Plant Biotechnology Journal

Plant Biotechnology Journal (2013), pp. 1–14 doi: 10.1111/pbi.12136

Research Article

De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology

Javier Canales[1†], Rocio Bautista[2†], Philippe Label[3†], Josefa Gómez-Maldonado[1], Isabelle Lesur[4,5,6], Noe Fernández-Pozo[2], Marina Rueda-López[1], Dario Guerrero-Fernández[2], Vanessa Castro-Rodríguez[1], Hicham Benzekri[2], Rafael A. Cañas[1], María-Angeles Guevara[7], Andreia Rodrigues[8], Pedro Seoane[2], Caroline Teyssier[9], Alexandre Morel[9], François Ehrenmann[4,5], Grégoire Le Provost[4,5], Céline Lalanne[4,5], Céline Noirot[10], Christophe Klopp[10], Isabelle Reymond[11], Angel García-Gutiérrez[1], Jean-François Trontin[11], Marie-Anne Lelu-Walter[9], Celia Miguel[8], María Teresa Cervera[7], Francisco R. Cantón[1], Christophe Plomion[4,5], Luc Harvengt[11], Concepción Avila[1,2], M. Gonzalo Claros[1,2] and Francisco M. Cánovas[1,2]*

[1]*Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Málaga, Spain*
[2]*Plataforma Andaluza de Bioinformática, Edificio de Bioinnovación, Parque Tecnológico de Andalucía, Málaga, Spain*
[3]*INRA, Université Blaise Pascal, Aubière Cedex, France*
[4]*INRA, Cestas, France*
[5]*Université de Bordeaux, Talence, France*
[6]*HelixVenture, Mérignac, France*
[7]*Departamento de Ecología y Genética Forestal, INIA-CIFOR, Madrid, Spain*
[8]*Forest Biotech Lab, IBET/ITQB, Oeiras, Portugal*
[9]*INRA, Unité Amélioration, Génétique et Physiologie Forestières, Orléans Cedex 2, France*
[10]*INRA de Toulouse Midi-Pyrénées, Auzeville, Castanet Tolosan cedex, France*
[11]*FCBA, Pôle Biotechnologie et Sylviculture, Cestas, France*

Received 20 July 2013;
revised 24 September 2013;
accepted 26 September 2013.
*Correspondence (Tel: +34 952131942;
fax: +34 952132376;
email: canovas@uma.es)
†These authors contributed equally to work.

Keywords: conifers, transcriptome sequencing, next-generation sequencing, full-length cDNA, transcription factors, single nucleotide polymorphism.

Summary

Maritime pine (*Pinus pinaster* Ait.) is a widely distributed conifer species in Southwestern Europe and one of the most advanced models for conifer research. In the current work, comprehensive characterization of the maritime pine transcriptome was performed using a combination of two different next-generation sequencing platforms, 454 and Illumina. *De novo* assembly of the transcriptome provided a catalogue of 26 020 unique transcripts in maritime pine trees and a collection of 9641 full-length cDNAs. Quality of the transcriptome assembly was validated by RT-PCR amplification of selected transcripts for structural and regulatory genes. Transcription factors and enzyme-encoding transcripts were annotated. Furthermore, the available sequencing data permitted the identification of polymorphisms and the establishment of robust single nucleotide polymorphism (SNP) and simple-sequence repeat (SSR) databases for genotyping applications and integration of translational genomics in maritime pine breeding programmes. All our data are freely available at *SustainpineDB*, the *P. pinaster* expressional database. Results reported here on the maritime pine transcriptome represent a valuable resource for future basic and applied studies on this ecological and economically important pine species.

Introduction

Forests are essential components of the ecosystems covering approximately one-third of the Earth's land area and playing a fundamental role in the regulation of terrestrial carbon sinks. Trees represent nearly 80% of the plant biomass (Olson *et al.*, 1983) and 50%–60% of annual net primary production in terrestrial ecosystems (Field *et al.*, 1998).

Conifers are the most important group of gymnosperms. Having diverged from a common ancestor more than 300 million years ago (Bowe *et al.*, 2000), gymnosperms and angiosperms have evolved very efficient and distinct physiological adaptations (Leitch and Leitch, 2012). Coniferous forests dominate large ecosystems in the Northern Hemisphere and include a broad variety of woody plant species, some of which are the largest, tallest and longest living organisms on Earth (Farjon, 2010).

Please cite this article as: Canales J., Bautista R., Label P., Gómez-Maldonado J., Lesur I., Fernández-Pozo N., Rueda-López M., Guerrero-Fernández D., Castro-Rodríguez V., Benzekri H., Cañas R. A., Guevara M.-A., Rodrigues A., Seoane P., Teyssier C., Morel A., Ehrenmann F., Le Provost G., Lalanne C., Noirot C., Klopp C., Reymond I., García-Gutiérrez A., Trontin J.-F., Lelu-Walter M.-A., Miguel C., Cervera M.T., Cantón F.R., Plomion C., Harvengt L., Avila C., Claros M.G. and Cánovas F.M. (2013) *De novo* assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnol. J.*, doi: 10.1111/pbi.12136

Conifer trees are also of great economic importance, as they are the primary source for timber and paper production worldwide. Total timber production in the European Union in 2011 was 427 million m^3 (UNECE, 2013). Approximately 22% was used to produce energy, while the rest was used to supply industrial demands. A study of United Nations Economic Commission for Europe/Food and Agriculture Organization (UNECE/FAO) point out that the future needs in forest biomass to meet the demands of industrial wood as an energy source will exceed production by 2020. The development of a more productive and sustainable forest plantation is essential to meet the increasing demand of wood worldwide together with minimizing environmental impacts (e.g. decreasing pressure on natural forests).

The extant conifers comprise 615 species classified into eight families within the division *Pinophyta* (Farjon, 2010). Some of the most important conifer trees such as pines, spruces and firs are included in the family *Pinaceae*. The genus *Pinus* comprises the largest number of diversified species (113). Maritime pine (*Pinus pinaster* Aiton) is a broadly planted species (4.2 million hectares) in the southwestern part of the Mediterranean Basin, especially along the Atlantic coast in France, Spain and Portugal where it is the dominant species on more than 2.3 million hectares (Sanz *et al.*, 2006). The maritime pine is particularly tolerant to abiotic stresses showing relatively high levels of intraspecific variability (Aranda *et al.*, 2010). The maritime pine is also one of the most genetically studied conifer species for genomic research in Europe (Mackay *et al.*, 2012; Neale and Kremer, 2011), and a large number of genomic resources and phenotypic data have been generated in the last few years and are available for the conifer research community (http://www.scbi.uma.es/sustainpine; https://w3.pierroton.inra.fr/PinusPortal). Furthermore, knowledge gained from studying this conifer species will potentially help to better understand gene function and diversity in closely related, economically significant species but also in other noneconomic but environmentally important gymnosperm species (Neale and Kremer, 2011).

Until recently, advances in the genomics of conifers were hampered by the large size of their genomes ranging from 20 to 40 Gb, which is more than 200-fold the *Arabidopsis* genome and roughly sevenfold the human genome (Mackay *et al.*, 2012; Ritland, 2012). As the conifer genome is extremely large, major research efforts were concentrated on transcriptomic analysis. The first large-scale EST (Expressed Sequence Tag) projects based on Sanger sequencing revealed that transcriptomes of pines and spruces are highly diverse and complex (Allona *et al.*, 1998; Cairney *et al.*, 2006; Li *et al.*, 2009; Pavy *et al.*, 2005; Ralph *et al.*, 2008).

The emergence of the next-generation sequencing (NGS) has profoundly transformed the landscape of genome analysis. Research efforts in conifers have provided growing catalogues of ESTs for several species of economic and ecological importance, including *Picea sitchensis* (Ralph *et al.*, 2008), *Picea glauca* (Rigault *et al.*, 2011), *P. pinaster* (Fernández-Pozo *et al.*, 2011) and *Picea abies* (Chen *et al.*, 2012). Recent landmark studies in conifer genomics reported the draft assembly of the first genome for a gymnosperm species. Specifically, massive parallel sequencing data were used for assembling the 20 Gb genomes of white (Birol *et al.*, 2013) and Norway (Nystedt *et al.*, 2013) spruces.

A full catalogue of genes and protein coding sequences in maritime pine is necessary to improve our understanding of biological functions and evolutionary relationships with other conifer trees and angiosperms. Comparative genomics provide a powerful mean to study gene structure and the evolution of gene function and regulation. Analysis of key genes and pathways allows scientists to better understand how complex biological processes are regulated and evolve (Koonin *et al.*, 2004; Soltis and Soltis, 2003).

In this study, we present results from deep sequencing of the *P. pinaster* transcriptome. The data were used to create a reference transcriptome in this pine species, containing an extensive set of genes expressed in a variety of actively growing tissues/organs (http://www.scbi.uma.es/sustainpinedb/sessions/new). Comparison to available public databases indicates that the transcriptome of *P. pinaster* is similar in size to that of previously characterized *P. glauca* (Rigault *et al.*, 2011) and *P. abies* (Nystedt *et al.*, 2013).

Results

Sequencing strategy and *de novo* assembly

Massive scale cDNA sequencing was performed to define a reference transcriptome for maritime pine. Long reads obtained by the 454 sequencing platform and short reads obtained by the Illumina platform were used for assembly of 210 513 putative unigenes from 18 cDNA libraries constructed from a variety of tree sources, tissues and experimental conditions (Table 1). Figure 1 shows a schematic representation of the workflow followed for preprocessing and assembling, as well as the combination of software algorithms utilized to obtain a final catalogue of unigenes. A total of 6 381 011 long reads provided 4 098 257 useful long reads (64.2%, mean length: 282 bp), and 591 174 069 short reads provided 292 596 546 useful short reads (49.5%, mean length: 86 bp). As shown in Figure 2, the recovery of useful 454 long reads was variable among the different libraries (37%–92%) with high recoveries (88%–92%) in EuroPineDB and UAGPF. Overall, recovery was around 70% of the initial reads. Regarding short reads, only 32% of single-end short reads were kept after data cleaning. The primary reason of out-filtering these single-end sequences was redundancy.

The assembly strategy used a combination of two completely different algorithms with the hypothesis that it may provide superior results than each algorithm by separate. MIRA3 and Euler-SR were selected for long-read assembly because the former is based on overlap-layout-consensus algorithm, while Euler-SR is based on de Bruijn graphs resolved by means of an Eulerian path. Furthermore, each algorithm produced the better results with simulated data (H Benzekri and MG Claros, unpublished results). Regarding short reads, we have used de Bruijn-based algorithms. ABySS was selected for two main reasons: (i) it can efficiently assemble using different k-mers (Lin *et al.*, 2011) and (ii) it is able to assemble vertebrate-sized genomes and transcriptomes (Li *et al.*, 2010). Finally, contigs separately obtained from short read and long read (355 483 in total) were reconciled using CAP3, resulting in 210 513 contigs to describe a *P. pinaster* transcriptome. Notably, 1241 unigenes were longer than 3000 nucleotides. Unigene length ranged from 40 to 7876 bp, with an average of 495 bp and a median of 361 bp.

The complete set of unigenes was compared using blast with the previous *P. pinaster* transcriptome available at EuroPineDB (Fernández-Pozo *et al.*, 2011), and 128 294 unigenes (61%) were not described in this earlier database, while 935 EuroPineDB unigenes were not present in the current unigene collection. When the *P. pinaster* transcriptome was compared with the *P. glauca* and *P. abies* draft genomes, 87.5% and 99.3% of

Table 1 Description of samples used for DNA sequencing

Gene library	Sequencing platform	Sampled plant material	Experimental conditions	SRA code
EuroPineDB	Sanger/454	Bud, xylem, phloem, stem, needles, roots, stem, embryos, callus, cone, male and female strobili	ESTs and SSH libraries from different tissues and conditions as described by Fernández-Pozo et al., 2011	SRS479769
Biogeco1	454	Xylem, bud and needle	ESTs from differentiating xylem, swelling bud and young needles	SRX032960, SRX032961, SRX032962, SRX032963
Biogeco2	454	Bud	EST from quiescent buds harvested on 2-year-old maritime pine (low growing family) in well-watered or drought-stress conditions	SRX031546
Biogeco3	454	Bud	EST from quiescent buds harvested on 2-year-old maritime pine (fast growing family) in well-watered or drought-stress conditions	SRX031589
UAGPF1	454	Embryome	ESTs from developing, immature embryos (1-week maturation)	SRX022618
INIA_PPIN	454	Bud	ESTs from buds	PRJNA221139
U_root	454	Root	ESTs from roots (1-month-old seedlings)	SRS480239
U_tip	454	Root tips	ESTs from root tips (1-month-old seedlings)	SRS480265
U_H	454	Hypocotyl	ESTS from hypocotyl (1-month-old seedlings)	SRS480236
U_N	454	Needle	ESTs from needles (1-month-old seedlings)	SRS480237
U_Cot_Os	454	Cotyledon	ESTs from cotyledons grown under dark conditions	SRS479771
U_H_Os	454	Hypocotyl	ESTs from hypocotyl grown under dark conditions	SRS480236
U_R_6	454	Roots	ESTs from roots (6-month-old seedlings)	SRS480238
U_S_8	454	Stem	ESTs from stem (8-month-old seedlings)	SRS480261
UAGPF2	Illumina	Somatic embryo	Paired-end ESTs from developing, immature embryos (1 week maturation)	SRR609713
BIOGECO4	Illumina	Bud	ESTs from young and aged buds	SRX031587
BIOGECO5	Illumina	Root	ESTs from drought-stressed and control roots in hydropony	SRX031592, SRX031590
BIOGECO6	Illumina	Bud	ESTs from young and aged buds	SRX031594
IBET	Illumina	Zygotic embryo	Paired-end ESTs from embryos	SRS481044

homology was found, respectively, confirming that most assembled unigenes were pine transcripts.

Annotation of unigenes

Unigene annotation was achieved by combining the results of several annotation processes. Each annotation is associated with an *E*-value to enable the empirical assessment of annotation quality. A preliminary analysis of the collection of unigenes using Full-LengtherNext (FLN) revealed that, from the 181 100 unigenes annotated (46.6%), 26 020 were nonredundant transcripts based on orthologue ID. It is also remarkable that 18 667 full-length (FL) unigenes were reconstructed with a mean length of 1495 nucleotides, representing 19% of the total annotated unigenes (Table 2). Of these, 9641 FL unigenes were different, unique genes (9.8% of the total annotated unigenes, and 37.0% of unique unigenes). The frequency distribution of FL unigenes (Figure 3) indicated a high proportion of unigenes ranging from 500 to 1500 nucleotides with the longest transcript being 7876 nucleotides.

Preliminary analysis using FLN also revealed that 111 577 unigenes (53.0%) did not possess significant homology to any other plant gene. This number includes new conifer genes as well as artefactual assemblies. To distinguish between both possibilities, FLN includes a TestCode analysis (Fickett, 1982) and a comparison with the noncoding RNA database (http://www.mirbase.org). As a result, at least 9799 nonredundant coding unigenes can worth a consideration of putative new conifer gene. In fact, 4608 unigenes had a homologue EST in the Pine Gene Index 9.0 database (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=pine). However, only 176 unigenes from this study were determined to be candidate noncoding RNAs. Therefore, the minimal *P. pinaster* transcriptome can be calculated as 26 020 unigenes having unique ID. In addition, it could also be considered the 9799 unigenes without homology but having coding characteristics, plus the 176 noncoding RNAs, that is 35 995 unigenes.

Because this transcriptome was deemed satisfactory, genes annotated as described in Experimental procedures (GO term, a definition or a KEGG code) were subjected to statistical analyses. Of total unigenes, 62.2% (130 845) were annotated, which indicated that the level of annotation was similar to previously published results (Fernández-Pozo *et al.*, 2011). Furthermore, the distribution of GO terms at level 2 of biological process and at level 3 of molecular function (Figure S1) shows that the putative transcriptome covers most important cell functions. A total of 58 296 unigenes possessed unknown sequences that could not be found in existing databases. The annotated transcriptome can be browsed, downloaded and queried at http://www.scbi.uma.es/sustainpinedb/.

Validation of full-length cDNA sequences in the transcriptome database

Full-length cDNA (FLcDNAs) are essential for gene annotation, unambiguous determination of intron–exon boundaries and gene

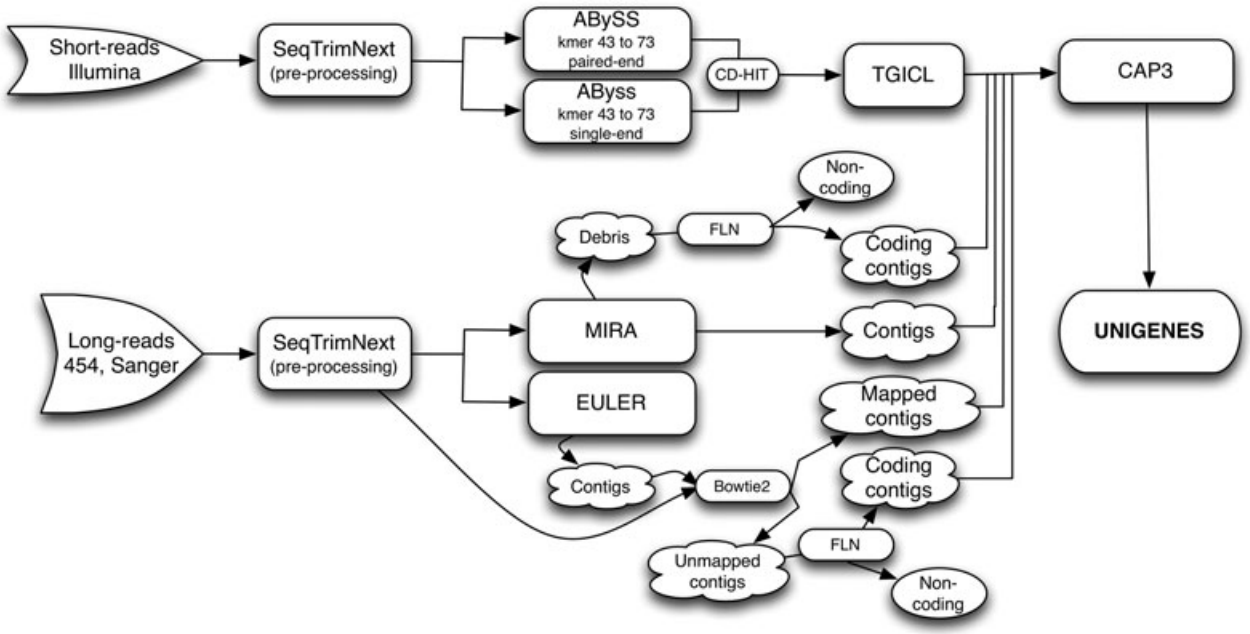

Fig. 1 Schematic representation of 454 and Illumina sequencing, trimming and assembling into unigenes of *Pinus pinaster* transcriptome short read and long read.

functional analysis. To examine the quality of the FLcDNA collection established in the maritime pine transcriptome database, the validation of a number of selected genes coding for proteins of a variety of sizes was advisable. Sequences encoding structural and regulatory genes were selected, appropriate primers based on available sequences were designed, and the corresponding FLcDNAs were amplified by RT-PCR from RNA samples extracted from a variety of maritime pine tissues. Table 3 shows a summary of this work. All selected genes were successfully amplified from cDNA samples using the sequence information available in the database. PCR products fit the theoretical, predicted size in most studied unigenes (Canales *et al.*, 2012; Cánovas *et al.*, 2007; Rueda-López *et al.*, 2013; Villalobos *et al.*, 2012).

Maritime pine regulatory genes

Transcription factors (TF) were specifically searched for in the Sustainpine database. The unique transcripts containing domains of plant TF in maritime pine using Pfam motifs were 877 distributed in 30 families (Table S1). Comparative analysis was performed with other woody plants including white spruce (*P. glauca*), poplar (*Populus trichocarpa*) and grapevine (*Vitis vinifera*), the herbaceous models *Arabidopsis* (*Arabidopsis thaliana)* and rice (*Oryza sativa*), and the model moss *Physcomitrella patens*. The total number of TF in maritime pine was similar to the number previously reported for white spruce (Rigault *et al.*, 2011), suggesting that the information quoted here could be close to the full representation of the maritime pine transcriptome. In addition, the TF gene number in *Physcomitrella* (802) and conifer species (877–892) is smaller than in angiosperms (Table S1), either woody (1337–2499) or herbaceous (1407–1828).

Overall, the relative representation of each TF domain among total TF domains was largely conserved in maritime pine and spruces (Figure 4 and Table S1) with only a few exceptions. For example, the SRF family is under-represented with only four members in maritime pine relative to the 45–71 reported in spruces or the 42–109 members reported in angiosperms. The AP2 family with 106 members is over-represented with regard to spruces (24–63), but contains a similar number of TF than in angiosperms (93–209). In contrast, the histone-like TF (CBF/NF-Y) family is under-represented in maritime pine (26) and Norway spruce (20) when compared to white spruce (52), but exhibits similar numbers to those found in angiosperm species.

When we compared this general distribution in maritime pine to six plant models, we observed conservation of TF categories that can be classified establishing a hierarchical clustering (Figure S2) where a closer conservation in members of each family is observed for *Physcomitrella*, spruce and pine with a second branch for *Arabidopsis*, rice and grapevine and a separate branch for poplar with a greater number of TF. Likely, the recent genome duplication increased the number of TF in each family of poplar (Tuskan *et al.*, 2006). Families not found in maritime pine and previously reported in white spruce are as follows: RNA polymerase II TF SIII, RNA polymerase III, TF IIIC subunit and alcohol dehydrogenase TF Myb/SANT-like (Rigault *et al.*, 2011).

Enzyme-encoding genes

The representation of gene family members involved in various metabolic pathways was determined, and data retrieved from the SustainpineDB were compared with those found in Norway spruce (*P. abies)*, *P. trichocarpa* and *A. thaliana* (Figure 5). A first set of genes studied were those participating in nitrogen acquisition and assimilation. The number of genes encoding nitrate reductase and nitrite reductase was similar in the three species (1–2). For genes involved in inorganic nitrogen transport, some important differences were observed among species. For instance, *Arabidopsis* contains almost twice genes encoding nitrate transporters that maritime pine, Norway spruce and poplar. For ammonium transporters, only the poplar genome presents an expanded gene family of 14 members.

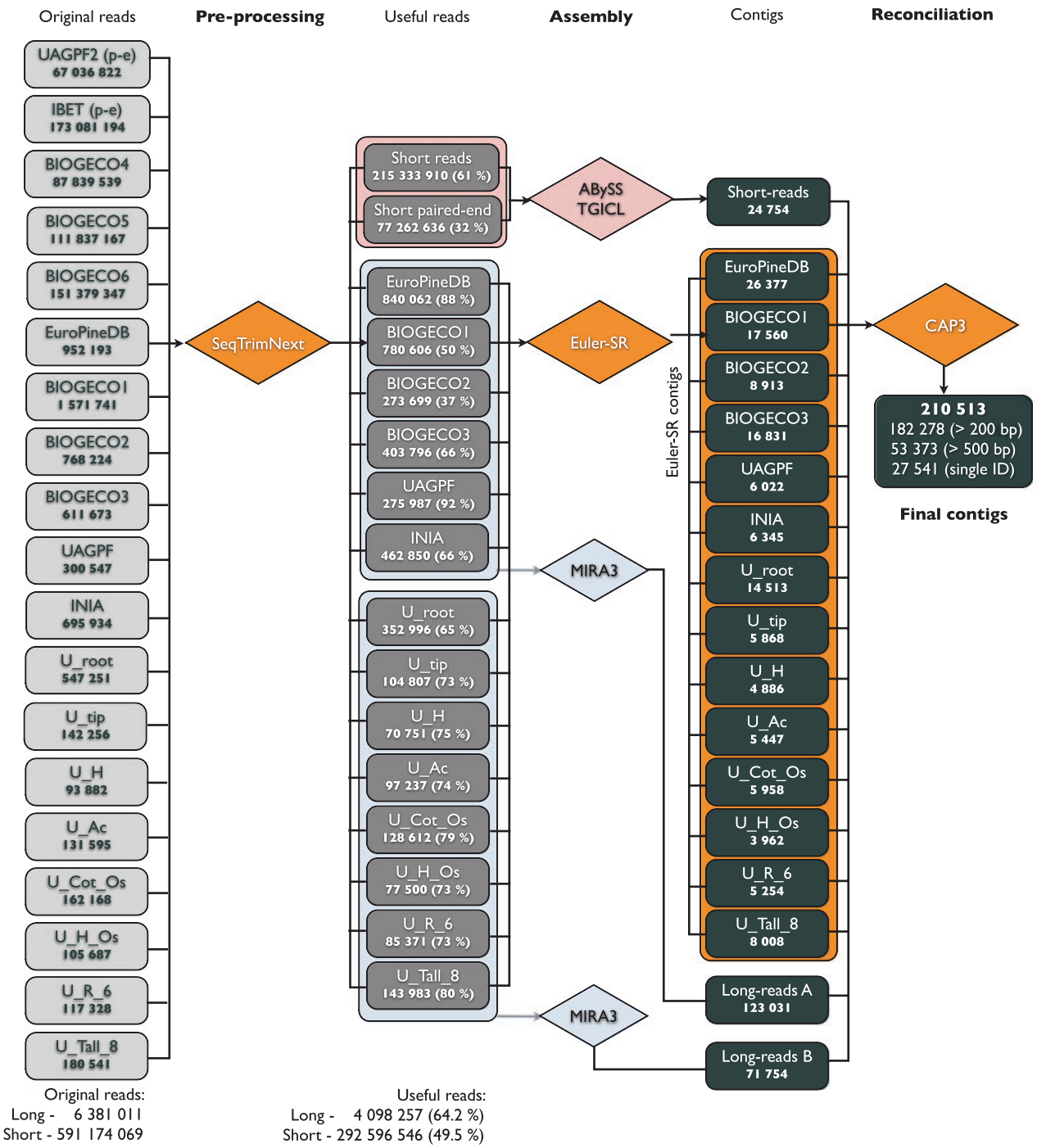

Fig. 2 Flow chart showing preprocessing into useful reads, assembly into contigs and overlap-based reconciliation into final unigenes of sequenced data from 5 (591 174 069 short reads, Illumina) or 14 (6 381 011 long reads, 454) cDNA libraries in maritime pine.

Fewer transcripts for genes encoding enzymes of ammonium assimilation were found in conifer species. Only 2–3 transcripts for glutamine synthetase (GS) were identified in gymnosperms (*P. pinaster* and *P. abies*), in accordance with previous results (Cánovas *et al.*, 2007). In contrast, genomes of angiosperm species are endowed with GS families with a higher number of members, eight in *Populus* and six in *Arabidopsis*. A single expressed gene was found for ferredoxin-glutamate synthase (Fd-GOGAT) and NADH-GOGAT in maritime pine and Norway spruce. In contrast, two genes encode Fd-GOGAT and NADH-GOGAT in poplar.

A second group of genes where those encoding enzymes involved in synthesis of methionine and S-adenosylmethionine (SAM), the activated form of methionine, which participate in a number of essential metabolic pathways in plants. In particular, we focused on three genes involved in the synthesis and recycling of SAM, a methyl donor in multiple cellular transmethylation reactions (Figure 5). The number of genes encoding cobalamin-independent

Table 2 Summary of final data for the *P. pinaster* transcriptome

	Absolute number	%
Unigenes	210 513	100.0
Artefacts	2010	0.95
Unigenes after resolving artefacts	209 928	99.7
Unigenes > 200 nt	181 100	86.0
Unigenes > 500 nt	52 550	25.0
Unigenes > 3000 nt	1241	0.59
Unigenes with orthologue*	98 175	46.6
Different orthologue ID	**26 020**	26.5
Complete transcripts (full length)	18 667	19.0
Different full-length transcripts	9641	9.8
Putative ncRNAs	**176**	0.08
Unigenes without orthologue*	111 577	53.0
Coding	29 736	26.6
Putative coding	23 545	21.1
Nonredundant coding	**9799**	8.8

Numbers in bold can be considered the representative amount of unigenes of the category.

*Percentages for this classification are calculated using this file as 100.0%.

methionine synthase and SAM synthase was identical in conifers and *Arabidopsis* but slightly higher in *P. trichocarpa*, whereas an additional S-adenosyl-L-homocysteine hydrolase gene was found in pine and Norway spruce in regard to angiosperms.

Cellulose, hemicellulose and lignin are major components of the secondary cell wall, a well-developed structure in woody perennials compared with *Arabidopsis,* particularly in trees such as maritime pine and poplar. A similar number of genes encoding the cellulose synthase catalytic subunit (CesA) were found in conifers compared with *Arabidopsis*, whereas the number increased to 18 in poplar. The number of genes encoding sucrose synthase and UDP-glucose 6-dehydrogenase in the maritime pine transcriptome was slightly lower compared with the other three plant species.

We also compared the number of genes encoding three key enzymes of the phenylpropanoid pathway that direct the carbon flow from the shikimate pathway to monolignol biosynthesis: phenylalanine ammonia-lyase, cinnamoyl-CoA reductase and cinnamyl alcohol dehydrogenase (Figure 5). The number of genes encoding the three enzymes was similar in maritime pine, Norway spruce and *Arabidopsis* and also in poplar in the case of phenylalanine ammonia-lyase. In contrast, the members of the cinnamoyl-CoA reductase and cinnamyl alcohol dehydrogenase families are considerably higher in poplar (40 and 21, respectively).

Single nucleotide polymorphism and simple-sequence repeat identification

Using CLCBio with stringent criteria both in terms of minimum allele frequency (MAF = 20%) and contig depth (minimum of 10 reads), we identified 55 607 in silico Single nucleotide polymorphism (SNPs) in 13 892 unigenes (Table S2), corresponding to an average of 1 SNPs per 2 kb surveyed in the assembled transcriptome. Setting the contig depth parameter to 4 or 10, yielded 542 830 or 20 208 SNPs, respectively, with GigaBayes. Suprisingly, only 10 829 and 2106 SNPs were simultaneously found between CLCBio and GigaBayes with relaxed (minimum contig depth, CRL = 4) and more stringent (CRL=10) detection criteria, respectively; corresponding to 2% and 10.4% of common SNPs with CLCBio. Despite a fivefold increase in commonly detected SNPs using the more stringent detection criteria in GigaBayes, both software detected highly divergent sets of SNPs. Given this inconsistency, we used a set of 5138 SNPs validated in natural populations and mapping pedigrees of maritime pine (Lepoittevin *et al.*, 2010; Chancerel *et al.*, 2011; Chancerel *et al.,* 2013) to evaluate the ability of both software to identify true SNPs. Prior to this analysis, we checked that the flanking sequences associated with the true SNPs were found in the present assembly using BLASTn analysis. Using CLCBio, 36.26% (1863 SNPs) of true SNPs were detected, while this rate dropped to 9.17% (471 SNPs) and 1.36% (70 SNPs) using relaxed and stringent GigaBayes criteria. By normalizing

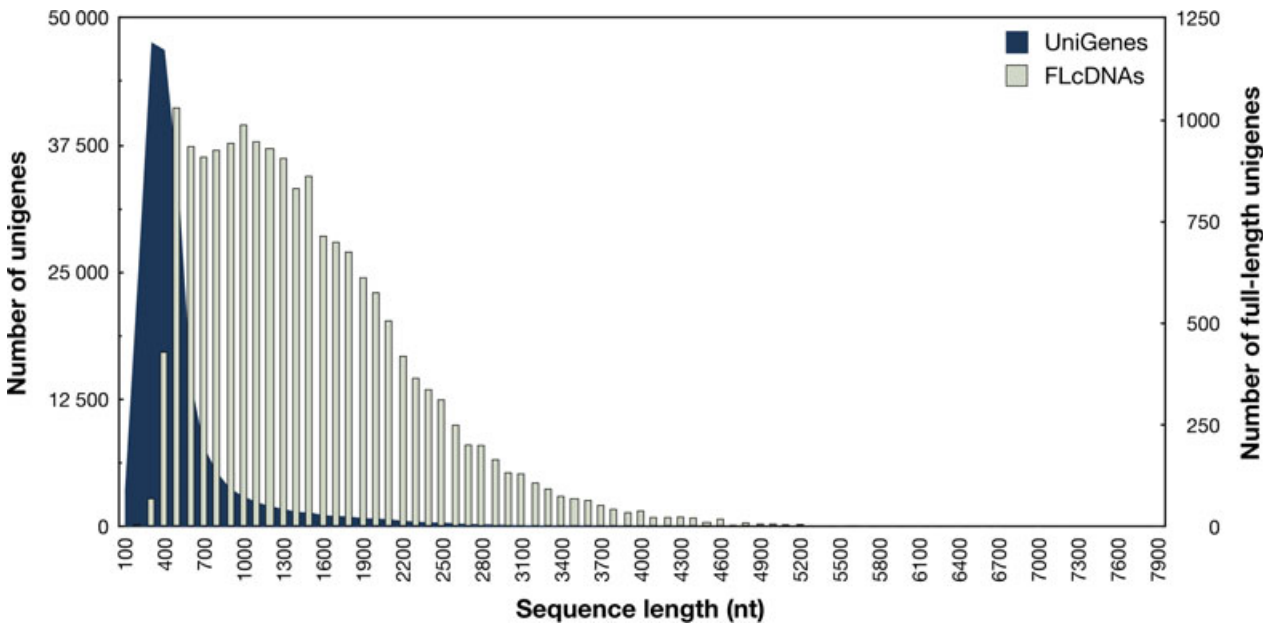

Fig. 3 Frequency distribution of total unigenes and full-length unigenes.

Table 3 Summary of FLcDNA used for sequence assembly validation by PCR amplification

Gene name	Theoretical size of ORF from assembly (bp)	Experimental size of ORF (bp)	Accession number*
Arginase	1026	1026	sp_v3.0_unigene23824
Xyloglucan endotransglycosylase	860	860	sp_v3.0_unigene29476
Phenylalanine ammonia-lyase	2265	2265	sp_v3.0_unigene17298 AY641535
PII-protein	714	732	sp_v3.0_unigene23578 AJ489604
Asparagine synthetase 1	1782	1782	sp_v3.0_unigene14147 HQ625490
Asparagine synthetase 2	1770	1773	sp_v3.0_unigene18231 HM222940
Glutamate decarboxylase	1530	1530	sp_v3.0_unigene11755 DQ125946
Glutamate dehydrogenase	1236	1236	sp_v3.0_unigene15901 HM222941
Sucrose synthase	1914	1914	sp_v3.0_unigene34880
Ammonium transporter 1.1	1584	1584	KC807907
Ammonium transporter 1.2	1464	1464	KC807908
Ammonium transporter 1.3	1539	1539	KC807909
Ammonium transporter 2.1	1461	1461	KC807910
Ammonium transporter 2.3	1446	1446	KC807911
Glutamine dumper 1	345	345	sp_v3.0_unigene97635
Glutamine dumper 2	384	384	sp_v3.0_unigene20421
†MYB1	1023	1023	sp_v3.0_unigene29297 EU482890
†MYB4	946	946	sp_v3.0_unigene127348
†MYB 8	1605	1605	FN868598
‡Dof2	1065	1065	KC688677
‡Dof3	1047	1047	KC688678
‡Dof4	1053	1053	KC688679
‡Dof6	907	907	KC688680
‡Dof7	1404	1404	KC688681
‡Dof8	822	822	KC688682
‡Dof9	897	897	KC688683
§NAC1	1559	1545	sp_v3.0_unigene7635
§NAC2	1476	1398	sp_v3.0_unigene14173
§NAC3	1457	1388	sp_v3.0_unigene18613
§NAC4	1630	1278	sp_v3.0_unigene20354
§NAC5	1824	1180	sp_v3.0_unigene1398
Methionine synthase	2301	2301	sp_v3.0_unigene34452 HE566045
S-adenosylmethionine synthase	1176	1176	sp_v3.0_unigene3027 HE574556
S-adenosylhomocysteine hydrolase	1458	1458	sp_v3.0_unigene19530 HE574555
Methylenetetrahydrofolate reductase	1785	1785	sp_v3.0_unigene15993 HE574560
Caffeate O-methyltransferase	1161	1095	sp_v3.0_unigene17184 HE574557
Hydroxycinnamoyl-CoA: shikimate hydr oxycinnamoyl transferase	1302	1302	sp_v3.0_unigene8683 HE574565
Glycine decarboxilase complex H-protein a	504	504	sp_v3.0_unigene30488 HE574553
Glycine decarboxilase complex H-protein b	516	516	sp_v3.0_unigene126824 HE574563
Mitochondrial serine hydroxymethyltransferase	1572	1572	sp_v3.0_unigene439 HE574554

Table 3 Continued

Gene name	Theoretical size of ORF from assembly (bp)	Experimental size of ORF (bp)	Accession number*
Cytosolic serine hydroxymethyltransferase	1413	1413	sp_v3.0_unigene17057 HE574564
D-3-Phosphoglycerate dehydrogenase	1947	1947	sp_v3.0_unigene543 HE574561
3-Phosphoserine aminotransferase	1302	1302	sp_v3.0_unigene37851 HE574562
Pinoresinol-lariciresinol reductase	939	939	sp_v3.0_unigene17681 HE574558
Phenylcoumaran benzylic ether reductase	927	927	sp_v3.0_unigene31659 HE574559
Phenylpropenal double-bond reductase	1056	1056	sp_v3.0_unigene22698 HE575885

*Accession number of unigene in Sustainpine and GenBank.
†MYB family of TF.
‡Dof family of TF.
§NAC family of TF.

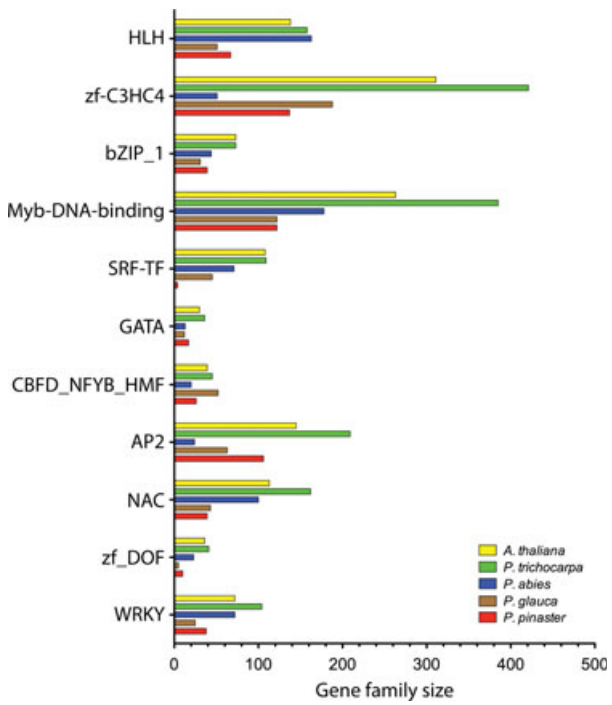

Fig. 4 Distribution of unique transcripts corresponding to TF gene families in *Pinus pinaster* and comparison to other plant transcriptomes. The number of different encoded transcripts with the conserved DNA-binding domain of each family is represented. The distribution of TF gene families in *P. pinaster*, *Picea glauca, Picea abies, Populus trichocarpa and Arabidopsis thaliana* is compared.

Gygabayses detection rates to the 55 607 putative SNPs identified by CLCBio, the relaxed and stringent GigaBayes detection procedures led to a success rate of 0.94% (48 true SNPs detected) and 3.75% (193 SNPs), respectively. In conclusion, CLCBio was found to perform nearly 10 times better than the most stringent GigaBayes SNP detection procedure with our data set.

A total of 5974 putative simple-sequence repeat (SSRs) were found, with trinucleotide repeats (3309) being the most common, and dinucleotide repeats (479) the less abundant. This is in agreement to previously published *P. pinaster* SSR abundance (Fernández-Pozo *et al.*, 2011).

Discussion

Maritime pine transcriptome assembly

Long-read sequence data sets are required for transcriptome assembly in nonmodel species for which a reference genome is not available. In conifers, 454 sequencing has been recently used to generate well-defined transcriptomes in several species of ecological and economic interest, that is, *Pinus contorta* (Parchman *et al.*, 2010), *P. glauca* (Rigault *et al.*, 2011), *P. pinaster* (Fernández-Pozo *et al.*, 2011), *Pinus taeda* and 11 other conifers (Lorenz *et al.*, 2012). In the present work, we used a combination of 454 and Illumina sequencing to define a minimal reference transcriptome for maritime pine (*P. pinaster*). A similar approach was recently used to characterize, for example, the globe artichoke transcriptome (Scaglione *et al.*, 2012). The nonredundant transcriptome resulting from the assembly contains 26 020 unique transcripts with orthologue ID in public databases, a number very close to the 27 720 unique cDNA clusters reported for the *P. glauca* transcriptome (Rigault *et al.*, 2011) and higher than the 17 000 unique coding genes obtained in the assembly of *P. contorta* transcriptome (Parchman *et al.*, 2010). The number of unique transcripts in maritime pine is also close to the number of genes (28 354) resulting from the draft assembly of the 20-gigabase genome of *P. abies* (Nystedt *et al.*, 2013). Considering all the available data, an elevated coverage of the maritime pine transcriptome is estimated.

FLcDNAs catalogues as genomic resources

The availability of large collections of FLcDNAs in several conifers, such as Sitka (Ralph *et al.*, 2008) and white spruces (Rigault *et al.*, 2011), as well as *Cryptomeria* (Futamura *et al.*, 2008), has greatly facilitated the assembly and annotation of FLcDNAs in maritime pine. FLcDNAs are crucial for accurate

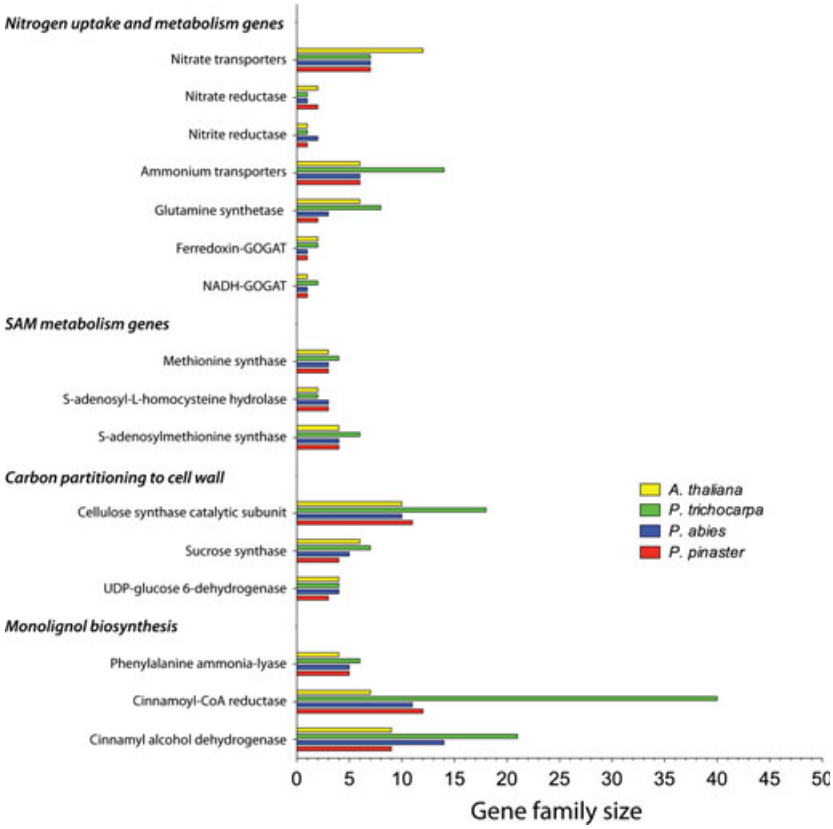

Fig. 5 Comparison of gene families for relevant enzymes in *Pinus pinaster, Picea abies, Populus trichocarpa* and *Arabidopsis thaliana.* The following databases were used in addition to *Sustainpine*DB*: P. abies* v1.0, *P. trichocarpa* v3.0, *A. thaliana* TAIR 10.

annotation, comparative analysis with other conifer species and also for functional analysis of relevant genes associated to maritime pine growth, development and response to environmental changes. Furthermore, this genomic resource will greatly facilitate protein identification as well as protein–protein interaction studies through proteomics approaches (Cánovas *et al.*, 2004). For all these reasons, it was of paramount importance to validate the assembly of the FLcDNA collection (9641 different transcripts).

Over the last few years, refined protocols have been developed for *Agrobacterium*-mediated genetic transformation of maritime pine embryogenic tissue, cryopreservation of transgenic lines and efficient transgenic plant regeneration through somatic embryogenesis (reviewed in Trontin *et al.*, 2013). These new developments and the availability of a large collection of FLcDNAs have paved the way for the application of reverse genetics towards functional dissection of traits of economic and ecological interest in maritime pine trees. Thus, the availability of both standardized transformation methods and FLcDNA catalogues is expected to significantly increase throughput in candidate gene analysis together with facilitating comparison across laboratories interested in maritime pine genomics. The functional analysis of key genes is crucial for future applications in tree improvement, new variety design and sustainable forest management (e.g. development of marker-assisted selection).

Maritime pine gene families and genome size

It has been suggested that the increased size and complexity of conifer genomes relative to angiosperms may be explained by the existence of large gene families (Kinlaw and Neale, 1997). However, this assumption is not fully supported by available data as most TF (Figure 4 and Table S1) or other gene families (Figure 5) present in maritime pine (this work) or spruce genomes (Birol *et al.*, 2013; Nystedt *et al.*, 2013; Rigault *et al.*, 2011) were of similar or even lower size compared with angiosperm species (*P. trichocarpa*, *A. thaliana* and *V. vinifera*). Meanwhile, the existence of large gene families in conifers coding for enzymes of secondary metabolism has been reported (Martin *et al.*, 2004), there are other families in primary metabolism that contain similar, or even shorter, number of functional members in conifers than in angiosperms (Cánovas *et al.*, 2007). High genome size and complexity in conifers may be more readily explained by divergence and accumulation of retrotransposons and pseudogenes (Morse *et al.*, 2009; Nystedt *et al.*, 2013). Retrotransposons and pseudogenes can be expressed and contribute in some extent to the collection of unigenes without orthologue in the maritime pine transcriptome. Accumulation of pseudogenes may have functional advantages in the regulation of gene expression if they are expressed. Recently, Poliseno *et al.* (2010) reported that expressed pseudogenes compete with authentic target transcripts for miRNA binding and, as such, modulate expression levels of their cognate genes.

Transcription factors

The identification of transcription factors (TF) and subsequent analysis of the composition and organization of TF families are necessary steps to understand the regulatory networks associated with key processes in conifer trees. The number of TF in *P. pinaster* appears to be similar to that of *P. glauca* (Rigault *et al.*, 2011), but considerably lower compared with those found in the genomes of several angiosperms. This fact is confirmed by studies carried out in specific families. For example, the Dof family has only ten members in maritime and loblolly pines (Rueda-López *et al.*, 2013); this is twofold to eightfold lower than gene

number in angiosperms, including 36 in *A. thaliana*, 30 in rice (Lijavetzky *et al.*, 2003), 22 in grape or 46 in poplar (Pérez-Rodríguez *et al.*, 2010). Moreover, it has also been argued that some subfamilies such as Knox I involved in meristem cell identity have pursued a distinct path of evolution in conifers. Three of the four groups delineated by phylogenetic analyses in angiosperms have no conifer sister groups (Guillet-Claude *et al.*, 2004), suggesting that conifers evolved in a nonlinear fashion compared with angiosperms. Recently, it has been proposed that expansion of the VASCULAR NAC DOMAIN (VND) gene family might be related to xylem vessel complexity in angiosperms (Nystedt *et al.*, 2013). Other example of this divergent pattern of evolution for TF is found in the HD-Zip III gene family involved in regulation of cambium, and primary and secondary vascular differentiation whose structure also has diverged considerably between angiosperms and gymnosperms (Côté *et al.*, 2010). Furthermore, distinct evolutionary trends in angiosperms and gymnosperms are evident by differential family gene expansion of subgroup 4 R2R3-Myb with more recent duplications in *P. glauca* (Bedon *et al.*, 2010). In this particular case, the higher number of members in conifers appears to be related with isoprenoid- and flavonoid-oriented stress responses.

Most of the existing knowledge of plant TF genes was obtained from studies in *Arabidopsis.* While *Arabidopsis* is a useful model for many developmental and environmentally regulated processes in higher plants, it lacks certain traits that are of immense value to agriculture or forestry. To better understand the evolutionary relationship and dependence of transcriptional regulation and morphological complexity in *Viridiplantae* is important to analyse how particular families of transcription factors did expand in correlation with the general increase in morphological complexity. In this sense, the availability of full-length sequences from gymnosperms come to close a huge gap and will help to understand how the land plants managed, in terms of transcriptional regulation, to become multicellular.

Gene families of metabolic pathways

A major contrast was observed for genes involved in nitrogen metabolism likely related with adaptations to nitrogen availability and use. For example, poplar, Norway spruce and pine possess almost half the number (7) of nitrate transporters encoded in the *Arabidopsis* genome (12). This difference may be due to differences among species on soil nitrate uptake and transport during active growth, compared with trees in which active growth is primarily supported by internal nitrogen storage and mobilization (Cantón *et al.*, 2005). In contrast, the family of ammonium transporters is expanded in *P. trichocarpa* (14) compared with pine, spruce and *Arabidopsis* (6). It has been suggested that this specific feature may be related to the peculiar physiology of a perennial and mycorrhizal tree (Couturier *et al.*, 2007). However, the observation of similar number of ammonium transporters in conifer trees and *Arabidopsis* does not support the above assumption anymore. Another major difference is the reduced number of the GS gene family in pine (2) and spruce (3) compared with *Populus* (8) and *Arabidopsis* (6). Deep transcriptome analysis in the current study supports previous reports describing only two genes for cytosolic GS pine and the lack of a plastidic isoform (Cánovas *et al.*, 2007).

In trees, consumption of methyl units during lignification implies the existence of an important carbon sink, and the S-adenosylmethionine availability may affect wood quality through alterations in lignin content and composition (Villalobos *et al.*, 2012). Nevertheless, no major differences were found in the number of genes in poplar and conifers with respect to *Arabidopsis*.

The number of genes related to carbon partitioning to cellulose and hemicellulose seems to be similar in the four species, with the exception of CesA that is expanded in *P. trichocarpa*. The *CesA* gene family in poplar includes 18 members that form nine phylogenetic groups, eight of which contain a pair of *CesA* genes with nearly identical sequence as a likely result of recent genome duplication. Some of them showed redundant expression xylem-specific, and it was suggested that more *CesA* genes could be required for the massive synthesis of cellulose in trees (Suzuki *et al.*, 2006). However, the number of genes encoding *CesA* subunits in *P. pinaster, P. abies* and *A. thaliana* seems to be similar.

Monolignol biosynthesis is a complex branch of the general phenylpropanoid pathway. The number of genes encoding phenylalanine ammonia-lyase was similar in all species (4–6). However, an increased number of genes encoding cinnamoyl-CoA reductase were found in *P. trichocarpa*, *P. pinaster and P. abies* compared with *Arabidopsis*. Previous work in angiosperms showed that cinnamoyl-CoA reductase family represents the largest lignin biosynthesis gene family in several species (Xu *et al.*, 2009). In conifers, transcriptomic data reported here suggest that the cinnamoyl-CoA reductase family is also expanded. In contrast, it has been previously reported that conifers may have a single copy of the cinnamyl alcohol dehydrogenase gene (Mackay *et al.*, 1997). However, different transcripts were found in maritime pine and Norway spruce.

Overall, gene families involved in metabolic pathways were of similar size in *P. pinaster* and *P. abies*. The number of genes in *P. pinaster* was either similar or lower to those existing in *Populus* or *Arabidopsis*. These findings should be confirmed when a reference genome for pine is available.

SNP identification

Based on previous unigene sets constructed with significantly fewer sequences obtained essentially from one ecotype (that of the Aquitaine provenance), highly multiplexed SNP arrays were constructed in maritime pine for linkage applications (Chancerel *et al.*, 2011, 2013) and association mapping (Lepoittevin *et al.*, 2012). The present study provides the most comprehensive SNP catalogue for maritime pine bringing together sequences from a diverse range of ecotypes (France, Spain, Portugal) and therefore allowing the use of high throughput genotyping technologies for applications in: i) genetic diversity and population structure analysis, with less ascertainment bias than was observed from the analysis of previous SNP data sets (Chancerel *et al.,* 2011) thank to a better coverage of the species diversity; and, ii) genomic selection, because about 14k different gene-loci are now covered by at least one SNP. Our study reveals a strong influence of the SNP calling algorithm on the number of detected SNPs. Such variability in the number of called SNPs between SNP detection software has been already reported in another conifer species (Muller *et al.*, 2012) and indicates that results should be interpreted with caution, especially if based on a single detection approach. In our case, using a set of already validated SNPs, we show that CLCBio was able to detect true SNPs at a rate that was ten times higher than the most stringent criteria implemented in GigaBayes. Trees have long generation times, and breeding (especially for pine) is a slow process (Mullin *et al.*, 2011). Genomic selection offers the possibility to increase genetic gain per time unit

in these long-lived organisms as illustrated recently for *P. taeda* (Resende *et al.*, 2012) and accelerate their domestication (Harfouche *et al.*, 2012). Genomic selection will also permit to control more precisely that sufficient level of diversity is maintained in future varieties to allow forest trees to cope with major biotic and abiotic constraints in rapidly evolving environment.

Concluding remarks

In this work, comprehensive characterization of the *P. pinaster* transcriptome was performed using a combination of two different next-generation sequencing platforms, 454 and Illumina. The *de novo* assembly of the maritime pine transcriptome provides a large catalogue of expressed genes (26 020 unigenes with orthologue, 9799 unigenes with coding characteristics but without orthologue, 176 ncRNA) and a relevant collection of FLcDNAs (9641). In addition, the sequencing data permitted the identification and establishment of robust SSR- and SNP-based databases for genotyping applications and translational integration in maritime pine breeding programmes. These genomic resources will facilitate genome sequencing, functional genomics and applied studies in maritime pine trees.

Experimental procedures

Tree source, tissues and experimental conditions

Pinus pinaster samples of developing xylem were collected from different genotypes of a Corsican clonal population planted in 1986 at the forestry station of INRA-Pierroton (Aquitaine, France), as described (Villalobos *et al.*, 2012). Miscellanea of maritime pine tissues (cones, male and female strobili, buds, xylem and phloem) were collected from adult trees located at Sierra Bermeja (Málaga). Roots, stems and needles of 2-week-old pine seedlings and zygotic embryos were also sampled as well as maritime pine embryonal masses. Somatic embryos were produced from embryonal masses line (AAY06006) of *P. pinaster* originating from a Landes × Morocco polycross. Embryonal masses were induced from immature zygotic embryo (Park *et al.*, 2006). Embryonal masses and somatic embryos were cultured according to Lelu-Walter *et al.* (2006). Zygotic embryos at several developmental stages isolated from immature seeds were also used. All samples were frozen in liquid nitrogen and stored at −80 °C untill further use.

RNA isolation, cDNA synthesis and construction of libraries

Total RNA was isolated as described by Canales *et al.* (2012). Double-stranded cDNA libraries were constructed using the MINT kit (Evrogen) with incorporated modifications proposed by Babik *et al.* (2010). The quality of cDNA libraries was checked on 1.5% agarose gels and Agilent 2100 Bioanalyzer. cDNA libraries were also constructed from poly(A) enriched RNA. Briefly, equal amounts of total RNA (50 μg) from different tissues were combined prior to mRNA purification. Poly(A)+ mRNA was double purified using the Oligotex mRNA Mini Kit (Qiagen), according to the manufacturer's instructions. First and second strand cDNA synthesis was conducted from 10 μg of mRNA following the protocol described by Durban *et al.* (2011). Five cDNA synthesis reactions were performed to obtain sufficient cDNA quantity for 454 pyrosequencing. RNA isolation, purification and library construction from somatic embryos were carried out by GATC company (Constanz, Germany) using Smart cDNA construction kit and standard procedures for Illumina GA IIx sequencing and using standard kit for construction of normalized cDNA library dedicated to Roche/454 GS FLX sequencing.

DNA sequencing and preprocessing analysis

Libraries were sequenced using different sequencing platforms, including Roche/454 Titanium, Roche/454 GS FLX, Illumina GAIIx and Illumina HiSeq 2000. Chloroplast, mitochondrial and ribosomal sequences in short reads were filtered out by reference mapping using BWA (Li and Durbin, 2009). All reads were preprocessed using the SeqTrimNext pipeline (http://www.scbi.uma.es/seqtrimnext; Falgueras *et al.,* 2010) available at the Genotoul's computer facilities (INRA Toulouse, France) and at the Plataforma Andaluza de Bioinformática (University of Málaga, Spain). Low quality, ambiguous and low complexity stretches, linkers, adaptors, vector fragments, organelle DNA, polyA/polyT tails, and contaminated sequences were removed while keeping the longest informative part of the read, discarding sequences below 20 (short reads) or 40 bp (long reads). The command line for long reads was *seqtrimnext -t transcriptomics_plants.txt -Q input_reads.fastq > output.txt*, where the configuration file *transcriptomics_plants.txt* is provided by default with SeqTrimNext and contains the configuration parameters. The command line for short reads was the same than above, but configuration file also indicates to filter out sequences shorter than 20 pb, skip preliminary statistic calculations and skip the read clonality analysis.

Transcriptome assembly

The bioinformatics process for obtaining the putative unigenes using the SeqTrimNext preprocessed reads is detailed in Figure 1. Short reads and long reads were assembled using well-established protocols (see Methods S1). The complexity and redundancy of short read and long read assemblies was reduced using CAP3 assembler (Liang *et al.*, 2000) to obtain the unigenes. A 95% cut-off for overlap per cent identity was applied to cope with both sequence variation (high heterozygosity of pine genes) and genome heterogeneity between samples as previously described (Fernández-Pozo *et al.*, 2011).

Annotation analysis

Unigenes were analysed using Sma3s (Pérez-Pulido, A., Muñoz-Mérida, A., Claros, M.G., Trelles, O., submitted; http://www.bioinfocabd.upo.es/node/11) with the command line *sma3s_v2.pl -a 123 -d uniprot_plants.dat -i Unigenes.fasta -b Unigenes.blast -p F > output* using as input the unigenes (.fasta file), the blast result of unigenes against the plant UniProtKB database (.blast file) and the metadata of the plant UniProtKB database (.dat file) to provide a gene description, GO terms, EC keys, KEGG maps and InterPro codes for every sequence. They were also analysed with FLN to provide gene description, identify, which unigenes corresponded to FLcDNAs, detect putative start and stop codons as well as the putative protein sequence, extract, which unigenes could be sRNAs, and obtain a quick preview of the pine unigene content. AutoFact (Koski *et al.*, 2005) was also used to provide a third gene description.

Identification of polymorphisms

Because nucleotide variation is considered frequent in plant genes, we screened the unigenes for SNPs and SSRs variation. SSRs were screened using MREPS (http://bioinfo.lifl.fr/mreps/; Kolpakov *et al.*, 2003) with default parameters. SNP detection

was performed based on 4 098 953 Roche/454 and Sanger reads using either CLC-Bio Workbench, v6.0 (CLC Bio, Aarhus, Denmark) or Gygabayes (http://bioinformatics.bc.edu/marthlab/wiki/index.php/Software), an implementation and expansion of the PolyBayes SNP detection algorithm from Marth *et al.* (1999). With respect to the former, we first used the reference mapping function to map the reads onto the 210 513 unigenes generated in the present study. The following parameters were used: similarity score = 80%, minimum length fraction = 90% and maximum number of hits for reads = 10. Then, we used the SNP detection function, based on the neighbourhood quality standard (NQS) algorithm, using the following parameters: minimum coverage = 10, minimum central base quality = 20, minimum neighbourhood quality over a window length of 11 nucleotides = 15, maximum gap and mismatch count = 2 and minimum allele frequency = 20%. With respect to the latter, we used the following parameters: (ploidy = diploid; CRL = 4 or 10; CAL2 = 2; PSL=0.9; O = 1; D = 0.001). In both software, no insertion/deletion variants (InDels) were considered.

Acknowledgements

We are indebted to the anonymous reviewers for their thorough evaluation and constructive recommendations that helped to improve this manuscript. We would like to thank Aranzazu Flores Monterroso, Remedios Crespillo, José Vega Bartol and Marta Simões for help in the preparation of samples. This work was supported by the SUSTAINPINE Project funded by the Plant KBBE programme, Scientific and Technological Cooperation in Plant Genome Research (PLE2009-0016). Zygotic embryo data production was supported by FCT (Portugal) through projects PTDC/AGRGPL/102877/2008 and P-KBBE/AGR-GPL/0001/2009, and grants PEst-OE/EQB/LA0004/2011 and SFRH/BD/79779/2011 (AR). Alexandre Aguiar and Isabel Carrasquinho from Instituto Nacional de Investigação Agrária e Veterinária (INIAV) are acknowledged for making zygotic embryos available. Somatic embryo data production was supported by the EMBRYOME project, contract number 33639, funded by the french 'Conseil Régional de la Région Centre'.

References

Allona, I., Quinn, M., Shoop, E., Swope, K., St Cyr, S., Carlis, J., Riedl, J., Retzel, E., Campbell, M.M., Sederoff, R. and Wetten, R.W. (1998) Analysis of xylem formation in pine by cDNA sequencing. *Proc. Natl Acad. Sci. USA,* **95**, 9693–9698.

Aranda, I., Alía, R., Ortega, U., Dantas, Â.K. and Majada, J. (2010) Intra-specific variability in biomass partitioning and carbon isotopic discrimination under moderate drought stress in seedlings from four *Pinus pinaster* populations. *Tree. Genet. Genome*, **6**, 169–178.

Babik, W., Stuglik, M., Qi, W., Kuenzli, M., Kuduk, K., Koteja, P. and Radwan, J. (2010) Heart transcriptome of the bank vole (*Myodes glareolus*): towards understanding the evolutionary variation in metabolic rate. *BMC Genomics,* **11**, 390.

Bedon, F., Bomal, C., Caron, S., Levasseur, C., Boyle, B., Mansfield, S.D., Schmidt, A., Gershenzon, J., Grima-Pettenati, J., Séguin, A. and MacKay, J. (2010) Subgroup 4 R2R3-MYBs in conifer trees: gene family expansion and contribution to the isoprenoid-oriented responses. *J. Exp. Bot.* **61**, 3847–3864.

Birol, I., Raymond, A., Jackman, S.D., Pleasance, S., Coope, R., Taylor, G.A., Yuen, M.M., Keeling, C.I., Brand, D., Vandervalk, B.P., Kirk, H., Pandoh, P., Moore, R.A., Zhao, Y., Mungall, A.J., Jaquish, B., Yanchuk., A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., Mackay, J., Bohlmann, J. and Jones, S.J. (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics,* **29**, 1492–1497 doi:10.1093/bioinformatics/btt178.

Bowe, L.M., Coat, G. and dePamphilis, C.W. (2000) Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA,* **97**, 4092–4097.

Cairney, J., Zheng, L., Cowels, A., Hsiao, J., Zismann, V., Liu, J., Ouyang, S., Thibaud-Nissen, F., Hamilton, J., Childs, K., Pullman, G.S., Zhang, Y., Oh, T. and Buell, C.R. (2006) Expressed sequence tags from loblolly pine embryos reveal similarities with angiosperm embryogenesis. *Plant Mol. Biol.* **62**, 485–501.

Canales, J., Rueda-López, M., Craven-Bartle, B., Avila, C. and Cánovas, F.M. (2012) Novel insights into regulation of asparagine synthetase in conifers. *Front. Plant Sci.* **3**, 100.

Cánovas, F.M., Dumas-Gaudot, E., Recorbet, G., Jorrin, J., Mock, H.-P. and Rossignol, M. (2004) Plant proteome analysis. *Proteomics*, **4**, 285–298.

Cánovas, F.M., Avila, C., Cantón, F.R., Cañas, R.A. and de la Torre, F. (2007) Ammonium assimilation and amino acid metabolism in conifers. *J. Exp. Bot.* **58**, 2307–2318.

Cantón, F.R., Suárez, M.F. and Cánovas, F.M. (2005) Molecular aspects of nitrogen mobilisation and recycling in trees. *Photosynth. Res.* **83**, 265–278.

Chancerel, E., Lepoittevin, C., Le Provost, G., Lin, Y.C., Jaramillo-Correa, J.P., Eckert, A.J., Wegrzyn, J.L., Zelenika, D., Boland, A., Frigerio, J.-M., Chaumeil, P., Garnier-Géré, P., Boury, C., Grivet, D., González-Martínez, S.C., Rouzé, P., Van de Peer, Y., Neale, D.B., Cervera, M.T., Kremer, A. and Plomion, C. (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. doi:10.1186/1471-2164-12-368.

Chancerel, E., Lamy, J.-B., Lesur, I., Noirot, C., Klopp, C., Ehrenmann, F., Boury, C., Le Provost, G., Label, P., Lalanne, C., Leger, V., Salin, F., Gion, J.-M. and Plomion, C. (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 Biol.* **11**, 50. doi:10.1186/1741-7007-11-50.

Chen, J., Uebbing, S., Gyllenstrand, N., Lagercrantz, U., Lascoux, M. and Källman, T. (2012) Sequencing of the needle transcriptome from Norway spruce (*Picea abies* Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genomics,* **13**, 589. doi:10.1186/1471-2164-13-589.

Côté, C.L., Boileau, F., Roy, V., Ouellet, M., Levasseur, C., Morency, M.-J., Cooke, J.E.K., Seguin, A. and MacKay, J. (2010) Gene family structure, expression and functional analysis of HD-Zip III genes in angiosperm and gymnosperm forest trees. *BMC Plant Biol.* **10**, 273. doi:10.1186/1471-2229-10-273.

Couturier, J., Montanini, B., Martin, F., Brun, A., Blaudez, D. and Chalot, M. (2007) The expanded family of ammonium transporters in the perennial poplar plant. *New Phytol.* **174**, 137–150.

Durban, J., Juárez, P., Angulo, Y., Lomonte, B., Flores-Diaz, M., Alape-Girón, A., Sasa, M., Sanz, L., Gutiérrez, J.M., Dopazo, J., Conesa, A. and Calvete, J.J. (2011) Profiling the venom gland transcriptomes of Costa Rican snakes by 454 pyrosequencing. *BMC Genomics,* **12**, 259. doi:10.1186/1471-2164-12-259.

Falgueras, J., Lara, A.J., Fernández-Pozo, N., Cantón, F.R., Pérez-Trabado, G. and Claros, M.G. (2010) SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics*, **11**, 38. doi:10.1186/1471-2105-11-38.

Farjon, A. (2010) *A Handbook of the World's Conifers*. Leiden/Boston: Brill.

Fernández-Pozo, N., Canales, J., Guerrero-Fernández, D., Villalobos, D.P., Díaz-Moreno, S.M., Bautista, R., Flores-Monterroso, A., Guevara, M.Á., Perdiguero, P., Collada, C., Cervera, M.T., Soto, A., Ordás, R., Cantón, F.R., Avila, C., Cánovas, M.G. and Claros, F.M. (2011) EuroPineDB: a high-coverage web database for maritime pine transcriptome. *BMC Genomics,* **12**, 366. doi:10.1186/1471-2164-12-366.

Fickett, J.W. (1982) Recognition of protein coding regions in DNA-Sequences. *Nucleic Acids Res.* **17**, 5303–5318.

Field, C.B., Behrenfeld, M.J., Randerson, J.T. and Falkowski, P. (1998) Primary production of the Biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.

Futamura, N., Totoki, Y., Toyoda, A., Igasaki, T., Nanjo, T., Seki, M., Sakaki, Y., Adriano, Mari.A., Shinozaki, K. and Shinohara, K. (2008) Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. *BMC Genomics*, **9**, 383. doi:10.1186/1471-2164-9-383.

Guillet-Claude, C., Isabel, N., Pelgas, B. and Bousquet, J. (2004) The Evolutionary implications of *knox-I* gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. *Mol. Biol. Evol.* **21**, 2232–2245.

Harfouche, A., Meilan, R., Kirst, M., Morgante, M., Boerjan, W., Sabatti, M. and Scarascia Mugnozza, G. (2012) Accelerating the domestication of forest trees in a changing world. *Trends Plant Sci.* **17**, 64–72.

Kinlaw, C.S. and Neale, D.B. (1997) Complex gene families in pine genomes. *Trends Plant Sci.* **2**, 356–359.

Kolpakov, R., Bana, G. and Kucherov, G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **31**, 3672–3678.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Rogozin, I.B., Smirnov, S., Sorokin, A.V., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7.

Koski, L.B., Gray, M.W., Lang, B.F. and Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics*, **6**, 151. doi:10.1186/1471-2105-6-151.

Leitch, A.R. and Leitch, I.J. (2012) Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629–646.

Lelu-Walter, M.-A., Bernier-Cardou, M. and Klimaszewska, K. (2006) Simplified and improved somatic embryogenesis for clonal propagation of *Pinus pinaster* (Ait.). *Plant Cell Rep.* **25**, 767–776.

Lepoittevin, C., Frigerio, J.M., Garnier-Géré, P., Salin, F., Cervera, M.-T., Vornam, B., Harvengt, L. and Plomion, C. (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**, 11034. doi:10.1371/journal.pone.0011034.

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

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics,* **25**, 1754–1760.

Li, X., Wu, H., Dillon, S. and Southerton, S. (2009) Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata* D. Don. *BMC Genomics,* **10**, 41. doi: 10.1186/1471-2164-10-41.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. and Wang, J. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L. and Quackenbush, J. (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* **28**, 3657–3665.

Lijavetzky, D., Carbonero, P. and Vicente-Carbajosa, J. (2003) Genome-wide comparative phylogenetic analysis of the rice and *Arabidopsis* Dof gene families. *BMC Evol. Biol.* **3**, 17. doi:10.1186/1471-2148-3-17.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J. and Deng, H.W. (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics,* **27**, 2031–2037.

Lorenz, W.W., Ayyampalayam, S., Bordeaux, J., Howe, G.T., Jermstad, K.D., Neale, D.B., Rogers, D.L. and Dean, J.D. (2012) Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. *Trees Genet. Genom.* **8**, 1477–1485.

Mackay, J., O'Malley, D.M., Presnell, T., Booker, F.-L., Campbell, M.M., Whethen, R.W. and Sederoff, R. (1997) Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. *Proc. Natl Acad. Sci. USA,* **94**, 8255–8260.

Mackay, J., Dean, J., Plomion, C., Peterson, D.G., Cánovas, F.M., Pavy, N., Ingvarsson, P.K., Savolainen, O., Guevara, M.Á., Fluch, S., Vinceti, B., Abarca, D., Díaz-Sala, C. and Cervera, M.T. (2012) Towards decoding conifer mega-genomes. *Plant Mol. Biol.* **80**, 555–569.

Marth, G.T., Korf, I., Yandeli, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hiller, L., Kwok, P.-Y. and Gish, W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456.

Martin, D.M., Fäldt, J. and Bohlmann, J. (2004) Functional characterization of nine norway spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d subfamily. *Plant Physiol.* **135**, 1908–1927.

Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., Kubisiak, T.L., Amerson, H.V., Carlson, J.E., Nelson, C.D. and Davis, J.M. (2009) Evolution of Genome Size and Complexity in *Pinus*. *PLoS ONE*, **4**, e4332.

Muller, T., Ensminger, I. and Schmid, K.J. (2012) A catalogue of putative unique transcripts from Douglas fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings. *BMC Genomics*, **13**, 673. doi:10.1186/1471-2164-13-673.

Mullin, T.J., Andersson, B., Bastien, J.-C., Beaulieu, J., Burdon, R.D., Dvorak, W.S., King, J.N., Kondo, T., Krakowski, J., Lee, S.D., McKeand, S.E., Pâques, L., Raffin, A., Russell, J., Skrøppa, T., Stoehr, M. and Yanchuk, A. (2011) Economic importance, breeding objectives and achievements. In *Genetics, Genomics and Breeding of Conifers Trees* (Plomion, C., Bousquet, J. and Kole, C., eds), pp. 40–127, New York, NY: Edenbridge Science Publishers and CRC Press.

Neale, D.B. and Kremer, A. (2011) Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* **12**, 111–122.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, A., Rilakovic, N., Ritland, C., Rosselló, J.A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Gil, R.G., Hvidsten, T.R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Thompson, S.L., Van de Peer, Y., Andersson, B., Nilsson, B., Ingvarsson, P., Lundeberg, J., and Jansson, S. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, doi:10.1038/nature12211.

Olson, J.S., Watts, J.A. and Allison, L.J. (1983) *Carbon in live vegetation of major world ecosystems*(Oak Ridge National Laboratory, Oak Ridge, TN), Report ORNL-. 5862.

Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W. and Buerkle, C.A. (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

Park, Y., Lelu-Walter, M., Harvengt, L., Trontin, J., MacEacheron, I., Klimaszewska, K. and Bonga, J. (2006) Initiation of somatic embryogenesis in *Pinus banksiana*, *P. strobus*, *P. pinaster*, and *P. sylvestris* at three laboratories in Canada and France. *Plant Cell, Tissue Organ Cult.* **86**, 87–101.

Pavy, N., Paule, C., Parsons, L., Crow, J., Morency, M.-J., Cooke, J., Johnson, J.E., Noumen, E., Guillet-Claude, C., Butterfield, Y., Barber, S., Yang, G., Liu, J., Stott, J., Kirkpatrick, R., Siddiqui, A., Holt, R., Marra, M., Seguin, A., Retzel, E., Bousquet, J. and MacKay, J. (2005) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics*, **6**, 144.

Pérez-Rodríguez, P., Riaño-Pachón, D.M., Guedes-Correa, L.G., Rensing, S.A., Kersten, B. and Mueller-Roeber, B. (2010) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* **38** (Database issue), D822–D827.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J. and Pandolfi, P.P. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumor biology. *Nature*, **465**, 1033–1040.

Ralph, S., Chun, H., Kolosova, N., Cooper, D., Oddy, C., Ritland, C.E., Kirkpatrick, R., Moore, R., Barber, S., Holt, R.A., Jones, S.J., Marra, M.A., Douglas, C.J., Ritland, K. and Bohlmann, J. (2008) A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). *BMC Genomics*, **9**, 484.

Resende, M.F.R., Muñoz, P., Resende, M.D.V., Garrick, D.J., Fernando, R.L., Davis, J.M., Jokela, E.J., Martin, T.A., Peter, G. and Kirst, M. (2012) Accuracy

of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*, **190**, 1503–1510.

Rigault, P., Boyle, B., Lepage, P., Cooke, J.E.K., Bousquet, J. and Mackay, J. (2011) A white spruce catalogue for conifer genome analyses. *Plant Physiol.* **157**, 14–28.

Ritland, K. (2012) Genomics of a phylum distant from flowering plants: conifers. *Tree Genet. Genomes,* **8**, 573–582.

Rueda-López, M., García-Gutiérrez, A., Cánovas, F.M. and Avila, C. (2013) The family of Dof transcription factors in pine. *Trees Struct. Funct*. doi:10.1007/s00468-013-0903-z.

Sanz, F., Latour, S., Neves, M., Bastet, E., Pischedda, D., Piñeiro, G., Gauthier, T., Lesbats, J., Plantier, C., Marques, A., Lanvin, J.-D., Santos, J.A., Touza, M., Pedras, F., Parrot, J., Reuliong, D. and Faria, C. (2006) *Industrial Applications of Pinus Pinaster*. 256 pp. Madeira: CIS Madeira, FIBA, AIMMP, CTBA.

Scaglione, D., Lanteri, S., Acquadro, A., Lai, Z., Knapp, S.J., Rieseberg, L. and Portis, E. (2012) Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotechnol. J.* **10**, 956–969.

Soltis, D.E. and Soltis, P.S. (2003) The role of phylogenetics in comparative genetics. *Plant Physiol.* **132**, 1790–1800.

Suzuki, S., Li, L., Sun, Y.-H. and Chiang, V.L. (2006) The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* **142**, 1233–1245.

Trontin, J.F., Debille, S., Canlet, F., Harvengt, L. and Lelu-Walter, M.-A., Label, P., Teyssier, C., Miguel, C., De Vega-Bartol, J., Tonelli, M., Santos, R., Rupps, A., Hassani, S.B., Zoglauer, K., Carneros, E., Díaz-Sala, C., Abarca, D., Arrillaga, I., Mendoza-Poudereux, I., Segura, J., Avila, C., Rueda, M., Canales, J. and Cánovas, F.M. (2013) Somatic embryogenesis as an effective regeneration support for reverse genetics in maritime pine: the Sustainpine collaborative project as a case study. In: *Proceeding of the IUFRO Working Party 2.09.02 conference on "Integrating vegetative propagation, biotechnology and genetic improvement for tree production and sustainable forest management", 25-28/06/2012 (Brno, Czech Republic)*, (Park, Y.S. and Bonga, J.M., eds), pp 184–187. Published online (http://www.iufro20902.org/).

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. and Rokhsar, D. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

UNECE (2013). *Forest product statistics 2007-2011*. Timber Bulletin ECE/TIM/BULL/65/2, Issue 2012, United Nations Economic Commission for Europe (UNECE), Trade and Sustainable Land management Division, http://www.unece.org/fileadmin/DAM/timber/statsdata/fps07-11.pdf.

Villalobos, D.P., Díaz-Moreno, S.M., El-Sayed, S.S., Cañas, R.A., Osuna, D., Van Kerckhoven, S.H., Bautista, R., Claros, M.G., Cánovas, F.M. and Cantón, F.R. (2012) Reprogramming of gene expression during compression wood formation in pine: coordinated modulation of S-adenosylmethionine, lignin and lignan related genes. *BMC Plant Biol.* **12**, 100. doi:10.1186/1471-2229-12-100.

Xu, Z., Zhang, D., Hu, J., Zhou, X., Ye, X., Reichel, K.L., Stewart, N.R., Syrenne, R.D., Yang, X., Gao, P., Shi, W., Doeppke, C., Sykes, R.W., Burris, J.N., Bozell, J.J., Cheng, M.Z., Hayes, D.G., Labbe, N., Davis, M., Stewart, C.N. Jr. and Yuan, J.S. (2009) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. *BMC Bioinformatics,* **10**, S3.

Supporting information

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

Figure S1 Distribution of GO terms among the annotated unigenes of *Pinus pinaster* transcriptome (Figure S1.pdf).

Figure S2 Hierarchical clustering analysis of TF families in plants (Figure S2.pdf).

Methods S1-Supplemental section on transcriptome assembly (Methods S1.pdf).

Table S1 A comparison of data set collection of transcription factor families present into the genomes of *Pinus pinaster, Picea glauca, Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa and Vitis vinifera*. (Table S1.docx).

Table S2 Maritime pine SNPs. (Table S2. Maritime pine SNP.txt).