MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 1446–1459 doi: 10.1111/1755-0998.12407

Single-nucleotide polymorphism discovery and validation in high-density SNP array for genetic analysis in European white oaks

C. LEPOITTEVIN,*† C. BODÉNÈS,*† E. CHANCEREL,*† L. VILLATE,*† T. LANG,*†‡ I. LESUR,*†§ C. BOURY,*† F. EHRENMANN,*† D. ZELENICA,¶ A. BOLAND,¶ C. BESSE,¶ P. GARNIER-GÉRÉ,*† C. PLOMION*† and A. KREMER*†

*UMR 1202 BIOGECO, INRA, Cestas F-33610, France, †UMR 1202 BIOGECO, University of Bordeaux, Pessac F-33600, France, ‡Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China, §HelixVenture, Mérignac F-33700, France, ¶CEA, Institut de Génomique, Centre National de Génotypage, 2 rue Gaston Crémieux, CP5721, Evry Cedex F-91057, France

Abstract

An Illumina Infinium SNP genotyping array was constructed for European white oaks. Six individuals of *Quercus petraea* and *Q. robur* were considered for SNP discovery using both previously obtained Sanger sequences across 676 gene regions (1371 *in vitro* SNPs) and Roche 454 technology sequences from 5112 contigs (6542 putative *in silico* SNPs). The 7913 SNPs were genotyped across the six parental individuals, full-sib progenies (one within each species and two interspecific crosses between *Q. petraea* and *Q. robur*) and three natural populations from south-western France that included two additional interfertile white oak species (*Q. pubescens* and *Q. pyrenaica*). The genotyping success rate in mapping populations was 80.4% overall and 72.4% for polymorphic SNPs. In natural populations, these figures were lower (54.8% and 51.9%, respectively). Illumina genotype clusters with compression (shift of clusters on the normalized *x*-axis) were detected in ~25% of the successfully genotyped SNPs and may be due to the presence of paralogues. Compressed clusters were significantly more frequent for SNPs showing *a priori* incorrect Illumina genotypes, suggesting that they should be considered with caution or discarded. Altogether, these results show a high experimental error rate for the Infinium array (between 15% and 20% of SNPs potentially unreliable and 10% when excluding all compressed clusters), and recommendations are proposed when applying this type of high-throughput technique. Finally, results on diversity levels and shared polymorphisms across targeted white oaks and more distant species of the *Quercus* genus are discussed, and perspectives for future comparative studies are proposed.

Keywords: cluster compression, genotyping, genotyping error rates, infinium, oaks, SNP detection

Received 17 December 2014; revision received 20 March 2015; accepted 20 March 2015

Introduction

Considerable effort has been made to identify genes of ecological significance in recent years, particularly in nonmodel species, such as forest trees, in response to research needs for biodiversity evaluation, management and conservation or adaptation to climate change (Grattapaglia *et al*. 2009; Abbott 2012). Extensive tree genomics studies have contributed to the large-scale sequencing of expressed sequence tags (ESTs) (Pavy *et al*. 2005; Canales *et al*. 2014) and PCR amplicons from candidate genes (Eveno *et al*. 2008; Ersoz *et al*. 2010), the ecological relevance of which remains to be confirmed in independent association or QTL detection studies (Neale & Kremer 2011). These experiments are based on analyses of statistical correlations between nucleotide diversity and phenotypic trait variation. We describe here the identification of single-nucleotide polymorphisms (SNPs) from previously sequenced amplicons and ESTs in oaks, the construction of a SNP array and the validation of the detected SNPs in mapping pedigrees and natural populations of European white oak species.

Oaks are widespread throughout the Northern Hemisphere and are collectively represented by several hundred species (Govaerts & Frodin 1998). They make a substantial contribution to the forest sector economy in many countries (Johnson *et al.* 2009) and provide important

Correspondence: Camille Lepoittevin, Fax: +33(0)5-57-12-28-81; E-mail: camille@pierroton.inra.fr

© 2015 John Wiley & Sons Ltd

ecological services in diverse environments, ranging from deserts to tropical humid forests and from lowlands to high-altitude environments (Nixon 2006). They are also used as case studies to address questions on species delineation, hybridization and evolutionary history (Petit *et al.* 2013). Recent transcriptomic and gene expression profiling studies in oaks have led to the construction of large cDNA libraries (Ueno *et al.* 2010; Kremer *et al.* 2012; Tarkka *et al.* 2013), and RNA-Seq studies have made it possible to identify genes involved in the response to waterlogging (Le Provost *et al.* 2012; Rasheed-Depardieu *et al.* 2012) and in bud dormancy (Ueno *et al.* 2013). However, current SNP catalogues are limited to genes from only a few pathways, such as those involved in water metabolism (Vornam *et al.* 2011), drought tolerance (Homolka *et al.* 2013), apical bud phenology (Derory *et al.* 2010) and citrate cycle metabolism (Vidalis *et al.* 2013). SNP diversity was recently explored in more detail, with the development of 384-plex arrays for phenology-related genes (Alberto *et al.* 2013) or genes involved in species differentiation (Guichoux *et al.* 2013).

In this study, we aimed to construct the first high-density Illumina® Infinium SNP array for oak. Our strategy involved the inclusion of SNPs for a large number of genes for future analyses of genetic diversity and differentiation across different oak species and the development of a high-density linkage map. Two major resources were used for SNP discovery: (i) a set of more than 676 gene regions previously sequenced by Sanger methods, for 24 individuals (referred to as *in vitro* SNPs) and (ii) the Roche 454 libraries used by Ueno *et al.* (2010) to construct the first oak Unigene OCV1 (referred to as putative *in silico* SNPs). We took a special care in comparing the efficiency of methods for putative SNPs detection and their consistency with the genotypes obtained from the Infinium array. This allowed us to propose filters to increase the quality of *in silico* SNP detection, estimate experimental genotyping error rates in a subset of the studied material, discuss the possible causes of inconsistencies and errors and finally make a few recommendations for developing similar resources in nonmodel species.

Methods

Plant material for SNP genotyping

We analysed the Mendelian segregation of SNPs in four full-sib families routinely used for genetic mapping and QTL detection in white oak species (Bodénès *et al.* 2012). They were obtained by four crosses within and between three *Q. petraea* (QS21, QS28 and QS29) and three *Q. robur* individuals (11P, 3P and A4) from the Pierroton and Arcachon populations (Table S1 and Figure S1, Supporting information): one cross within *Q. robur* (P1 pedigree, 3P × A4: 369 full-sibs), *Q. robur* × *Q. petraea* crosses (P2 pedigree, 11P × QS28: 178 full-sibs and P3 pedigree, 11P × QS29: 114 full-sibs), and one cross within *Q. petraea* (P4 pedigree, QS28 × QS21: 398 full-sibs). We included the six parents of the different crosses as controls in the genotyped material.

The genetic diversity of the selected SNPs was also assessed within natural populations of four white oak species (*Q. petraea*, *Q. robur*, *Q. pubescens* and *Q. pyrenaica*, see Table S1 and Figure S1, Supporting information). We sampled 283 trees in total. Samples were collected from eight *Q. petraea*, 65 *Q. pyrenaica*, 66 *Q. pubescens* and 73 *Q. robur* trees from a single population of mixed composition (Briouant). Another set of 71 *Q. petraea* trees was sampled in two oak populations located on the northern slopes of the Pyrenean foothills (49 trees from Ade and 22 trees from Ibos). These two populations are separated by less than 15 km and are actually of mixed composition, as they also include *Q. robur*. The three populations (Ade, Ibos and Briouant) are less than 100 km apart. Hybridization events between the four species have repeatedly been inferred (Lepais *et al.* 2009), despite evidence of strong, asymmetric barriers to reproduction (Abadie *et al.* 2012; Lepais *et al.* 2013). However, the sampled trees were selected on the basis of a prior admixture analysis, which made it possible to exclude any individuals resulting from recent hybridization events (Lepais *et al.* 2009).

Single-nucleotide polymorphism discovery and array design

The selection of *in vitro* SNPs for the Infinium assay was optimized by extracting data for the six parents of the oak mapping pedigrees as a subset of the 24 individuals of *Q. robur* and *Q. petraea* (see Kremer *et al.* 2002) included in a previous allelic resequencing project for oak species in the framework of EVOLTREE network of excellence activities (http://www.evoltree.eu/). These resources will be fully published in a companion work, but we summarize what has been valued in this study: Sanger sequences were produced from gene fragments selected from a set of over 100 000 oak ESTs deposited at the NCBI (Ueno *et al.* 2010). Sequence data were assembled and extracted with the PhredPhrap suite of programs assembled in our own bioinformatics pipeline, SeqQual which was designed to ensure the detection of high-quality *in vitro* SNPs, excluding all bases with a Phd score below 30 (see http://www.phrap.org/phredphrapconsed.html, Brousseau *et al.* 2014; El Mujtar *et al.* 2014). For the particular subset of six pedigree individuals, we further removed unnecessary alignment gaps inserted due to the use of the previous broader discovery

© 2015 John Wiley & Sons Ltd

panel and checked the presence of SNPs using several programs from the SeqQual pipeline (*i.e. pick-seq.pl* and *remove1-bad-pos_aln.pl* from Brousseau *et al.* 2014 and *make_consensus_IUPAC.pl*, *print_source_SNP-statistic-haplo.pl* and *SNP-statistic0/1/2-haplo.pl* from El Mujtar *et al.* 2014; Table S2, Supporting information). The *snp2Illumina* Perl script (Lepoittevin *et al.* 2010, available as Appendix S1, Supporting information) was then used to extract bi-allelic SNPs, which then output in the form of a SequenceList file compatible with Illumina Assay Design Tool (ADT) software (available at http://www.Illumina.com). The ADT software assigned a functionality score to each SNP, corresponding to a predicted probability of genotyping success, taking into account the sequence conformation around the SNP and the lack of repetitive elements in the surrounding sequence (Shen *et al.* 2005). One to three *in vitro* SNPs per contig were selected on the basis of their functionality score (>0.6) and the distance between them (>60 nucleotides), in accordance with Illumina recommendations. The final set of *in vitro* SNPs consisted of 1371 SNPs spread across 676 gene regions (corresponding to 709 gene fragments, consensus sequences provided in Appendix S2 (Supporting information); SNP characteristics are further detailed in Table S3 (Supporting information) and in the data accessibility part below).

The set of putative *in silico* SNPs was detected in the oak unigene OakContigV1 (OCV1) described by Ueno *et al.* (2010). This assembly brings together 125 925 Sanger and 1 578 013 454 reads from 34 cDNA libraries of *Q. robur* and *Q. petraea* and is available from http://ngspipelines.toulouse.inra.fr:9024/ngspipelines/#!/NGS-pipelines/Quercus%20robur%20-%20qrobur. In this assembly, the duplicated reads were identified using mega BLAST search (http://www.ncbi.nlm.nih.gov/blast/html/megablast.html) with minimum hit score of 100, per cent identity cut-off of 98, alignment of reads starting exactly at the same position and ending in a 70-bp window of the end of the longest sequence. This process was implemented by the Genotoul bioinformatics platform along with the Pyrocleaner tool from NG6 (http://vm-bioinfo.toulouse.inra.fr/ng6/) which is dedicated to handle multiple copy reads. Some of the 454 reads used in OCV1 were obtained by sequencing mRNA extracted from the six parental lines of the mapping pedigrees. Aligned sequences from these six pedigree individuals were extracted in batches from each contig and also masking polymorphisms likely to be sequencing errors (*i.e.* either with a minor allele frequency (MAF) below 5%, whatever the alignment depth, or singletons with a depth of between five and 20). The same pipeline and programs were used than for *in vitro* SNPs above (with the additional program *maskAln.pl* from Brousseau *et al.* 2014, Table S2, Supporting information). We selected one or two putative *in silico* SNPs per contig, on the basis of the following criteria: detection depth >6, minor allele appearing at least three times, MAF > 20%, functionality score >0.6 and consecutive SNPs located at least 60 nucleotides apart. We finally checked by blast analysis that *in silico* and *in vitro* SNPs did not target the same gene fragment. The final set of putative *in silico* SNPs contained 6542 polymorphisms located in 5112 contigs (fasta alignments deposited at the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fd862; SNP characteristics are further detailed in Table S3 (Supporting information) and in the data accessibility part below).

DNA extraction and quality controls

For each tree, buds or young leaves were collected and stored either in silica gel or at −80°C until DNA extraction. Plant material was crushed in a mixer mill (Retsch MM300; Haan, Germany). Genomic DNA was isolated with the Invisorb Plant DNA 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 by fluorescence assays (Quant-IT kit; Invitrogen, Carlsbad, CA, USA).

Genotyping assay

The SNP genotyping assay was performed on an Illumina® Infinium iSelect Custom Genotyping Array (Illumina Inc., San Diego, CA, USA), according to the standard manufacturer's protocol, using 200 ng of genomic DNA per sample. All steps were performed on a Tecan Genesis and Tecan Evo2 150 LIMS ready liquid handler (Tecan, Männedorf, Switzerland) with Illumina® IAC Robot Control software. Fluorescence intensities were read with Illumina® iScan Control Software (ICS). A total of 7913 SNPs were genotyped on 1065 individuals at the French National Genotyping Centre (CNG, Evry, France).

SNP calling was performed with Genome Studio v2010.3 software (Illumina), with a GenCall score cut-off of 0.15, in accordance with Illumina recommendations. Genotyped SNPs were called five times, independently: once for each of the four mapping pedigrees and once for the overall sample of oak trees collected from natural populations. These independent analyses made it possible to check the consistency of Mendelian segregation between parents and offspring in mapping pedigrees. The scatter plots of all genotyped SNPs were inspected, to ensure that the data were of sufficiently high quality. Scatter plots displaying more than three clusters, unclear

© 2015 John Wiley & Sons Ltd

cluster delineation or inconsistent Mendelian segregation in mapping pedigrees were discarded (see examples in Figure S2, Supporting information). Scatter plots displaying compression, which has been described by Hyten *et al.* (2008) as a shift of clusters on the normalized theta *x*-axis, probably due to the presence of target locus paralogues, were either discarded or manually determined depending on cluster position and Mendelian segregation consistency (see examples in Figure S2, Supporting information). We considered homozygous clusters with theta values outside the [0, 0.2] or [0.8, 1] range and heterozygous clusters with theta values outside the [0.4, 0.6] range to be compressed.

Validation of sequencing and genotyping data

We refer in the following to three different groups of genotypes that were available for the six parents of the controlled crosses in the three different techniques used in this study to assess SNPs: 'Sanger genotypes' for *in vitro* SNPs from Sanger sequences, '454 genotypes' for putative *in silico* SNPs from Roche 454 reads and 'Illumina genotypes' from the Illumina assay which could be compared *a posteriori* to the corresponding genotypes from initial *in vitro* and *in silico* SNPs. For each combination of technique and bi-allelic SNP position, three different genotypes (two homozygotes and one heterozygote) can potentially be observed. We compared Sanger, Illumina and 454 genotypes that corresponded to the same *in vitro* or putative *in silico* SNP position to detect inconsistencies reflecting either sequencing or genotyping errors.

Sanger genotypes for *in vitro* SNPs were considered to be mostly reliable (*i.e.* with very low base call error rates), given the quality procedure implemented in the SeqQual pipeline, which generally masked base calls with Phred scores below 30 (*i.e.* error rates below 1/1000) or located within low-quality flanking sequences (Phred score <30). Moreover, sequence quality was double checked, for both the forward and reverse sequences, for each fragment in each sample. By contrast, 454 genotypes for putative *in silico* SNPs could potentially be incorrect more often, given the higher error rate of this technique than that of the Sanger method (Shendure & Ji 2008; Gilles *et al.* 2011), and the low coverage achieved with the 454 technique for some loci. Indeed, if few reads are available for a true heterozygote, the probability of sampling only one allele (and thus detecting a homozygote) is high. Conversely, in situations in which a large number of reads are available for a SNP position and the number of reads for the second allele is very low, the second allele may potentially correspond to a sequencing error. Thus, defining higher quality 454 homozygote and heterozygote genotypes was needed. The criteria initially used for the selection of putative *in silico* SNPs (depth > 6, minor allele appearing at least three times, MAF > 20%) were a compromise between preventing the targeting of sequencing errors rather than real polymorphisms and the rather low coverage observed overall at the individual level, so they were applied to the reads for the six parents in bulk, rather than for each parental individual. We thus applied several filters *a posteriori*, to improve the determination of correct 454 genotypes at the individual level and assess their impact on the number of inconsistencies between 454 and Illumina genotypes.

The first filter (filter A) involved the discarding of all the 454 homozygote genotypes with a depth lower than seven. The probability of incorrectly detecting a homozygote rather than a heterozygote at a given SNP, which corresponds to sampling the same allele seven times, thus decreased to less than 1% ($0.5^7 = 0.8\%$).

The second filter (filter B) on 454 reads was based on a probabilistic approach implemented by Brousseau *et al.* (2014). When applied to our data, this approach involved the discarding of a minor allele if its associated depth was significantly lower than that of the major allele, suggesting that it might be due to a sequencing error. When sequencing a heterozygote, the probability of observing the second allele no more than t times in n reads is given by $P(X \leq t) = \Sigma_{i=0}^{t} \frac{i!}{n! \times (n-i)!} 0.5^n$. All the heterozygous genotypes with a configuration of t (number of observations for the second allele) such that $P(X \leq t) \leq 0.01$ were considered to be false heterozygotes. This filter was ineffective for depths of fewer than 11 reads and dismissed the minor allele when it appeared only once in 11–13 reads, twice or less in 14–16 reads, less than four times in 17 or 18 reads *etc.* Finally, as filter B was ineffective for lower detection depths, we applied filter C, which involved discarding the data for depths <11, in addition to filter B.

Transferability of SNPs across more distant taxa

The transferability of 277 SNPs that have been mapped in the P1 pedigree (see Table S3, Supporting information) was additionally tested in nine other oak species included in the phylogenetic study of Hubert *et al.* (2014) with the Sequenom iPLEX technology (Jurinke *et al.* 2002). These species represent four major groups of the *Quercus* genus. From six to 12 individuals from natural populations of *Q. cerris*, *Q. coccifera*, *Q. coccinea*, *Q. ellipsoidalis*, *Q. ilex*, *Q. macrocarpa*, *Q. suber* and *Q. velutina*, and 10 individuals from a mapping population of *Q. alba* were genotyped (Table S1, Supporting information).

© 2015 John Wiley & Sons Ltd

Results

SNP assay genotyping statistics

From the initial set of 7913 *in vitro* or putative *in silico* SNPs used for the Illumina ADT, 903 (11.4%) did not pass Illumina production quality control due to weak or ambiguous signals and were discarded from subsequent analyses. In mapping populations, visual inspection of all the scatter plots (according to the procedure described in the methods) resulted in the validation of 6363 genotyped SNPs (80.4% of the total number of SNPs on the chip, in Table 1). Significantly more loci corresponding to *in vitro* SNPs than to putative *in silico* SNPs could be retained after visual inspection of the Illumina clusters (+4.8% in Table 1, with χ^2-test *P*-value of 5.2×10^{-5}). The proportion of genotyped SNPs discarded after visual inspection varied from 26% to 30% on a single-cross-basis, and 25% to 31% were monomorphic (Table 2). Finally, 5726 genotyped SNPs (72.4%) were polymorphic in at least one mapping pedigree. The reproducibility of the Infinium assay results, estimated with the use of 16 positive controls, was 99.998%, based on all segregating SNPs (92 106 data points were used).

In natural populations, 4334 SNPs were scored after visual inspection (Table 1). The success rate was significantly lower than that in the mapping populations (−25.6% with χ^2-test *P*-value $<2.2 \times 10^{-16}$). Genotyping success was higher for *in vitro* than for *in silico* chosen SNPs (+9.2% with a χ^2-test *P*-value of 4.3×10^{-10}, Table 1). There were more than 1.7 times more monomorphic SNPs in *Q. pubescens* and *Q. pyrenaica* than in *Q. petraea* and *Q. robur* (Table 3). Similarly, MAFs were generally lower for *Q. pubescens* and *Q. pyrenaica* than in the two other species (Fig. 1). Around 75% of the successfully genotyped SNPs were polymorphic in all four species (Fig. 2). *Q. petraea* and *Q. robur* had the highest proportion of shared polymorphic SNPs (91.1%), and *Q. pyrenaica* and *Q. pubescens* had the lowest proportion of shared polymorphic SNPs (80.4%). The percentage of

Table 1 Genotyping success for SNPs in the mapping pedigrees (including *Q. petraea* and *Q. robur*) and in natural populations (including all 4 species) as indicated by absolute counts of SNPs and percentages (in brackets) with respect to the total number of SNPs

SNP category	Total number of SNPs	Illumina quality control	Visual control in mapping populations	Visual control in natural populations
In vitro	1371	1207 (88.0)	1157 (84.4)	856 (62.4)
In silico	6542	5803 (88.7)	5206 (79.6)	3478 (53.2)
Total	7913	7010 (88.6)	6363 (80.4)	4334 (54.8)

Table 2 Breakdown of the 7913 SNPs in four full-sib families. Absolute counts of SNPs and percentages (in brackets) relative to the total number of SNPs

SNP categories	P1	P2	P3	P4	4 crosses
NS	2386 (30.2)	2059 (26.0)	2026 (25.6)	2273 (28.7)	1550 (19.6)
M	2012 (25.4)	2120 (26.8)	2488 (31.4)	2222 (28.1)	637 (8.0)
P	3515 (44.4)	3734 (47.2)	3399 (43.0)	3418 (43.2)	5726 (72.4)

NS, nonscorable SNP (includes the SNPs discarded after quality control of Illumina or visual inspection); M, monomorphic SNP; P, polymorphic SNP.

Table 3 Breakdown of the 7913 SNPs in four white oak species. Absolute counts of SNPs and percentages (in brackets) with respect to the total number of SNPs

SNP category	*Quercus petraea*	*Quercus robur*	*Quercus pubescens*	*Quercus pyrenaica*	All species
NS	3579 (45.2)	3579 (45.2)	3579 (45.2)	3579 (45.2)	3579 (45.2)
M	426 (5.4)	419 (5.3)	779 (9.8)	717 (9.1)	229 (2.9)
P	3908 (49.4)	3915 (49.5)	3555 (44.9)	3617 (45.7)	4105 (51.9)

NS, nonscorable SNP (includes the SNPs discarded after quality control of Illumina or visual inspection); M, monomorphic SNP; P, polymorphic SNP.

© 2015 John Wiley & Sons Ltd

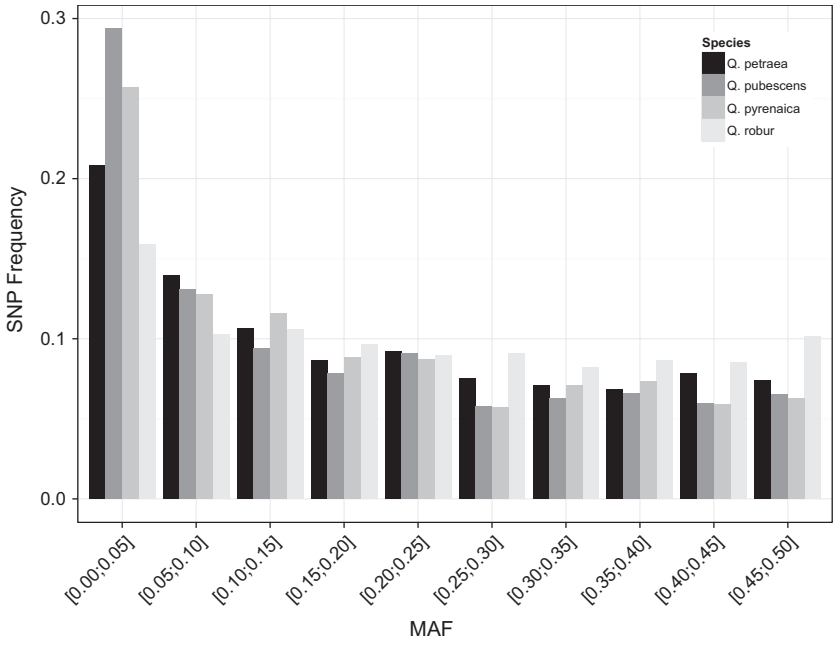

Fig. 1 Distribution of MAF for 4105 polymorphic SNPs in four white oak species.

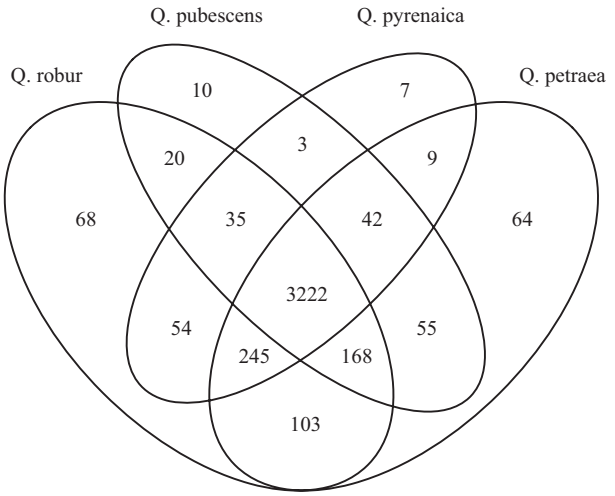

Fig. 2 Venn diagram showing the number of shared polymorphisms in the four oak species.

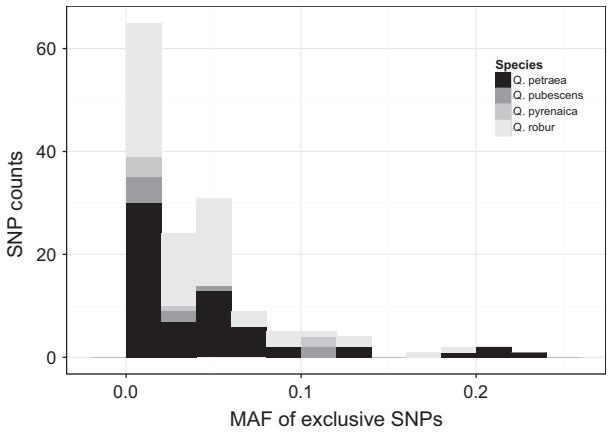

Fig. 3 Distribution of MAF for exclusive SNPs in each species

private polymorphic SNPs varied between 0.2% and 1.7%, depending on the species (Fig. 2), and around 75% of them had very low MAFs (below 5% in Fig. 3).

We compared Sanger, 454 and Illumina genotype calls for the six parents of the mapping populations, to check the consistency of initial SNP discovery and Illumina genotyping results. For *in vitro* SNPs, 93.5% of the 5814 data point pairs for which both Sanger and Illumina results were available were consistent (Table 4). For *in silico* SNPs, the consistency rate was significantly lower (81.8%, χ^2-test P-value $<2.2 \times 10^{-16}$, Table 5). We therefore investigated the potential sources of errors leading to these inconsistencies.

Error detection for Sanger and/or Illumina genotypes

The most common type of inconsistency was a heterozygote by Sanger sequencing, but a homozygote by Illumina methods (205 data point pairs, Table 4). Visual checking of the Sanger chromatograms confirmed that almost all cases (202) were correctly called by the automatic SeqQual pipeline. These data points had an overall Phred score above 40 (error probability below 1/10 000), and using Polyphred tags, a high genotype score (>90) and an overall threshold score for the SNP position (>60 and >90 respectively, see class 1 in Table 6, Nickerson *et al.* 1997). Three Sanger genotypes were incorrectly interpreted as having a second allele by the automatic call, due to unusually high levels of background (class 3 in Table 6).

© 2015 John Wiley & Sons Ltd

Table 4 Comparison of Sanger and Illumina genotypes for the six parents of the mapping populations (*in vitro* SNPs). Percentages are expressed relative to the total number homozygotes or heterozygotes called by Sanger sequencing

Sanger genotype	Illumina genotype	Number of data points pairs	%
AA	AA	3760	95.6
AA	AB	157	4.0
AA	BB	15	0.4
Total no. of Sanger homozygotes		3932	100
AB	AB	1677	89.1
AB	AA	205	10.9
Total no. of Sanger heterozygotes		1882	100
Total no. of data point pairs		5814	
Overall consistency rate			93.5

There were 157 inconsistencies involving SNPs identified as homozygous by Sanger sequencing, but heterozygous by Illumina methods (Table 4). Overall, 108 (~69%) of these Sanger calls were validated as correct, because the chromatograms displayed no background noise and belonged to class 1 (in Table 6). Others were probably true heterozygotes (automatic Sanger calls giving an incorrect homozygote), because a second allele was clearly visible with a second peak on the chromatogram, but the peak height for this allele was too low for automatic detection (class 2 in Table 6).

Inconsistencies in which Sanger method identified a homozygote of one type and Illumina method identified a homozygote of the other type were rare (only 15 data point pairs). In 14 of these cases, the Sanger base call was correct (see criteria used above for class 1). In the remaining case, the Sanger chromatogram suggested that the SNP was probably heterozygous, the second allele not being detected due to a weak signal (Phred score ~ 30 and Polyphred genotype score <90, class 3 in Table 6). However, correcting the Sanger genotype into a heterozygote did not remove the inconsistency, with a probable incorrect homozygous call with the Illumina technique.

Thus, visually checking of the Sanger chromatograms identified a large majority (86%) of incorrect Illumina genotypes among the 377 Sanger–Illumina inconsistencies. Considering all 5814 data points involved in the comparison, 5.6% of Illumina genotypes were therefore not correct while less than 1% of Sanger genotypes were not correct. Moreover, across the 1158 SNPs concerned, one to five (of six) incorrect Illumina genotype calls were detected for 19.52% of them, which is nearly five times higher than the proportion of SNPs (4.15%) showing one or two incorrect Sanger genotype calls.

Error detection for 454 and/or Illumina genotypes

Comparing 454 and Illumina genotypes, ~20% of inconsistencies concerned SNPs called as heterozygous by the 454 method and as homozygotes by the Illumina method (Table 5). Then, ~15% of inconsistencies involved SNPs called as homozygotes by the 454 method and as heterozygotes by the Illumina method, and the last type of inconsistency concerned SNPs identified as one type of homozygote by the 454 method but as the alternative homozygote by the Illumina technique (1.4% of 454 homozygotes).

Table 5 Comparison of 454 genotypes for putative *in silico* SNPs and corresponding Illumina genotypes for the six parents of the mapping populations. Percentages are expressed relative to the total number of 454 homozygotes or 454 heterozygotes. The various filters used to improve the determination of 454 genotypes are explained in the Methods section

		Observed data		Filter A		Filter B		Filters B+C	
454 genotype	Illumina genotype	Number of data point pairs	%	Number of data point pairs	%	Number of data point pairs	%	Number of data point pairs	%
AA	AA	7 878	83.7	1 120	96.4	8 012	83.6	499	92.4
AA	AB	1 406	14.9	31	2.7	1 436	15.0	36	6.7
AA	BB	130	1.4	11	0.9	131	1.4	5	0.9
Total no. of 454 homozygotes		9 414	100	1 162	100	9 579	100	540	100
AB	AB	6 496	79.6	6 496	79.6	6 466	80.8	479	89.0
AB	AA	1 669	20.4	1 669	20.4	1 534	19.2	59	11.0
Total no. of 454 heterozygotes		8 165	100	8 165	100	8 000	100	538	100
Total no. of data point pairs		17 579		9 327		17 579		1 078	
Overall consistency rate			81.8		81.7		82.4		90.7

© 2015 John Wiley & Sons Ltd

Table 6 Distribution of Sanger–Illumina inconsistencies between different classes, according to the quality of Sanger chromatograms

	Overall Sanger chromatogram quality (Phred score)	Data point Sanger quality (Phred score)	Polyphred threshold overall score	Polyphred genotype site score	Number of data point pairs	%
Class 1: High-quality Sanger, correct call	>40	>40*	>60	>90	324	86.0
Class 2: High-quality Sanger but differential amplification of second strand, incorrect call	>40	~ 30	>60	<90	50	13.2
Class 3: Lower quality Sanger (score just above thresholds), incorrect call	~ 30	~ 30*	>60	<90	3	0.8

Phred score: >40 indicates an error rate below 1/10 000, ~30 indicates an error rate ~1/1000, Polyphred threshold overall score sets a lower limit below which the position is not considered as a SNP, it accounts for individual genotype site score at each double-stranded Sanger sequence, genotype site score is the Polyphred software quality score for the peak pattern.
*Except for heterozygotes, which have lower scores.

Table 7 Breakdown of the 277 SNPs in nine oak species. Absolute counts of SNPs and percentages (in brackets) with respect to the total number of SNPs

	Mapping population	Natural populations								
Taxonomic group*	*Q. alba* Quercus	*Q. macrocarpa* Quercus	*Q. cerris* Cerris	*Q. suber* Cerris	*Q. coccifera* Ilex	*Q. ilex* Ilex	*Q. coccinea* Lobatae	*Q. ellipsoidalis* Lobatae	*Q. velutina* Lobatae	Overall
Nonscorable SNP	24 (8.7)	20 (7.2)	39 (14.1)	35 (12.6)	27 (9.7)	30 (10.8)	36 (13.0)	35 (12.6)	33 (11.9)	14 (5.1)
Monomorphic SNP	229 (82.7)	197 (71.1)	227 (81.9)	228 (82.3)	232 (83.8)	231 (83.4)	226 (81.6)	225 (81.2)	222 (80.1)	149 (53.8)
Polymorphic SNP	24 (8.7)	60 (21.7)	11 (4.0)	14 (5.1)	18 (6.5)	16 (5.8)	15 (5.4)	17 (6.1)	22 (7.9)	114 (41.2)

*According to Hubert *et al*. (2014).

Filter A discarded 87.7% of the homozygous SNP data points but is not applicable for *in silico* SNP detection and just illustrates the fact that not enough depth (below seven reads) prevents homozygote genotype calls with enough certainty (see Methods). Filter B allowed to correct 165 heterozygotes for which the minor allele was a likely error, into homozygous genotypes (Table 5): 134 of these genotypes were then consistent with the Illumina genotypes, whereas 31 genotypes remained inconsistent. This filter increased the consistency rate for 454 heterozygotes by 1.2% compared with the observed data. Finally, we applied filter C in addition to filter B (Table 5, filter B + C), and this gave a consistency rate for 454 heterozygotes similar to that for *in vitro* SNPs (89.0% compared with 89.1%), and the highest overall consistency rate across observed data and filters. The use of filter B+C was stringent and entailed a high degree of data loss (~94% of the data point pairs), but it allowed to assume that the subset of 566 high-quality retained SNPs (hereafter referred to as filtered *in silico* SNPs) corresponded *a priori* to 1078 correct 454 genotype calls. Therefore, we observed that 9.3% of Illumina genotype calls were probably incorrect, and this corresponded to 15.2% of the SNPs showing at least one (up to five) incorrect Illumina calls. These high-quality filtered SNPs are used below for understanding the impact of cluster compression to potential Illumina genotyping errors.

Presence of compression in Illumina clusters

For 39 data point pairs where Illumina genotypes did not match validated Sanger or filtered 454 genotypes, we observed that the incorrect assignment of genotypes to the different Illumina clusters was due to compression.

© 2015 John Wiley & Sons Ltd

This occurred when only two clusters were present, a first shifted cluster considered to correspond to a heterozygote rather than a homozygote and a second cluster considered to correspond to a homozygote rather than a heterozygote (see example on Figure S3, Supporting information). The detection of such errors was confirmed using data from other mapping pedigrees, when three clusters were present.

Overall in mapping populations, compression was observed for around 25% of validated Sanger or filtered 454 SNPs (*a priori* true). It was more frequent among SNPs for which at least one data point was *a priori* false with the Illumina technique than among SNPs displaying no inconsistency at all (+18.3% with a χ^2-test *P*-value of 3.06×10^{-11}).

Preliminary diversity comparisons and transfer across different species

The polymorphism of 277 SNPs that have been mapped in the P1 pedigree was tested in nine different oak species from the *Quercus* (white oaks), *Ilex*, *Cerris* and *Lobatae* (red oaks) groups (Hubert *et al.* 2014). Within each species, more than 85% of the SNPs were successfully genotyped, but only 4% to 22% were polymorphic depending on the species (Table 7, see details on the allele type and polymorphism at each of the 277 SNPs across species in Supplemental Table 3). Moreover, when considering the nine species overall, an additional 20% of the SNPs were polymorphic due to fixed observed differences in at least one species compared with the others.

Discussion

Exploring existing genomic resources to discover large numbers of SNPs is particularly useful for population genetics studies and in breeding, conservation or management applications. In this study, we aimed to detect a large number of SNPs in oaks and to validate them by investigating their segregation and variation in different species. Of the 7913 initially chosen SNPs from Sanger or 454 data, we successfully genotyped 80.4% and 54.8% of them, in mapping and natural oak populations, respectively. These success rates correspond to conversion rates (considering only polymorphic SNPs) of 72.4% and 51.9%. A large range of conversion rate values has been obtained with the same technology in other species: from 13.1% to 24.4% in maritime pine (Chancerel *et al.* 2013), 42.9% in walnut (You *et al.* 2012), between 55.8% and 67.6% in spruce (Pavy *et al.* 2013), from 32.5% to 70.6% in apple (Antanaviciute *et al.* 2012; Chagné *et al.* 2012) and 91.0% in soya bean (Song *et al.* 2013). Among the factors affecting the variation of success or conversion rates between studies, we have shown here the importance of all the technical criteria for choosing the SNPs (quality of the sequences, sequencing depth, MAF in the discovery panel, level of polymorphism in the flanking sequences, GenomeStudio parameters for SNP calling such as GenCall or GenTrain score cut-off or criteria used for additional visual inspection, if any). Success rates can also depend on the type of genomic resources or sequences from which the SNP originates (either *in silico* or *in vitro* SNPs) and the sequencing discovery panel that could be more or less representative of the genetic diversity of the targeted species or populations. We discuss below our results in the light of the different factors at stake and in comparison with other studies.

Experimental error rates, causes and recommendations

As in earlier studies comparing *in silico* and *in vitro* SNPs (Pavy *et al.* 2008; Lepoittevin *et al.* 2010; Chancerel *et al.* 2011), overall genotyping success rates were lower for *in silico* than for *in vitro* SNPs (−4.8% in mapping pedigrees, *P*-value $<5.2 \times 10^{-5}$, and −9.2% in natural populations, *P*-value $<4.3 \times 10^{-10}$). Genotyping success depends on the quality of SNP flanking sequences which needs to be high enough for designing primer sequences. This has been identified in previous studies as a problem likely to be less marked at higher coverage (Wang *et al.* 2008; Pavy *et al.* 2013). Using our results, we noticed that the genotyping success rate was independent to 454 data sequencing depth (Fig. 4), given the minimum depth of six reads used at the array design step for *in silico* SNP discovery and primer design. This tends to confirm that

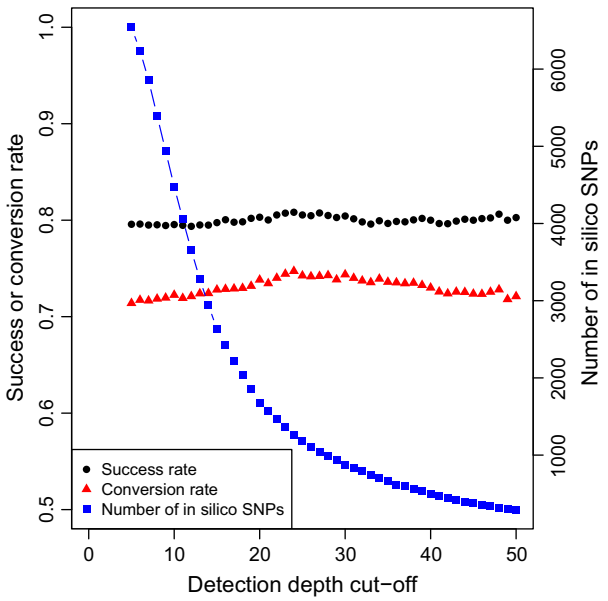

Fig. 4 Success rate, conversion rate and number of *in silico* SNPs as a function of detection depth cut-off.

© 2015 John Wiley & Sons Ltd

the overall quality of the consensus sequence on which the primers were designed was good enough with a minimum of six reads. However, as pointed out in Wang *et al.* (2008), the presence of undetected introns in the middle of primer binding sites may be a major cause of genotyping failure for *in silico* SNPs, because primer design is based on mRNA sequences.

We compared Sanger, 454 and Illumina genotypes for the six parents of the mapping pedigrees and examined the consistency across methods and possible causes for different genotype calls. Observed consistency rates for Sanger and Illumina techniques (93.5%) were much higher than those for 454 and Illumina techniques (between 81.8% without any filter to 90.7% with the most stringent filter), as expected given the higher error rate of 454 than of Sanger sequencing (Shendure & Ji 2008). Importantly, using comparisons of validated Sanger data and filtered high-quality 454 data across more than 1000 SNPs, our results revealed a potential error rate for Illumina genotypes of more than 5% affecting between 15% and 20% of the corresponding SNPs (including some retained compressed clusters).

As validated Sanger or filtered 454 genotypes were considered to be *a priori* true, we refer to Illumina genotypes presenting inconsistencies as errors in the rest of the discussion, and we examine the possible causes of these errors. For incorrect Illumina heterozygotes (versus homozygotes from sequencing methods) in 4% of all comparisons (see Tables 4 and 5), one possible cause is the binding of Illumina probes to different paralogous regions in the oak genome, generating mixed allele signals. Compression has been linked to the presence of paralogues or homeologues in several studies (Hyten *et al.* 2008; Akhunov *et al.* 2009; Sandve *et al.* 2010). In our study, compression rates were also significantly higher for SNPs displaying at least one potential Illumina error than for SNPs showing consistent genotype calls (+23% and +10% when comparing to validated *in vitro* SNPs or filtered *in silico* SNPs, respectively), which is consistent with this hypothesis. Compression also accounted for some clustering errors which could be detected in mapping pedigrees (Fig. S3, Supporting information). We therefore recommend discarding data if only two compressed clusters are available for a SNP in a mapping pedigree, and more generally to be very cautious with compressed cluster patterns if no other validation method is available.

Another important type of error was the identification of Illumina homozygous genotypes while the sequencing methods showed *a priori* true heterozygotes (almost 11% of all comparisons, see Tables 4 and 5). The Infinium probe may have failed to bind to one of the alleles, preventing the enzymatic base extension and resulting in a null allele. A different or paralogous region of the genome could also have been targeted by the designed primers, and a combination of both hypotheses is also possible. Even though the availability of pedigree data in a study can help ruling out some hypotheses by studying Mendelian inconsistencies, a combination of causes can always go undetected.

Excluding all compression cases, potential Illumina errors compared with validated sequencing data were still observed for more than 10% of the SNPs. Mammadov *et al.* (2012) showed that only 162 of 618 SNPs (26.2%) successfully genotyped in four maize genotyping assays (GoldenGate, Infinium, TaqMan and KASPar) displayed consistent clustering patterns across techniques. The major attrition came from the Illumina Infinium assay, the same technique used in our study. This suggests as in our case that in nonmodel species, the error rate for Illumina genotyping is potentially substantial (more than 10% of SNPs showing genotyping errors). Additional information from the genome sequence when available for *Quercus* genus will probably help to better design primer sequences and lower the number of errors.

For scientists interested in developing SNP arrays in nonmodel species, we recommend the use of next-generation sequencing methods which save both time and money for the discovery panel. Even if the Sanger sequencing in our study was used as a reference method to estimate error rates, our results show that high-quality *in silico* SNPs can be better identified using appropriate filters applied on reads which are individually tagged. The probabilistic approach used in one of the filters proposed is necessary to avoid the detection of false-positive SNPs and can also be applied at the population level in case of untagged libraries, taking into account the number of gametes sequenced (Brousseau *et al.* 2014). We also recommend a minimal sequencing depth target of 11 reads that may correspond to a much higher mean sequencing coverage depending on the technology used, the genome length and the number of gametes sequenced, to avoid the loss of data during the filtering steps. It remains that the experimental error rates observed here are relatively high and should be accounted for when making inferences from SNP data with the Illumina Infinium technique. This suggests that methods of genotyping by sequencing, which allow to better validate the genomic regions targeted, are likely to be of higher value for nonmodel species for different aims and applications.

Impact of discovery panel on genotyping success rates

Our study clearly illustrates that the presence of samples from the targeted populations in the SNP discovery panel greatly increases the array success rate. SNPs were

© 2015 John Wiley & Sons Ltd

chosen initially from sequences from the six parents of the pedigrees P1, P2, P3 and P4, we therefore had full knowledge of the polymorphism around the targeted SNPs in these mapping populations. This ensured both an optimal primer design and an efficient genotyping of progenies with known patterns of marker segregation between parents and offspring for the definition of clusters (see above for the validation strategy) and is consistent with the higher success rate in mapping *versus* natural populations (+25.6%, see Table 1). Indeed in natural populations, we expect that undetected polymorphisms present at binding sites would cause the loss of data due to inefficient primer design. Differences in levels of genetic diversity between discovery panel and natural populations can be due to both differences in sampling and geographic origins structure. Despite the relatively close geographic proximity of sampled natural populations to pedigree parents (Table S1 and Fig. S1, Supporting information), the larger sample of individuals assessed (283 compared to 6 in pedigrees) led to a drop of more than 20% of genotyping success and argues for larger sample sizes in discovery panels in nonmodel species (Lepoittevin *et al.* 2010). Remarkably, genotyped SNPs in the studied natural populations were scorable either across all four species or in none of them, confirming their close genetic proximity, as illustrated also by their shared proportions of polymorphisms (see below), and interfertility based on controlled crosses experiments from material of similar local origin (e.g. Lepais *et al.* 2013).

The issue of transferability of the markers developed here to more distant natural populations within the same species remains to be addressed more thoroughly. In *Q. petraea* and *Q. robur*, a low molecular genetic structure among populations across a larger species range was observed on average but with a smaller set of molecular markers, and despite hot spots of localized higher differentiation levels (Mariette *et al.* 2002; Scotti-Saintagne *et al.* 2004). Given the larger genome exploration conducted in our study, we could expect a higher proportion of those islands of differentiation. This could lower overall success rates further in more distant geographic populations but also provide interesting insights into the dissection of the divergence across this species complex in future studies.

SNP across the different oak species studied and transfer to more distant taxa

We found that *Q. robur* and *Q. petraea* had a higher frequency of polymorphic SNPs than *Q. pyrenaica* and *Q. pubescens*. This could be due to an ascertainment bias as SNP discovery was conducted in *Q. petraea* and *Q. robur* and not in the two other species. Also, earlier comparative assessments of diversity based on microsatellites showed that diversity levels were similar in *Q. petraea* and *Q. pyrenaica* (Valbuena-Carabana *et al.* 2005) or in *Q. robur*, *Q. petraea* and *Q. pubescens* (Curtu *et al.* 2007; Höltken *et al.* 2012). Similarly, when comparing differences in patterns of shared polymorphisms across polymorphic SNPs in all four species, *Q. petraea* and *Q. robur* had the largest number of shared polymorphisms (91.1%), compared to each of those two species and *Q. pubescens* or *Q. pyrenaica* (from 83.9% to 86.6%). This proportion was much larger than in spruce species (Pavy *et al.* 2013; values of 2–65%), maybe because the sampling of spruce species extended over a larger phylogenetic range. They also came from very distant sites and were not all interfertile, compared to oak species being interfertile and preferentially sampled from the same mixed natural forest stands. Both diversity and sharing patterns could be partly due to ascertainment bias. However, given the already mentioned genetic proximity of the four species, the larger set of markers developed here could also have revealed patterns not previously detected. Therefore, more studies and analyses would be needed on much larger samples of individuals to conclude on the relative magnitudes of diversity and divergence across these four species of European white oaks. The high levels of shared diversity among these species preclude their straightforward distinction, but a recent study on over 400 gametes in both *Q. petraea* and *Q. robur* illustrates the usefulness of this type of SNP resource for hybridization studies and evolutionary inferences in the species history (Guichoux *et al.* 2013).

Finally, we could assess the transfer of a subset of 277 SNPs in more distant taxa. As expected, the best transferability results were obtained with *Q. macrocarpa* and *Q. alba* that belong to the same taxonomic group (Quercus genus, white oaks) as the four species studied here, despite only two individuals (four gametes) for the *Q. alba* pedigree. This suggests a general interest of these markers for more than one hundred species of white oaks distributed across Europe, North Africa, America and Asia (Hubert *et al.* 2014). Besides, we observed fixed allele differences among several species pairs, suggesting their potential use for phylogenetic analyses. Together with the results of previous studies on the transferability of microsatellites across oak species (Steinkellner *et al.* 1997) and even closely related genera (Castanea and Quercus, Barreneche *et al.* 2004; Bodénès *et al.* 2012), the SNP resource developed in this study provides a highly valuable tool for genetic and genomic investigations in oaks.

Acknowledgements

We thank all those who helped in the preparation of the list of candidate genes from the oak allelic resequencing project for

© 2015 John Wiley & Sons Ltd

in vitro SNP detection (P. Abadie, C. Burban, T. Decourcelle, J. Derory, M.-L. Desprez-Loustau, G. Le Provost, C. Robin and J-M Frigerio). We also thank all people providing DNA samples of *Quercus* from a large set of species for the Sequenom transferability experiment: Giuseppe Vendramin (CNR Firenze, Italy) for *Q. ilex*, *Q. cerris*, *Q. suber* and *Q. coccifera* DNA samples, Andrew Hipp and Marlene Hahn (The Morton Arboretum, Illinois, USA) for *Q. macrocarpa*, *Q. velutina* and *Q. ellipsoidalis* DNA samples, Nicole Zembower and John Carlson (Pennstate University, Pennsylvania, USA) for *Q. alba* DNA samples, Jeanne Romero-Severson (University of Notre Dame, Indiana, USA) and Mark V. Coggeshall (University of Missouri, Columbia, USA) for *Q. rubra* samples (data excluded because of low Genecall scores). The sequencing of oak amplicons and ESTs and genotyping were funded by the European Union (EVOLTREE Network of Excellence, EU FP7 project 016322). EC was supported by the EVOLTREE Project. IL was supported by the ANR GENOAK Project (11-BSV6-009-01), and LV was supported by the GENOAK, FORESTTRAC and FORGER EU FP7 Projects. We would like to thank three anonymous reviewers for their insightful comments, as these led us to a significant improvement of the manuscript.

References

Abadie P, Roussel G, Dencausse B *et al.* (2012) Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *Journal of Evolutionary Biology*, **25**, 157–173.

Abbott A (2012) The genomes of the giants: a walk through the forest of tree genomes. *Tree Genetics & Genomes*, **8**, 443–443.

Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theoretical and Applied Genetics*, **119**, 507–517.

Alberto FJ, Derory J, Boury C *et al.* (2013) Imprints of natural selection along environmental gradients in phenology-related genes of *Quercus petraea*. *Genetics*, **195**, 495–512.

Antanaviciute L, Fernández-Fernández F, Jansen J *et al.* (2012) Development of a dense SNP-based linkage map of an apple rootstock progeny using the *Malus* Infinium whole genome genotyping array. *BMC Genomics*, **13**, 203.

Barreneche T, Casasoli M, Russell K *et al.* (2004) Comparative mapping between *Quercus* and *Castanea* using simple-sequence repeats (SSRs). *Theoretical and Applied Genetics*, **108**, 558–566.

Bodénès C, Chancerel E, Gailing O *et al.* (2012) Comparative mapping in the Fagaceae and beyond with EST-SSRs. *BMC Plant Biology*, **12**, 153.

Brousseau L, Tinaut A, Duret C *et al.* (2014) High-throughput transcriptome sequencing and preliminary functional analysis in four neotropical tree species. *BMC Genomics*, **15**, 238.

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.

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

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 J-B, 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.

Curtu A, Gailing O, Leinemann L, Finkeldey R (2007) Genetic variation and differentiation within a natural community of five oak species (*Quercus* spp.). *Plant Biology*, **9**, 116–126.

Derory J, Scotti-Saintagne C, Bertocchi E *et al.* (2010) Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. *Heredity*, **104**, 438–448.

El Mujtar VA, Gallo LA, Lang T, Garnier-Gere P (2014) Development of genomic resources for *Nothofagus* species using next-generation sequencing data. *Molecular Ecology Resources*, **14**, 1281–1295.

Ersoz ES, Wright MH, González-Martínez SC, Langley CH, Neale DB (2010) Evolution of disease response genes in loblolly pine: insights from candidate genes. *PLoS ONE*, **5**, e14234.

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.

Gilles A, Meglecz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.

Govaerts R, Frodin DG (1998) World checklist and bibliography of Fagales. *Kew: Royal Botanic Gardens, Kew vii*, 407 p. ISBN: 1900347466.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. *Current Opinion in Plant Biology*, **12**, 148–156.

Guichoux E, Garnier-Gere P, Lagache L *et al.* (2013) Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, **22**, 450–462.

Höltken M, Buschbom J, Kätzel R (2012) Species integrity of *Quercus robur* L., *Q. petraea* (Matt.) Liebl. and *Q. pubescens* Willd. from the genetic point of view. *Allgemeine Forst und Jagdzeitung*, **183**, 100–110.

Homolka A, Schueler S, Burg K, Fluch S, Kremer A (2013) Insights into drought adaptation of two European oak species revealed by nucleotide diversity of candidate genes. *Tree Genetics & Genomes*, **9**, 1179–1192.

Hubert F, Grimm GW, Jousselin E *et al.* (2014) Multiple nuclear genes stabilize the phylogenetic backbone of the genus Quercus. *Systematics and Biodiversity*, **12**, 405–423.

Hyten D, Song Q, Choi I-Y *et al.* (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics*, **116**, 945–952.

Johnson PS, Shifley SR, Rogers R (2009) *The ecology and silviculture of oaks*. CABI publishing, New York, NY, USA.

Jurinke C, Van Den Boom D, Cantor CR, Köster H (2002) Automated genotyping using the DNA MassArray™ technology. In: *PCR Mutation Detection Protocols* (eds Theophilus BDM & Rapley R), pp. 179–192. Humana Press, Totowa, New Jersey.

Kremer A, Dupouey JL, Deans JD *et al.* (2002) Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Annals of Forest Science*, **59**, 777–787.

Kremer A, Abbott A, Carlson J *et al.* (2012) Genomics of Fagaceae. *Tree Genetics & Genomes*, **8**, 583–610.

Le Provost G, Sulmon C, Frigerio JM *et al.* (2012) Role of waterlogging-responsive genes in shaping interspecific differentiation between two sympatric oak species. *Tree Physiology*, **32**, 119–134.

Lepais O, Petit R, Guichoux E *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology*, **18**, 2228–2242.

Lepais O, Roussel G, Hubert F, Kremer A, Gerber S (2013) Strength and variability of postmating reproductive isolating barriers between four European white oak species. *Tree Genetics & Genomes*, **9**, 841–853.

Lepoittevin C, Frigerio J-M, 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.

Mammadov J, Chen W, Mingus J, Thompson S, Kumpatla S (2012) Development of versatile gene-based SNP assays in maize (*Zea mays* L.). *Molecular Breeding*, **29**, 779–790.

Mariette S, Cottrell J, Csaikl UM *et al.* (2002) Comparison of levels of genetic diversity detected with AFLP and microsatellite markers

© 2015 John Wiley & Sons Ltd

within and among mixed *Q. petraea* (Matt.) Liebl. and *Q. robur* L. stands. *Silvae Genetica*, **51**, 72–79.
Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews: Genetics*, **12**, 111–122.
Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.
Nixon K (2006) Global and neotropical distribution and diversity of oak (genus *Quercus*) and oak forests. In: *Ecology and Conservation of Neotropical Montane Oak Forests* (ed. Kappelle M), pp. 3–13. Springer, Berlin.
Pavy N, Paule C, Parsons L *et al.* (2005) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics*, **6**, 19.
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**, 17.
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.
Petit RJ, Carlson J, Curtu AL *et al.* (2013) Fagaceae trees as models to integrate ecology, evolution and genomics. *New Phytologist*, **197**, 369–371.
Rasheed-Depardieu C, Parent C, Crèvecoeur M *et al.* (2012) Identification and expression of nine oak aquaporin genes in the primary root axis of two oak species, *Quercus petraea* and *Quercus robur*. *PLoS ONE*, **7**, e51838.
Sandve SR, Rudi H, Dørum G, Berg PR, Rognli OA (2010) High-throughput genotyping of unknown genomic terrain in complex plant genomes: lessons from a case study. *Molecular Breeding*, **26**, 711–718.
Scotti-Saintagne C, Mariette S, Porth I *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615–1626.
Shen R, Fan J-B, Campbell D *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research*, **573**, 70–82.
Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.
Song Q, Hyten DL, Jia G *et al.* (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE*, **8**, e54985.
Steinkellner H, Lexer C, Turetschek E, Glössl J (1997) Conservation of (GA) n microsatellite loci between *Quercus* species. *Molecular Ecology*, **6**, 1189–1194.
Tarkka MT, Herrmann S, Wubet T *et al.* (2013) OakContigDF159.1, a reference library for studying differential gene expression in *Quercus robur* during controlled biotic interactions: use for quantitative transcriptomic profiling of oak roots in ectomycorrhizal symbiosis. *New Phytologist*, **199**, 529–540.
Ueno S, Le Provost G, Leger V *et al.* (2010) Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics*, **11**, 650.
Ueno S, Klopp C, Leplé JC *et al.* (2013) Transcriptional profiling of bud dormancy induction and release in oak by next-generation sequencing. *BMC Genomics*, **14**, 236.
Valbuena-Carabana M, González-Martínez S, Sork V *et al.* (2005) Gene flow and hybridisation in a mixed oak forest (*Quercus pyrenaica* Willd. and *Quercus petraea* (Matts.) Liebl.) in central Spain. *Heredity*, **95**, 457–465.
Vidalis A, Curtu A, Finkeldey R (2013) Novel SNP development and analysis at a NADP+-specific IDH enzyme gene in a four species mixed oak forest. *Plant Biology*, **15**, 126–137.
Vornam B, Gailing O, Derory J *et al.* (2011) Characterisation and natural variation of a dehydrin gene in *Quercus petraea* (Matt.) Liebl. *Plant Biology*, **13**, 881–887.
Wang S, Sha Z, Sonstegard TS *et al.* (2008) Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, **9**, 450.
You FM, Deal KR, Wang J *et al.* (2012) Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. *BMC Genomics*, **13**, 354.

C.L. performed bioinformatic analysis for *in silico* SNP detection and array design, with the help of P.G.G., I.L. and T.L.; P.G.G. and T.L. performed *in vitro* SNP detection and development of bioinformatic tools; E.C. extracted DNA and checked its quality; D.Z., A.B. and C.B. took part in genotyping; C.Bod, C.Bou, C.L., E.C., C.P. and L.V. did SNP scoring; L.V., C.Bod, C.P. and A.K. did genetic analysis; C.L. and P.G.G. did comparison of Sanger, 454 and Illumina genotypes and statistical analysis of consistency rates; F.E., I.L. and P.G.G. performed database compilation and analysis; C.Bod and C.Bou did Sequenom transferability experiment; C.L., P.G.G., C.Bod, L.V., A.K. and C.P. wrote the manuscript; and A.K., P.C. and C.L. were involved in overall coordination.

Data accessibility

Sanger consensus sequences of the amplicons for the 6 parental individuals of the mapping pedigrees are provided in Appendix S2 (Supporting information), and polymorphic *in silico* and *in vitro* SNPs are available from the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP). Accession nos are listed in Table S3 (Supporting information).

The 454 sequencing data are available from the short-read archive of the NCBI database (http://www.ncbi.nlm.nih.gov/sra) (SRA012448). The 454 alignments for the 5112 contigs from which *in silico* SNPs were extracted are available at Dryad Digital Repository doi:10.5061/dryad.fd862. The oak contigs, annotations and data mining results obtained with BioMart can be browsed at http://ngspipelines.toulouse.inra.fr:9024/ngspipelines/#!/NGSpipelines/Quercus%20robur%20-%20qrobur.

The particular scripts used in this study from the SeqQual pipeline and the *snp2Illumina* script are compiled and described in Table S2 (Supporting information). The *snp2Illumina* Perl script is available in Appendix S1 (Supporting information).

Supporting Information

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

Fig. S1 Distribution area of *Q. petraea* and *Q. robur*, and location of sampled populations.

Fig. S2 Examples of scatter plots for SNPs that were discarded or manually adjusted after visual inspection. A: Discarded SNP,

© 2015 John Wiley & Sons Ltd

more clusters than expected. B: Discarded SNP, the delineation between clusters is unclear. C: Discarded SNP, the AA homozygote cluster should not be observed because the parental individuals are AB and BB. A null allele may be present in the female parent. D: Discarded SNP, there is some compression and the homozygote and heterozygote clusters cannot be distinguished. Such SNPs should be discarded if only two clusters are available, even if the Mendelian segregation is consistent (see Fig. S3). E: Manually adjusted SNP. There is some compression, but the homozygote cluster BB is in the usual place.

Fig. S3 Cluster plot for *in vitro* SNP CL8450CT11856_03 + 04-625 in the P2 and P3 mapping pedigrees. When clustered alone, pedigree P3 shows a heterozygote cluster and a homozygote BB cluster to the right. When clustered together with pedigree P2, it shows a homozygote cluster AA and a heterozygote cluster to the right. Compression (here of the AA and AB clusters towards the BB cluster) may cause clustering errors when only two clusters are present. This type of error cannot be detected on the basis of Mendelian segregation patterns.

Table S1 List of the samples used for sequencing and genotyping.

Table S2 List of the programs used in this study.

Table S3 SNP statistics and NCBI ss accession nos.

Appendix S1 *snp2Illumina* perl program.

Appendix S2 Sanger consensus sequences of the amplicons (709 gene fragments corresponding to 676 gene regions) for the 6 parental individuals of the mapping pedigrees.

© 2015 John Wiley & Sons Ltd