First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill

S. PINOSIO,*† S. C. GONZÁLEZ-MARTÍNEZ,†‡ F. BAGNOLI,¶ F. CATTONARO,† D. GRIVET,‡ F. MARRONI,** Z. LORENZO,‡† M. VERD†† and G. G. VENDRAMIN*  
*Institute of Biosciences and Bioresources, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze, Italy, †IGA Technology Services s.r.l., Via J. Linussio, 51, 33100 Udine, Italy, ‡Department of Forest Ecology and Genetics, National Institute for Agriculture and Food Research and Technology (INIA), Forest Research Centre (CIFOR), Carretera de A Coruña km 7.5, 28040 Madrid, Spain, §Department of Ecology and Evolution, University of Lausanne, Biophore Building, 1015 Lausanne, Switzerland, ¶Plant Protection Institute, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze, Italy, **Dipartimento di Scienze Agrarie e Ambientali, Università di Udine, via delle Scienze 208, 33100 Udine, Italy, ††Centro de Investigaciones sobre Desertificación (CIDE-CSIC/UV/GV), Ctra. Naquera Km. 4, 46113 Moncada, Valencia, Spain

Abstract

Aleppo pine (*Pinus halepensis* Mill.) is a relevant conifer species for studying adaptive responses to drought and fire regimes in the Mediterranean region. In this study, we performed Illumina next-generation sequencing of two phenotypically divergent Aleppo pine accessions with the aims of (i) characterizing the transcriptome through Illumina RNA-Seq on trees phenotypically divergent for adaptive traits linked to fire adaptation and drought, (ii) performing a functional annotation of the assembled transcriptome, (iii) identifying genes with accelerated evolutionary rates, (iv) studying the expression levels of the annotated genes and (v) developing gene-based markers for population genomic and association genetic studies. The assembled transcriptome consisted of 48,629 contigs and covered about 54.6 Mbp. The comparison of Aleppo pine transcripts to *Picea sitchensis* protein-coding sequences resulted in the detection of 34,014 SNPs across species, with a $K_a/K_s$ average value of 0.216, suggesting that the majority of the assembled genes are under negative selection. Several genes were differentially expressed across the two pine accessions with contrasted phenotypes, including a glutathione-s-transferase, a cellulose synthase and a cobra-like protein. A large number of new markers (3,334 amplifiable SSRs and 28,236 SNPs) have been identified which should facilitate future population genomics and association genetics in this species. A 384-SPN Oligo Pool Assay for genotyping with the Illumina VeraCode technology has been designed which showed an high overall SNP conversion rate (76.6%). Our results showed that Illumina next-generation sequencing is a valuable technology to obtain an extensive overview on whole transcriptomes of nonmodel species with large genomes.

Keywords: association genetics, next-generation sequencing, population genomics, SNPs, SSRs, transcriptome

Received 12 November 2013; revision received 15 January 2014; accepted 17 January 2014

Introduction

Mediterranean forests are key ecosystems that provide numerous products and services, constitute long-term reservoirs of genetic and species diversity and are characterized by highly heterogeneous and fragmented environments. These forests include many populations located at ecological and geographical range margins (Hampe & Petit 2005; Hampe & Jump 2011) and are often threatened by, but in some cases adapted to, factors such as climatic change, forest fires, overexploitation, pollution, land-use change and fragmentation (Underwood *et al.* 2009). Reduction of water availability in Mediterranean forests is leading to massive die-backs of trees often followed by pest outbreaks that further increase tree mortality (Heres *et al.* 2011) and modifies the susceptibility to fire (Hicke *et al.* 2012). In this context, standing genetic variation, that is the genetic variation that is already present in the population (in contrast to new mutation), and the pivotal effect of gene flow and migration are fundamental for the persistence of the
Mediterranean forests (Aitken et al. 2008, Thompson et al. 2009). The relative roles of selection, gene flow and genetic drift on local adaptation in heterogeneous environments are, however, not yet well documented (Savolainen et al. 2007; Alberto et al. 2013).

Aleppo pine (Pinus halepensis Mill.), being highly tolerant to drought and growing in a vulnerable ecosystem (where decreasing precipitation, and increasing temperatures and intensity and frequency of fires are observed), is an important model species for studying adaptive responses to changes in drought and fire regimes. Widespread in the Mediterranean region, Aleppo pine is a thermophilous, shade-intolerant and highly plastic tree species (e.g. for biomass allocation, Chambel et al. 2007; vegetative growth, Santos-del-Blanco et al. 2013), which escapes drought by an effective stomata control (Borghetti et al. 1998). Nevertheless, growth processes in this species seem to be significantly impacted by the frequency and duration of drought events (Girard et al. 2012). Common garden tests and genomic studies showed that Aleppo pine is characterized by high variability in intrinsic water-use efficiency (IWUE) and growth performance (e.g. Voltas et al. 2008), and that local adaptation is recognizable in candidate genes involved in responses to drought, for example aquaporins and dehydrins (Grivet et al. 2009). Another important factor shaping Aleppo pine ecology, population dynamics and distribution is wildfire. Aleppo pine is considered a pyrophyte species that regenerates well after fire (Pausas et al. 2004); it presents notable fire-related adaptations, such as early flowering, high serotiny, high resin content and a branching pattern (with no self-pruning) that favours crown fires (He et al. 2012; Hernández-Serrano et al. 2013). Under high fire recurrence, this species is able to expand its range, replacing other less-pyrophyte Mediterranean vegetation (Barbéro et al. 1998).

A relatively simple recolonization history has been described for Aleppo pine; this species would have spread westwards from a unique refugium in Southern Balkans by long-distance colonization events possibly followed by severe bottlenecks (Grivet et al. 2009). In fact, eastern populations harbour larger genetic variation than western populations (Bucci et al. 1998; Grivet et al. 2011). However, new genomic resources are needed to accurately date this colonization event. Furthermore, research focused on drought-response candidate genes provided evidence that both selection and demography may have played a role in genetic diversity reduction during Aleppo pine range expansion (Grivet et al. 2009, 2011). But these studies were based on a low number of genes (6 and 10, respectively), and ‘gene surfing’ in the front wave of colonization (Excoffier & Ray 2008) would have produced similar signatures to selective processes. A larger number of genes would thus be necessary to study the colonization process of Aleppo pine and to effectively distinguish between demography and selection in this species. Furthermore, this information would help to unravel the factors associated with wildfire that shape fine-scale pine population structure (Hernández-Serrano et al. 2013). Quantitative genetic experiments (i.e. common gardens) are readily available in Aleppo pine (e.g. Climent et al. 2008; Voltas et al. 2008; Santos-del-Blanco et al. 2013). However, we lack enough markers to study the molecular basis of adaptation using large-scale genome-wide association studies (GWAS): despite the importance of Aleppo pine, only about thirty candidate genes, mainly related to drought response, are currently available (Grivet et al. 2009, 2011, 2013). These genes were originally identified on the basis of functional studies performed in Pinus taeda and other conifers or derived from model species such as Arabidopsis thaliana (Gonzalez-Martinez et al. 2006; Grivet et al. 2009; Wachowiak et al. 2009), and may provide a biased view of Aleppo pine genome.

Increased genome-wide resolution in population genomics and association genetics has been greatly facilitated by the emergence of next-generation sequencing (NGS) technologies, such as RNA-Seq. Here, we utilize the Illumina pair-end technology to sequence the transcriptome of Aleppo pine with the main aim of developing genomic resources to further support population genomic and association genetic studies in this and in other closely related Mediterranean pines (e.g. Pinus brutia). We also aim to provide a full description of Aleppo pine transcriptome, including evolutionary rates and expression patterns, as well as the identification of relevant candidate genes for further research. Specifically, we aim to (i) perform a functional annotation of the transcriptome, (ii) identify genes or groups of genes with accelerated evolutionary rates, (iii) study the expression levels of the annotated genes and (iv) develop both neutral and potentially functional gene-based markers (SSRs and SNPs). For this purpose, the cDNA of two Aleppo pine individuals, phenotypically divergent for traits related to fire strategy (as indicated by the serotiny level), was sequenced. Transcriptome sequencing is an attractive alternative to whole genome sequencing, especially for those species with large and complex genomes (~26 Gbp in pine), because it enables to efficiently explore the functional part of a genome with affordable sequencing effort.

Materials and methods

Plant material, RNA extraction

Two Aleppo pine individuals with contrasting fire-response phenotypes (as gauged by serotiny levels, i.e.
the retention of closed cones till opened by fire) were sampled in two populations located in eastern Spain, a region known for its high variability for fire traits in this species (Hernández-Serrano et al. 2013; Budde et al. 2014). The needles were collected at the same time of day under similar climatic conditions and were immediately stored in liquid nitrogen prior to RNA extraction. Accession 1716 was collected in Sinarcas (latitude: 39.79535, longitude: −1.20522) and had extremely low serotiny levels (4.35% of cones were serotinous), whereas Accession 1445 was collected in Serra Calderona (latitude: 39.73807, longitude: −0.48061) and represented a tree at the other extreme of the serotiny distribution (96.77% of cones were serotinous). Because serotiny is correlated with other fire-related traits (Schwilk & Ackerly 2001) and to summer drought (Hernández-Serrano et al. 2013), a known driver of fire (Pausas 2004; Pausas & Paula 2012), trees with extreme levels of serotiny are expected to have also high variability for other adaptive traits in Aleppo pine. Serotiny is a highly heritable trait (Wymore et al. 2011; Budde et al. 2014 and reference therein) showing also large phenotypic variation within and among species and populations of Mediterranean pine species (e.g. He et al. 2012).

Total RNA was extracted from 3 to 5 needles per tree using the Spectrum Plant Total RNA kit (SIGMA, Saint Louis, MO, USA) following the manufacturer’s protocol.

Library preparation and sequencing

RNA-Seq libraries were generated using the TruSeq RNA-Seq Sample Prep kit according to the manufacturer’s protocol (Illumina Inc., San Diego, CA, USA). Briefly, poly-A RNA was isolated from total RNA and chemically fragmented. First- and second-strand synthesis were followed by end repair, and adenosines were added to the 3’ ends. Adapters were ligated to the cDNA, and 200 ± 25 bp fragments were gel purified and enriched by PCR. Libraries were quantified using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA), pooled in equimolar amount together with other two samples not included in the present study and sequenced in a single lane on the Illumina HiSeq2000 (Illumina Inc.) as paired-end reads of length 101 bp.

Assembly

The filter-for-assembly module implemented in rna (Vezzi et al. 2012) was used to trim the bases with a quality score below 20 from the read ends and to filter for chloroplast contaminants by aligning the short reads to the Pinus contorta chloroplast complete genome (NCBI Reference Sequence: NC_011153.4). After quality trimming and filtering, only pairs having both reads longer than 90 bp were retained. Trimmed reads of both samples were assembled using a combination of reference-based and de novo assembly. The reference-based assembly was performed by aligning the short reads to a set of 18 092 unigenes constructed from P. taeda Sanger ESTs (P. taeda: UniGene Build #13. http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3352; overall genetic divergence, K, between Aleppo pine and P. taeda of ~5%, Grivet et al. 2011). The alignment was carried out with BWA (Li & Durbin 2009), a tool for mapping low-divergent sequences to a reference, with the aim to reconstruct the sequences of Aleppo pine genes having high homology with P. taeda ones. We used BWA allowing at most five mismatches and one gap per read. A pileup file describing the base-pair information at each position of the reference (i.e. P. taeda genes) was obtained using the samtools mpileup utility (Li et al. 2009), considering only bases with quality equal or >12. Consensus calls were obtained using the pileup2cns command included in Varscan v2.2.8 (Koboldt et al. 2009) requiring a minimum coverage of one. Aleppo pine consensus sequences were created only for the unigenes covered in at least 75% of their length. Reads not aligning to the unigenes were de novo assembled using the CLC Genomic Workbench 5.0 software with automatically generated parameters (word size = 23 and bubble size = 50) and requiring a minimum contig size of 200. Contigs obtained from short-read alignment and de novo assembly were included in a single multifasta file and processed with the CAP3 sequence assembly programme (Huang & Madan 1999) with default parameters to obtain a nonredundant list of contigs. To remove any possible source of contamination from the final assembly, the list of contigs was used as query for a BLASTn (Altschul et al. 1990) search against the nt database. All contigs for which the best BLASTn hit was not included in the Viridiplantae subset (Taxid: 33090) were removed from the assembly.

Annotation

The transcriptome assembly was used as query for a BLASTx (Altschul et al. 1990) analysis against the Viridiplantae section of the RefSeq database (Pruitt et al. 2009) using an E value threshold of 10−6. BLASTx results were imported into BLAST2GO version V.2.5.0 (Conesa et al. 2005) for the Gene Ontology (GO) assignment. The Annex tool implemented in BLAST2GO was used to improve the annotation by deriving terms due to verified links from ‘Molecular Function’ terms to ‘Biological Process’ and ‘Cellular Component’ ones. Annotation results were summarized through the mapping to the Plant GO-Slim, a reduced version of the Gene Ontology containing a selected number of nodes relevant for plants. The percentages of annotated contigs for the three
GO categories ‘Molecular Function’, ‘Biological Process’ and ‘Cellular Component’ were reported only for categories represented in at least the 1% of the annotated contigs. Enzyme codes and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway annotations were generated in Blast2GO from the direct mapping of the GO terms to their enzyme code equivalents.

SSRs and SNPs identification

Simple sequence repeats (SSRs) were identified in the final assembly using Sputnik (http://espressosoftware.com/sputnik/index.html). The analysis was performed by requiring a minimum number of four contiguous repeats and a minimum score (calculated as per cent perfection) of 80. SSR motifs were named using the convention of alphabetical ordering (Echt & May-Marquardt 1997). SSR density was calculated as the number of SSRs per Mb of the analysed sequence. The programme BatchPrimer3 (You et al. 2008) was employed to design PCR primers in the flanking regions of the detected SSRs, setting a minimum primer length of 20 bp, a minimum GC content of 30%, a melting temperature between 52 and 62 °C and a maximum melting temperature difference between primers of 4 °C. Primer pairs amplifying single fragments in the transcriptome assembly were identified with an in silico PCR using the tool jPCR (Kalender et al. 2001).

To identify single nucleotide polymorphisms (SNPs), the short reads of the two sequenced accessions were separately aligned to the reference transcriptome using the short-read aligner BWA (Li & Durbin 2009) with default parameters. For each accession, SNP detection on uniquely aligned reads was carried out using the variant caller VARScan v2.2.8 (Koboldt et al. 2009) by setting the minimum read depth at a position to make a call to 10 and the minimum frequency of reads carrying the variant allele to 0.3. VARScan was employed with the same settings to perform also small indel detection in the two accessions, and the obtained indel list was used to filter for likely miscalled SNPs nearby indels with the Varscan command filter.

SNP analysis

To assess the long-term selective pressure acting on assembled genes (i.e. evolutionary rates), the Aleppo pine-assembled transcriptome was compared with 18 762 *Picea sitchensis* protein-coding sequences retrieved from the European Nucleotide Archive (ENA) database (http://www.ebi.ac.uk/ena/home). To select a list of homologous genes between the two species, the Aleppo pine contigs were aligned to the *P. sitchensis* sequences using BLASTn with an E value threshold of 10^-6. We selected only the portions of the Aleppo pine contigs that aligned for at least 250 consecutive bases to *P. sitchensis* genes. Then, polymorphic positions between Aleppo pine and *P. sitchensis* were identified with a custom R function and functionally annotated with Annovar (Wang et al. 2010) using the *P. sitchensis* ORFs as reference to predict the effects of SNPs on protein sequence. The $K_s/K_a$ ratio was calculated with the kaks function included in the R library ‘seqinr’. This function makes an unbiased estimate of the ratio of nonsynonymous ($K_a$) to synonymous ($K_s$) nucleotide substitution for a set of aligned sequences (Li et al. 1993). The *P. sitchensis* orthologous genes were stratified by $K_a/K_s$ values ($K_s/K_a = 0, 0 < K_s/K_a \leq 0.15, 0.15 < K_s/K_a \leq 0.5, K_s/K_a > 0.5$). Over-representation of GO terms in the four subsets, as compared with the whole assembly, was tested using a Fisher’s exact test ($P = 0.05$) implemented in the Gossip package integrated in Blast2GO (Blüthgen et al. 2005). A false discovery rate (FDR) correction for multiple testing (Benjamini & Hochberg 2007) was applied, and only differences with a corrected $P$-value <0.05 were considered.

Design a 384-SNP Oligo Pool Assay (OPA)

As a first attempt to verify the efficiency of high-throughput transcriptome sequencing to select new molecular markers in this species, a subset of 240 selected SNPs identified in this study together with 144 additional SNPs obtained by standard Sanger sequencing within the EvolTree EU project (www.evoltree.eu; CRIEC initiative) was used to design a 384-SNP Oligo Pool Assay (OPA) (Table S4, Supporting information) for genotyping with the Illumina VeraCode technology. SNPs to include in the chip were selected in order to prioritize the following: a) SNPs fixed for different alleles in the two sequenced trees; b) SNPs belonging to genes with differential expression across individuals; c) SNPs belonging to genes with elevated nonsynonymous ($K_a$) to synonymous ($K_s$) divergence ratio and d) SNPs from genes with annotation corresponding to adaptive loci in forest trees. OPA performance was evaluated by standard conversion rates (i.e. the percentage of polymorphic SNPs with call rates higher than 0.85).

Expression analysis

Overall, gene expression levels of the assembled genes were estimated as the mean per base sequence coverage obtained by aligning to the transcriptome assembly the short reads of both accessions and combining results in a single BAM file. Coverage was measured by dividing the total number of aligned bases in each gene by the number of the positions of the gene covered by at least one read. Genes were classified in three gene expression...
categories on the bases of their mean sequence coverage: low (mean coverage <10×), medium (mean coverage 10×–100×) and high (mean coverage >100×). Sample-specific BAM files were employed to perform differential expression analysis between the two sequenced accessions. The DESeq R package (Anders & Huber 2010) was employed to identify up-regulated and down-regulated genes between the two sequenced trees. As only two trees were used in analyses of differential expression (i.e. no biological replicates), results have to be interpreted with caution. Normalization was made using size factors after calculation of relative library sizes (see manual for details). After dispersion estimation for each gene (pooling the two trees), the binomial test described in Anders & Huber (2010) was used to identify differentially expressed genes, after correction for multiple testing using the Benjamini-Hochberg procedure.

Results and discussion

Sequencing and assembly

The two cDNA libraries prepared using RNA of the two Aleppo pine trees included a total of 83 922 352 (paired-end) reads 101 bp long corresponding to ~8.5 Gb. After filtering for low-quality sequences, ~69 million high-quality reads were retained, corresponding to 6.9 Gb (Table S1, Supporting information). Reads were aligned to the Pinus contorta chloroplast genome and covered about the 90% of its length with a mean coverage of 188× (Fig. S1, Supporting information). A reference-based assembly was performed by aligning the high-quality filtered reads to a reference composed by a set of approximately 18 000 P. taeda unigenes, and 12 884 unigenes were covered for at least 75% of their length and used to create the corresponding Aleppo pine consensus sequences with a mean length of ~800 bp that covered in total ~10.3 Mb (Table 1). Unaligned reads were employed to perform a de novo assembly using the CLC Workbench Assembler. An assembly composed of 47 811 contigs, having a mean length of 1040 bp, a minimum contig length of 128 bp and corresponding to a total length of 49.7 Mb, was obtained. Unigene consensus sequences and de novo assembled contigs were joined in 49 612 contigs using CAP3. Out of those, 983 contigs showed high homology with genes not included in the clade Viridiplantae and were removed from the assembly as probable contaminants. The majority of them showed homology with fungal genes, possibly indicating the presence of endophytic fungi in the sampled tissues. The final assembly was composed of 48 629 contigs having a mean length of 1122 bp, a N50 size of 1792 bp and covering about 54.6 Mb. The number of genes in Aleppo pine is unknown, but our results are comparable with those obtained in P. taeda (Lorenz et al. 2011), for which an assembly composed of 48 751 contigs with a mean length of about 859 bp was obtained. In addition, transcriptome size of Aleppo pine is also consistent with the estimate of 47 Mb obtained for the white spruce transcriptome (Rigault et al. 2011) using a combination of Sanger and NGS. Similarly, Howe et al. (2013) obtained a transcriptome composed of about 40 000 contigs with an average length of 1390 bp using 454 sequences in Douglas fir, while Nystedt et al. (2013) reconstructed a high-confidence gene set in Norway spruce covering 27 Mb of protein-coding sequence.

Annotation

The Aleppo pine transcriptome was annotated by aligning the contigs to the Viridiplantae section of the RefSeq database: of 48 629 contigs, 27 671 (~57%) had a BLASTx hit to proteins included in the database; the majority of them matched to Glycine max, Vitis vinifera or Populus trichocarpa proteins, which are better represented in RefSeq database than conifer proteins (Fig. S2, Supporting information). The high proportion of contigs matching to known proteins indicates the high quality of assembled genes. As expected, nonannotated contigs had a smaller average length (809 bp) and thus were less probably to have a BLASTx match. Using Blast2GO, we assigned a total of 105 499 gene ontology (GO) terms to 19 849 (72%) of the 27 671 contigs for which we obtained a BLASTx match. Most of the assignments (48 096; 46%) were assigned to GO terms related to the photosynthesis process, basal metabolic process and metal ion binding (Table S2, Supporting information).

Table 1  Reference-guided and de novo assembly phases

<table>
<thead>
<tr>
<th>Assembly phase</th>
<th>Number of contigs</th>
<th>Assembly length (bp)</th>
<th>Average contig length (bp)</th>
<th>N50 (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference-guided assembly with Pinus taeda unigenes</td>
<td>12 884</td>
<td>10 284 477</td>
<td>798</td>
<td>836</td>
</tr>
<tr>
<td>CLC Genomic Workbench de novo assembly</td>
<td>47 811</td>
<td>49 713 723</td>
<td>1040</td>
<td>1758</td>
</tr>
<tr>
<td>Merge of reference-guided and de novo assemblies</td>
<td>49 612</td>
<td>55 169 568</td>
<td>1112</td>
<td>1777</td>
</tr>
<tr>
<td>Filtering for contaminants</td>
<td>48 629</td>
<td>54 571 354</td>
<td>1122</td>
<td>1792</td>
</tr>
</tbody>
</table>

Number of contigs, the total length of the assembly, the average contig length and the N50 statistic for each of the assembly phases.
belonged to the ‘Biological Process’ category, while the remaining were equally shared between the ‘Molecular Function’ (28 427; 27%) and the ‘Cellular Component’ categories (28 965; 27%). The distribution of gene ontology annotations in the different functional categories showed a large diversity of the annotated transcripts and evidenced the high representativeness of our Aleppo pine transcriptome (Fig. 1). This GO distribution was consistent with the one obtained in Pinus contorta (Parchman et al. 2010). To generate the corresponding Enzyme Codes (ECs) and KEGGs, the GO terms were mapped to their enzyme code equivalents: 7775 Enzyme Codes were assigned to 6276 contigs and were included in 139 different KEGG pathways (Table S2, Supporting information).

The most represented enzymes belonged to the following KEGG classes: transferase activity (32%), oxidoreductase activity (29%) and hydrolase (17%) (Fig. S3, Supporting information). The most represented pathways were ‘Purine metabolism’ with 52 ECs assigned to 419 different contigs and ‘Starch and sucrose metabolism’ with 36 ECs assigned to 399 contigs (Table S2, Supporting information). Another highly represented pathway was the ‘Phenylpropanoid biosynthesis’ one (map00940), which generates an enormous array of secondary metabolites with a wide variety of functions both as structural and signalling molecules and includes all the enzymes required for the biosynthesis of monolignols (i.e. lignin formation; see Fig. S4, Supporting information).

SSR and SNP identification

A total of 8248 di-, tri-, tetra- or penta-nucleotide SSRs composed of at least four contiguous repeated units were identified (Table 2). Di- and tri-nucleotide repeats comprised the most abundant SSR motif classes, while tetra- and penta-nucleotide repeats were detected at much lower frequencies. Among all, the dinucleotide SSRs (AT)\textsubscript{n} and (AG)\textsubscript{n} were the most abundant motifs and
Identification of common SNPs was based on two diploid genomes (i.e. four gametes). Considering only variants for which we had a sufficient coverage to call the genotype in both individuals, we detected a total of 28 236 polymorphic positions distributed in 6475 different contigs. Of them, 9509 were polymorphic in the two accessions (i.e. the two trees were heterozygous), while the remaining were specific polymorphisms of one of them. Considering that 17 949 332 positions of the assembly had a sufficient coverage in both individuals, we detected one SNP every 636 bp. Comparing with the SNP density reported in classical Sanger sequencing studies for Aleppo pine (one SNP every 71 bp; Grivet et al. 2009, 2011), we were able to identify, using only two accessions with contrasted phenotypes, about one every nine SNPs expected in the transcriptome. The significant difference in polymorphism between the Illumina and the Sanger data is associated with the low discovery panel used in this study (only four gametes compared with over 48 in Grivet et al. 2009, 2011).

A 384-SNP Oligo Pool Assay (OPA) (Table S4, Supporting information) for genotyping with the Illumina VeraCode technology was designed and tested. The overall SNP conversion rate, estimated by genotyping 1332 individuals from 45 populations covering the full distribution range of the species, was high (76.6%) and comparable (or superior) to those in other pines (e.g. 66.9% in loblolly pine, Eckert et al. 2009; 51% in maritime pine, Lepoittevin et al. 2010 and Douglas fir (72.5%, Howe et al. 2013). SNPs originated in transcriptome sequencing had only slightly lower conversion rates if compared with those obtained from traditional Sanger sequencing (74.2% vs. 80.55%), thus confirming the efficiency of this method for high-quality marker development. However, it should be stressed that the low panel size used to discover SNPs (only two individuals) may render this OPA unsuitable for some applications as those based on reconstruction of site frequency spectrum or where rare polymorphism is of particular interest.

**Evolutionary divergence analysis**

We selected 1291 *P. sitchensis* genes showing high homology with Aleppo pine transcripts, corresponding to about 673 Kbp of homologous sequences. A total of 34 014 polymorphic positions between the two species, one every 20 bp, was identified. Most of the identified SNPs (~66%) were synonymous; among the 11 430 non-synonymous polymorphisms, 47 caused a premature stop codon in *P. halepensis* with respect to the *P. sitchensis* coding frame, while 2 produced a stop loss (Table S5, Supporting information). Some of the genes with premature stop codons encoded for proteins that have a known relevant functional role in plants. Among them, we
found a beta-tubulin (contig Phal_PtaS20503129), which is implicated in the formation of plant cell walls (Spokevicius et al. 2007) and a glutathione-s-transferase (contig cle_contig_411119), an enzyme implicated in the response to many forms of biotic and abiotic stress in plants (Marrs 1996).

The average values of nonsynonymous substitutions per nonsynonymous sites (Ka) and synonymous substitutions per synonymous sites (Ks) across the 1291 genes were 0.025 (range 0–0.123), 0.143 (0.016–0.678) and 0.216 (0–3.416), respectively. A simple calculation (as in Brown et al. 2004) based on synonymous divergence (Ks), generation time (~25 years, Grivet et al. 2009), synonymous nucleotide diversity for eastern Spain (~0.001; Grivet et al. 2009, 2011) and split time between Picea and Pinus (~123 Ma, Gernandt et al. 2008; Chen et al. 2012) produced an approximate estimate of effective population size (Ne) of ~17 000, about one-fifth of that found in other conifers (e.g. Williard et al. 2007).

A Ka/Ks average value of 0.216 suggested that the majority of the assembled genes are under negative (purifying) selection (Table 3). Sixty-five genes had Ka/Ks values equal to zero, or only slightly greater than zero, indicating that these genes have evolved under strong selective constraint. The majority of the sequences with the lowest Ka/Ks values showed significant similarity to housekeeping genes that typically evolve slowly (Li 1993, 1997), such as ribosomal proteins, histone proteins and cytochromes. About 1.2% of sequences showed Ka/Ks >1, suggesting that positive selection could have contributed to the interspecific sequence divergence of the corresponding genes. Our results agree with those by Chen et al. (2012) in Picea abies where Ka/Ks ratios were in general much lower than 1, and only few genes showed a ratio larger than 1; this contrasts with what was observed by Buschiazzo et al. (2012) in P. sitchensis and P. taeda analysing relatively short alignments. As in Chen et al. (2012), no significant correlation between Ka/Ks ratios and alignment lengths was observed in our study. Some of the genes that displayed values higher than 1 in Aleppo pine encode for proteins of interest in plants such as a phytochrome, implicated in flower phenology, and a glutathione-s-transferase, involved in biotic and abiotic stress response. We divided the studied genes in four categories on the basis of the Ka/Ks value (Table 3) and tested for enrichment in gene ontology terms involved in ‘Biological Processes’, ‘Cellular Components’ and ‘Molecular Function’. For genes with low Ka/Ks values, we identified a number of GO terms that were overrepresented (FDR <0.05) compared with the entire data set (Fig. S5, Supporting information). The most overrepresented ones were associated with essential biological processes such as those involved in ribosome assembly, which are highly conserved across species (McIntosh & Bonham-Smith 2001), those involved in translation, which is an essential step for protein biosynthesis, and those involved in protein binding that mediates many essential cellular processes such as signal transduction, transport, cellular motion and most regulatory mechanisms (Sharan et al. 2005). Similar results have been reported in a previous work on Eucalyptus grandis (Novaes et al. 2008). At the other extreme, genes under positive selection were not enriched for specific GO categories. The proportion of annotated genes decreased when Ka/Ks increases (Table 3). A possible explanation is the abundance of duplicated genes in high Ka/Ks categories, whose significant sequence divergence may have prevented Blast2GO annotation. Another possibility is that these categories were enriched for pseudogenes and/or transcribed but not translated loci.

### Expression analysis

About 90% of the sequenced reads were aligned to the assembly, producing a mean coverage of 70×, with a decreasing proportion of genes at higher expression levels (Fig. S6, Supporting information). Only 15% of the genes were highly expressed (mean coverage >100×), while 50% of the genes showed a low expression (mean coverage <10×). No significant correlation between the expression level and the selective constraint measured by Ka/Ks was observed This is in contrast to previous studies in other species (Subramanian et al. 2010; Yang & Gaut 2011) where highly expressed genes were subjected to stronger selection but is in agreement with what Chen et al. (2012) found in Picea abies.

Short reads of the two sequenced Aleppo pine accessions were also separately aligned to the transcriptome in order to perform a differential expression analysis between the two individuals. The use of only two individual without biological replicates is not ideal to perform differential expression analysis and must be taken with caution. Therefore, the differentially expressed genes that we detected should be considered as

<table>
<thead>
<tr>
<th>Ka/Ks range</th>
<th># contigs</th>
<th>Average contig length</th>
<th>% annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>65</td>
<td>571</td>
<td>97</td>
</tr>
<tr>
<td>0–0.15</td>
<td>553</td>
<td>764</td>
<td>92</td>
</tr>
<tr>
<td>0.15–0.5</td>
<td>578</td>
<td>826</td>
<td>80</td>
</tr>
<tr>
<td>&gt;0.5</td>
<td>57</td>
<td>809</td>
<td>62</td>
</tr>
</tbody>
</table>

Number of contigs by Ka/Ks value. For each group, the average length of the contigs and the percentage of annotated contigs are also reported.
candidates deserving further investigation. We identified a total of 541 genes that showed a signature of differential expression in the two samples, including a glutathione-s-transferase (contig825, 134-fold change, p.adj = 7.5E-09), a cellulose synthase (contig6777, only expressed in the fire-adapted tree, p.adj = 2.3E-18) and a cobra-like protein (contig5088, only expressed in the fire-adapted tree, p.adj = 4.2E-10); about 150 of these were completely absent in one of the two trees suggesting that they were expressed only in one tree.

Conclusions and perspectives

This sequence collection represents the first major genomic resource for Aleppo pine. The large number of genes, SNPs and SSRs detected in this study will facilitate population genomics and association genetics in this species. In particular, they will help addressing two major research questions in Mediterranean Aleppo pine: the genetic effects of range expansion and the molecular basis of genetic responses to global change. This is particularly important considering that this species is one of the major components of the Mediterranean ecosystem, which is severely threatened by climatic change, forest fires and land-use changes (Pausas 2004; Pausas & Fernández-Muñoz 2012). As already stressed by Hao et al. (2011) in Taxus mairei, this study suggests that Illumina second-generation sequencing is a valid technology to obtain whole transcriptomes, being especially useful in nonmodel species with large genomes (such as conifers). Aleppo pine transcriptome has similar characteristics to other conifers (e.g. Pinus contorta, Parchman et al. 2010; Taxus mairei, Hao et al. 2011; Picea abies, Chen et al. 2012), suggesting a conserved pattern in transcription levels across species. The large set of SSR and SNP markers discovered in this work will help to better understand genome-wide patterns of adaptive variation in this keystone Mediterranean forest tree species.

Acknowledgements

This research was funded by the Spanish National Research Plan (VAMPIro, CGL2008-05289-C02-01/02, TrEvol, CGL2012-39938-C02-01, and AdapCon, CGL2011-30182-C02-01), and the ERA-Net BiodivERsA (LinkTree project, EU2008-03713), which included the Spanish Ministry of Economy and Competitiveness as national funder (part of the 2008 BiodivERsA call for research proposals). SP and GGV were also supported by a grant of the Italian Ministry of Education and Scientific Research (‘Biodiversitalia’, RBAP10A2T4). DG acknowledges the support of the Spanish Ministry of Economy and Competitiveness through a ‘Ramón y Cajal’ fellowship and SCGM the support of a Senior Marie Curie Intra European Fellowship within the 7th European Community Framework Programme (PIEF-GA-2012-328146).

References


© 2014 John Wiley & Sons Ltd
Novae E, Drost DR, Farmerie WG et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics, 9, 312.


Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold related


S.C.G.-M. and G.G.V. provided funding and organized and planned the research. S.P., F.B., S.C.G.-M. and G.G.V. drafted the manuscript. S.P. and F.C. did the wet laboratory procedures. S.P., F.C. and F.M. performed the bioinformatic analysis of the data. J.P. and M.V. selected the material and did the phenotypic characterization. D.G., Z.L. and S.C.G.-M. selected the SNPs for the OPA design. Z.L., J.P., M.V. and S.C.G.M. performed the SNP genotyping. All authors contributed to and approved the final version of the manuscript.

Data Accessibility

Raw sequence data are freely available from the NCBI short read archive, Accession numbers SRR942848 e SRR942867 (see Table S1, Supporting information). The assembled transcriptome and the annotations (GOs and KEGGs assignments) are available from Dryad entry doi: 10.5061/dryad.vb1311

Supporting Information

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

**Fig. S1** Graphic representation of the coverage distribution obtained in the *Pinus contorta* chloroplast genome.

**Fig. S2** Species distribution of BLASTx matches.

**Fig. S3** Catalytic activity distribution in annotated Aleppo pine contigs.

**Fig. S4** The phenylpropanoid biosynthesis pathway taken from KEGG.

**Fig. S5** Overrepresented GO terms in *Ko/Ks* categories.

**Fig. S6** Aleppo pine transcript expression levels.

**Table S1** Sample, library identification and sequence metrics.

**Table S2** List of all the pathways with the number of annotated ECs.

**Table S3** Information on SSRs primer sequences and working conditions.

**Table S4** Design file (Oligo Pool Assay) for genotyping 384 Aleppo pine SNPs with Illumina VeraCode technology.

**Table S5** List of all the identified SNPs.