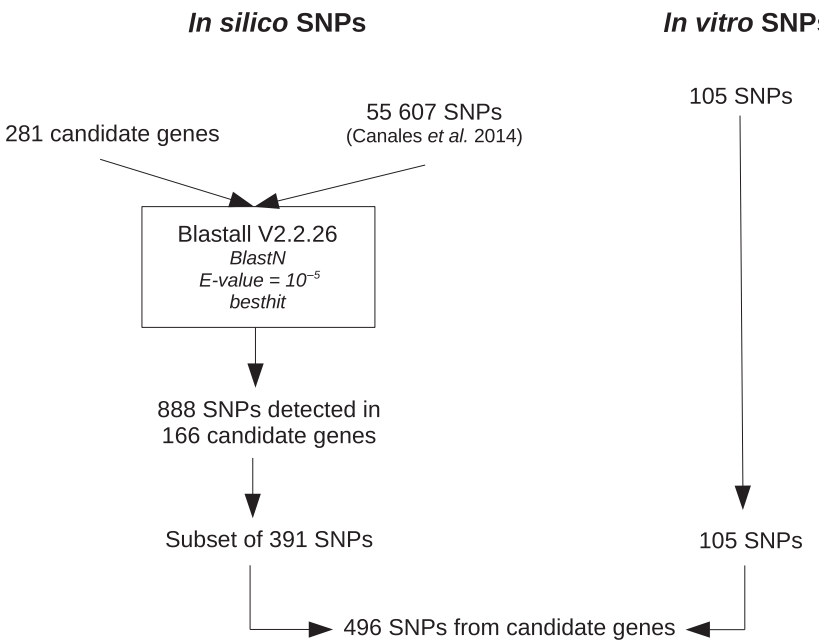

Fig. 2 Strategy for developing new SNPs in candidate genes.

CLC GENOMICS WORKBENCH version 6.0 (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 discovered SNPs from candidate genes: We obtained 105 in vitro SNPs from two full-length candidate genes with functions relating to phenology (*col*1 and *gia*) and 64 amplicons sequenced as part of the Comparative Re-sequencing in European Conifers (CRIEC) project, an EVOLTREE (http://www.evoltree.eu) initiative (D. Grivet, unpublished data). 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 & Savolainen 2012), (iii) 53 transcripts displaying differential expression in the presence and absence of pine wood nematode infection (Santos *et al.* 2012), (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 PineV3 UniGene to retrieve the best matching contig and the SNP flanking sequences from the catalogue of 55 607 available SNPs obtained by Canales *et al.* (2014; Fig. 2). Redundant SNPs were removed.

As described above, the corresponding PineV3 contigs were retrieved for each newly discovered SNP. A single contig was found for 2446 SNPs (58.3%), whereas several contigs equally aligned (same *E*-value, alignment length and %identity) with the flanking sequences of the remaining SNPs. We selected the first contig of the BLASTN output for the functional annotation, but the whole list is available in Table S3 (Supporting information).

Annotation of synonymous and nonsynonymous 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 characterized), nonsynonymous 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 Table S3 (Supporting information).

Populations studied and genetic analysis

The SNP assays were tested and validated on the basis of Mendelian segregation in a mapping pedigree (92 genotypes) and genetic diversity analysis in an elite breeding population from France (50 genotypes) and natural pop-

ulations of different origins (French Atlantic coast with 50 genotypes and Portugal with 42 genotypes).

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 × 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 seedlings were transplanted into the field in March 1999 (4 × 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 (a different set compared to that used by Chancerel *et al.* 2013) to test for the Mendelian segregation of SNP markers and to associate them with a particular linkage map 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 genotyping 477 F2 plants for 248 SNPs (unpublished) distributed over 12 linkage groups (LG), the haploid chromosome number in pines.

The R package *onemap* version 2.0-3 (Margarido *et al.* 2007; Mollinari *et al.* 2009) was used for linkage mapping. All polymorphic SNPs and individuals were considered in the analysis as they passed the missing data threshold of 5% and 1%, respectively. SNPs were clustered into LGs on the basis of a LOD score >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 & Yekutieli 2001) within each LG. The same procedure was also applied to a previous SNP data set 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 & Plomion 2014).

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

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

We first tested departure from Hardy–Weinberg equilibrium, using standard chi-squared tests 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 (H_O) and expected heterozygosity (H_E, 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 H_E only. Genetic differentiation (F_{ST}) was assessed for pairs of populations, with a set of 3981 polymorphic SNPs common to both populations, using GDA software (Lewis & 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 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, 3513 SNPs in total, corresponding to 88% of the original data set (3981 SNPs). Finally, two individuals were discarded from the French group as they presented 19% (INIA_PET11) and 10% (INIA_PET12)

missing data across the 3981 SNPs. The estimate of folded site-frequency spectrum (SFS) was therefore based on 3513 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 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 & 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 >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 initial SNP genotyping array included 8949 SNPs, but only 8410 were assayed because of the loss of bead types during array manufacture. A total of 4712 (i.e. 52.65%) were already available (166 from Lepoittevin *et al.* 2010; 835 SNPs from Chancerel *et al.* 2011; 3378 SNPs from Chancerel *et al.* 2013, 434 SNPs from de Miguel *et al.* 2014 and 184 unpublished SNPs; Fig. S1, Supporting information). The other 4237 (i.e. 47.35%) corresponded to newly discovered SNPs.

Newly discovered SNPs

In silico SNPs from 454 sequence reads: Pyrosequencing of the C (Corsica), M (Morocco) and L (Landes) libraries provided 188 494, 611 673 and 458 882 reads, respectively (Fig. 1). After cleaning, we retained 107 089, 484 348 and 225 842 reads for C, L and M, respectively. These reads were aligned on the PineV2 unigene. We identified 2039 SNPs in the C data set, 4937 in the L data set and 4752 in the M data set. We then selected SNPs in a test-cross configuration (1:1 segregation), that is 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 2830, 1695 and 2957 SNPs for the C × M, L × C and L × M crosses, respectively. In total, 784, 2046 and 911 SNPs were informative for the C, M and L parents, respectively. This procedure resulted in the retention of 3741 unique SNPs as particularly suitable for linkage mapping.

In vitro and in silico SNPs from candidate genes: Of the 281 candidate genes, 166 were associated with at least one contig of the maritime pine UniGene (Fig. 2). We retrieved 888 SNPs from these 166 contigs. Finally, after removing duplicated sequences and SNPs located <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 nonsynonymous substitutions

In total, 4257 (47.6%) SNPs were found in contigs for which a full-length protein was predicted, with 1557 (36.6%) of these SNPs located in noncoding regions. SNPs in coding sequences (2700, i.e. 63.4%) were characterized in terms of their synonymous (S) or nonsynonymous (NS) nature and codon responsible for the NS change (Fig. S2A and Table S3, Supporting information). 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 2133 (79%) nonsynonymous SNPs with the mutation affecting the first codon position in 21.5% of cases and the second codon position in 70.6%. For NS mutations, the change in the protein is indicated in Table S3 (Supporting information). Most of the SNPs were transitions (59.3%). The most common transition was C ↔ T and the most common transversion G ↔ T (Fig. S2B, Supporting information).

Linkage mapping

Of the 8410 SNPs used to genotype the 92 trees of the F2 mapping population, 4634 (55.1%) were monomor-

phic (including 2231 SNPs, about half, from the newly developed set), 2052 (from 1672 different contigs) were polymorphic (24.4%) and the remaining 1724 (20.5%) corresponding to failed assays due to poor signal or poor clustering performance as already presented in Chancerel *et al.* (2013). While lower failure rates (<10%) were usually obtained in crop plants where more information are available to optimize the selection of the SNPs (e.g. Delourme *et al.* 2013 in oilseed rape, Sim *et al.* 2012 in tomato, Li *et al.* 2014 in alfalfa), a similar failure rate was reported in oak (19.6%; Lepoittevin *et al.* 2015) and Douglas-fir (27.5%; Howe *et al.* 2013) using the iselect Infinium genotyping platform. Besides, as expected, the rate of failed assays was much lower in the set of already validated assays (11.9%) compared to that found in the set of newly developed ones (27.3%). All polymorphic SNPs passed the quality threshold and were therefore used in linkage analysis. The total map length was 1993 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 1672 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; Fig. 3). The order of SNPs was similar (Spearman $\rho = 0.98$) to that obtained from the linkage map reconstructed with the data set of Chancerel *et al.* (2013) based on different genotypes of the same progeny (F2_O in Table 1). Only 56 of the 1180 markers common to the two genetic maps differed in order between the two maps (Fig. 3). 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 2353 SNPs (including 1121 SNPs in the same contigs). This composite map included 1661 different loci and is therefore the densest yet gene-based linkage map for this species. The

Table 1 Characteristics of the F2 genetic linkage maps

	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	1992.9
	Number of SNPs	157	152	190	176	199	163	181	155	184	148	156	191	2052
	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	1687.6
	Number of SNPs	106	117	147	138	126	113	124	93	134	107	130	146	1481
	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	1711.7
	Number of SNPs	179	177	221	191	219	188	201	174	207	173	194	229	2353
	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

F2_N: map obtained from the 9 k-array (this study), F2_O: map constructed with data from Chancerel *et al.* (2013) and F2_C: composite map generated from the F2_N and F2_O maps.

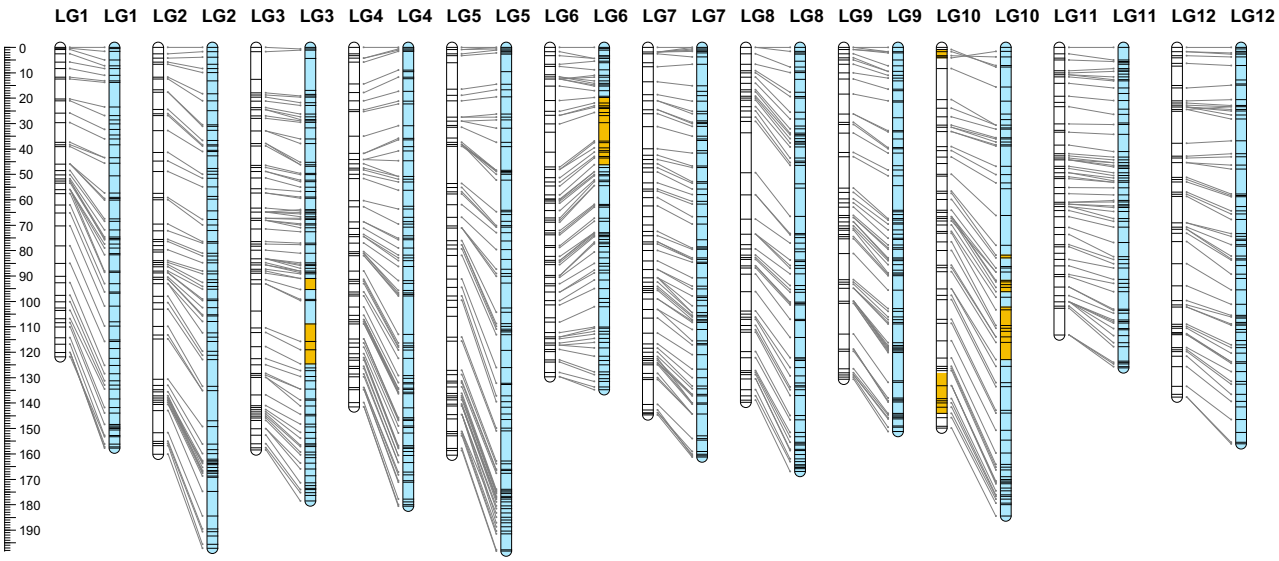

Fig. 3 Comparison between the two F2 genetic linkage maps (F2_N in blue and F2_O in white). Segregation distortion regions are indicated in orange.

total map length was 1711.7 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: six (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 5016 SNPs (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 (H_E 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 ecotype was broad enough to collect most of the standing genetic variation. 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 (H_E 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} above 0.25. These SNP markers are good candidates to replace the biochemical assay (based on terpene content analysis, Baradat & 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, seedlots 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,

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

			All SNPs			Common polymorphic SNPs		
Population	*n*	% Mono	Loci #	MAF	H_E	Loci #	MAF	H_E
French Atlantic								
Breeding population	46	0.242	5016	0.201	0.266	3981	0.254	0.336
Natural population	50	0.241	5016	0.198	0.263	3981	0.250	0.332
Portugal	42	0.227	5016	0.187	0.253	3981	0.236	0.319

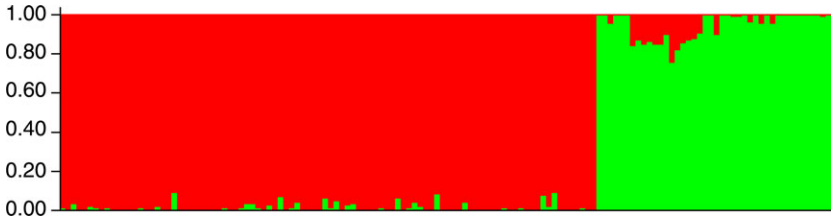

Fig. 4 Bayesian clustering analysis for $K = 2$, showing the French Atlantic (red) and Portuguese (green) maritime pine gene pools. The red and green portions of each bar (individual tree) indicate the probability of genetic ancestry (y-axis) for each cluster.

with the identification of only two gene pools ($K = 2$), including the French natural and breeding populations grouped together into a single gene pool significantly separated from the Portuguese population (Fig. 4), is a 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 optimized for this species (Chancerel *et al.* 2013).

As expected, the SFS of both the French and Portuguese populations showed a deficit of the observed low-frequency variants (Fig. S3, Supporting information). 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), and the sampling scheme (a small population size), 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 might 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. We also discussed the limitations that this SNP array represents for genetic applications.

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 and 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 cost-efficient mapping of hundreds of genes, as the bin set of 92 highly recombinant F2s represented 95% of the recombination occurring among the initial set of 477 genotypes used to construct the framework linkage map. The genetic map obtained with this strategy was highly reliable and accurate, as revealed by comparison with a previously developed map based on different genotypes from the same progeny (Chancerel *et al.* 2013). A low fraction of molecular markers were distorted, and they did not alter marker order. The use of markers common to the two sets of offspring made it possible to construct a composite genetic linkage map confirming the positions of most of the previously mapped markers. The composite linkage map obtained in this study is the densest gene-based map for maritime pine and is in the range of recently published linkage maps for other conifer species (Pavy *et al.* 2012; Neves *et al.* 2014). However, a considerable fraction of SNPs clustered in the same position. Once high-throughput genotyping of thousands of molecular markers has been achieved, the main limitations for genetic linkage mapping will come from the high number of individuals needed to detect recombination events between closely positioned genes (Bartholomé *et al.* 2015).

Association mapping

In long-lived outbreeding organisms, such as forest trees, association mapping is an attractive alternative to QTL mapping, as it can be used to identify molecular variation underlying phenotypes from multiple genetic backgrounds without the need to produce segregating progenies (Neale & Savolainen 2004). Under certain circumstances (e.g. moderate–high heritability), association mapping can also be applied to natural populations. This approach is useful for studies of adaptive traits in the precise environments in which they evolved (see 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 (Gonzalez-Martinez *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; Cabezas *et al.* 2015). Apart from including all SNPs reported to display significant genotype:phenotype associations in previous studies, our new SNP array provides the largest genotyped platform developed to date in maritime pine, with 7252 workable SNPs. Despite its modest size, this resource has the potential to stimulate new association studies for a wide range of adaptive and production traits in this species.

Given the large size of conifer genomes (24.5 Gb for maritime pine, Chagné *et al.* 2002), our SNP array would only covers a small fraction of the potential relevant variation underlying phenotypes. However, this assay is enriched in SNPs from candidate genes, based on all the available information (published or unpublished) concerning genes displaying signatures of natural selection (Eveno *et al.* 2008; Grivet *et al.* 2011), involved in environmental associations (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 expect the use of this SNP genotyping array to generate highly informative data. For example, Westbrook *et al.* (2013) found that a small number of significantly associated SNPs (~20 to 30) had the same predictive power as the full data set (4854 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 & Stephens 2011). Although largest SNP assays would be desirable, some thousands of well-selected SNPs have been found to be enough to provide relevant insights into polygenic adaptation patterns (Berg & Coop 2014).

Genomic selection

High-throughput genotyping platforms, such as the new SNP array developed here, can guide breeding and selection decisions (Eggen 2012). 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 & 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. GS is expected to revolutionize 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 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 SNP array presented here constitutes a valuable tool to carry out GS proof of concept in maritime pine population of limited effective size. However, because of the relatively largebreeding effective population size (~135 for the French breeding programme), operative implementation of GS would still need further development of SNP resources. Based on deterministic simulations, Grattapaglia & Resende (2011) suggested a marker density of over 10 SNPs per cM to reach an accuracy of 0.7 with GS. This would translate to at least 15 000 SNP markers for maritime pine, half of which provided in the SNP array developed here.

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, particu-

larly in the Iberian Peninsula (Burban & 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. Our 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), SFS statistics (Eckert *et al.* 2009b), correlations with environmental variables (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 & Borenstein 2012). 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 SFS (Albrechtsen *et al.* 2011; reviewed by Helyar *et al.* 2011). Ascertainment bias has the potential of skewing the SFS towards common alleles, as is the case for our SNP array. A correction, explicitly incorporating SNP ascertainment bias into population genetics models, or the use of methods robust to ascertainment bias are necessary to accurately assess population genetic inferences (Albrechtsen *et al.* 2011; Excoffier *et al.* 2013). Furthermore, most of the SNPs were obtained from independent contigs of the UniGene, precluding the use of linkage disequilibrium (LD) and haplotype diversity as metrics for population genetic inferences. This limitation could therefore potentially reduce the success in identifying loci underlying local adaptation, as this approach – similarly to genome-wide association – exploits LD to indirectly identify adaptive SNPs by relying upon the premise that some markers in LD with a causal SNP should be associated with important ecological variables as well. Despite these limitations, thanks to the methodological advances, our SNP array should allow to interpret the signatures left in the genome by different evolutionary forces, and to make inferences about complex population processes and geographic patterns of genome-wide genetic variation.

Conclusion and future direction in genomic resource development

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. We found this technology highly reliable considering the extremely low level of missing data compared to alternatives such as restriction site-associated DNA sequencing (RAD-Seq, Baird *et al.* 2008), which made unnecessary imputing data. Infinium technology is also unaffected by allele drop out, a major drawback for genetic inferences with RAD-Seq 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). The SNP array presented here represents a major step forward for population and conservation genetics and for breeding in maritime pine; however, we also discussed a number of limitations, mostly derived from the large conifer genome sizes and low levels of LD.

In the near future, we will develop new generation genotyping-by-sequencing methods, which should make it possible to decrease the problem of ascertainment bias, because the individuals of interest are sequenced directly. Considering missing data allele drop out concerns, 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) and 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); and 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.

References

Albrechtsen A, Nielsen FC, Nielsen R (2011) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, **27**, 2534–2547.

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

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

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

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

Bartholomé J, Mandrou E, Mabiala A *et al.* (2015) High-resolution genetic linkage maps of *Eucalyptus* improve BRASUZ1 reference genome assembly. *New Phytologist*, **206**, 1283–1296.

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

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics*, **10**, e1004412.

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, Uppsala, Sweden.

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*. *Tree Physiology*. doi:10.1093/treephys/tpv050.

Canales J, Bautista R, Label P *et al.* (2014) *De novo* assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnology Journal*, **12**, 286–299.

Cappa EP, El-Kassaby YA, Garcia MN *et al.* (2013) Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. *PLoS One*, **8**, e81267.

Chagné D, Lalanne C, Madur D *et al.* (2002) A high density linkage map of *Pinus pinaster* based on AFLPs. *Annals of Forest Science*, **59**, 627–636.

Chancerel E, Lepoittevin C, Le Provost G *et al.* (2011) Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. *BMC Genomics*, **12**, 368.

Chancerel E, Lamy JB, Lesur I *et al.* (2013) High density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology*, **11**, 50.

Chen J, Källman T, Ma X *et al.* (2012) Disentangling the roles of history and local selection in shaping clinal variation of allelee frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics*, **191**, 865–881.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

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

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**, 225–234.

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.

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.

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

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

Eveno E, Collada C, Guevara MA *et al.* (2008) Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution*, **25**, 417–437.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Ganapathy G, Uyenoyama MK (2009) Site frequency spectra from genomic SNP surveys. *Theoretical Population Biology*, **75**, 346–354.

Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **21**, 3165–3178.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, **136**, 245–257.

Gonzalez-Martinez SC, Dillon S, Garnier-Géré P *et al.* (2011) Patterns of nucleotide diversity and association mapping. In: *Genetics, Genomics and Breeding of Conifer Trees* (eds Plomion C, Bousquet J, Kole C), pp. 239–275. Edenbridge Science Publishers and CRC Press, 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.

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. *Tree Genetics and Genomes*, **7**, 241–255.

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.

Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *The 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**, 162–176.

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

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. *Genome*, **53**, 876–883.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

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

Illy G (1966) Recherches sur l'amélioration génétique du pin maritime. *Annal des Sciences Forestières*, **23**, 769–948.

Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forests*, **45**, 379–401.

Isik F, Bartholomé J, Farjat A *et al.* (2015) Genomic selection in maritime pine. *Plant Science*. doi: 10.1016/j.plantsci.2015.08.006

Jaramillo-Correa JP, Grivet D, Terrab A *et al.* (2010) The Strait of Gibraltar as a major biogeographic barrier in Mediterranean conifers: a comparative phylogeographic survey. *Molecular Ecology*, **19**, 5452–5468.

Jaramillo-Correa JP, Grivet D, Lepoittevin C *et al.* (2015) Molecular proxies of climate maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). *Genetics*, **199**, 793–807.

Jermstad KD, Eckert AJ, Wegrzyn JL *et al.* (2011) Comparative mapping in *Pinus*: Sugar pine (*Pinus lambertiana* Dougl.) and loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes*, **7**, 457–468.

Jorge FC, Pascoal Neto C, Irle C *et al.* (2002) Wood adhesives derived from alkaline extracts of maritime pine bark: preparation, physical characteristics and bonding efficacy. *European Journal of Wood and Wood Products*, **60**, 303–310.

Kosambi DD (1943) The estimation of map distances from recombination values. *Annals of Human Genetics*, **12**, 172–175.

Kujala S, Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genetics and Genomes*, **8**, 1451–1467.

Lalagüe H, Csilléry K, Oddou-Muratorio S *et al.* (2014) Nucleotide diversity and linkage disequilibrium at 58 stress response and phenology candidate genes in a European beech (*Fagus sylvatica* L.) population from southeastern France. *Tree Genetics and Genomes*, **10**, 15–26.

Lamy JB, Bouffier L, Burlett R, Plomion C, Cochard H, Delzon S (2011) Uniform selection as the primary evolutionary force of cavitation resistance across a species range. *PLoS One*, **6**, e23476.

Lamy JB, Delzon S, Bouche P *et al.* (2014) Limited genetic variability and phenotypic plasticity for cavitation resistance in a Mediterranean pine. *New Phytologist*, **201**, 874–886.

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.

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

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.

Lepoittevin C, Bodénès C, Chancerel E *et al.* (2015) Single-nucleotide polymorphism discovery and validation in high density SNP array for genetic analysis in European white oaks. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12407.

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

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

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.

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**, 1529–1535.

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.

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.

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

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, New York.

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

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, 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.

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, 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.

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, 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.

Perdiguero P, Barbero MdelC, 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, 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.

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.

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.

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.). *Global Ecology and Biogeography*. doi: 10.1111/geb.12369.

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

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Touriño S, Selga A, Jiménez A *et al.* (2005) Procyanidin fractions from pine (*Pinus pinaster*) bark: radical scavenging power in solution, antioxidant activity in emulsion, and antiproliferative effect in melanoma cells. *Journal of Agricultural and Food Chemistry*, **53**, 4728–4735.

Vision TJ, Brown DG, Shmoys DB *et al.* (2000) Selective mapping: a strategy for optimizing the construction of high-density linkage maps. *Genetics*, **155**, 407–420.

Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genetics and Genomes*, **5**, 117–132.

Weigel D (2012) Natural variation in *Arabidopsis*: from molecular genetics to ecological genomics. *Plant Physiology*, **158**, 2–22.

Westbrook JW, Resende MFR, Munoz P *et al.* (2013) Association genetics of oleoresin flow in loblolly pine: discovering genes and predicting phenotype for improved resistance to bark beetles and bioenergy potential. *New Phytologist*, **199**, 89–100.

I.L. and H.L. contributed to in silico SNP detection in mapping populations; I.L. to SNP array design and SNP annotation; C.B., L.B., J.M.G., F.B., G.G.V., D.G., I.R.Q. and S.C.G.M. to logistical aspects of genotyping; C.B., J.B. I.R.Q., S.C.G.M. and D.G. to SNP scoring; C.P., I.L., M.d.M., M.T.C., N.d.M. and F.H. to database compilation from published and unpublished data; J.B. to linkage mapping; S.C.G.M. and D.G. to population genetics analysis; and C.P. to 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.

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. The assembled contigs of the third maritime pine UniGene are available from: http://www.scbi.uma.es/sustainpinedb/home_page. Raw data (SNP genotypes for the breeding population, natural populations from France and Portugal, F2 mapping populations) and SNP accession numbers are available in Table S3 (Supporting information).

Supporting Information

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

Fig. S1 Number of SNPs developed over the last 5 years for maritime pine and interdependence of studies (for more details see Table S2).

Fig. S2 Synonymous and nonsynonymous substitutions.

Fig. S3 Distribution of minor allele frequency [MAF from France (A) and Portugal (B)] and folded site-frequency spectrum [SFS from France (C) and Portugal (D)].

Table S1 List of SNP arrays already developed for maritime pine and associated research questions.

Table S2 List of SNP markers already developed for maritime pine.

Table S3 Characteristics of the data sets used in the present study.