biology and life sciences forum

MDPI

Proceeding Paper

Transcriptome Characterization of Different Tissues of Stone Pine (*Pinus pinea* L.): *de novo* Assembly †

Ana Usié [1,2,‡,*], Bruna Mendes [1,‡], Marta Antunes [1,3,§], Célia Leão [1,2,4,§], Liliana Marum [1,2,*] and António Marcos Ramos [1,2]

[1] Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo (CEBAL) Instituto Politécnico de Beja (IP-Beja), 7801-908 Beja, Portugal; bruna.mendes@cebal.pt (B.M.); fc48389@alunos.fc.ul.pt (M.A.); celia.leão@iniav.pt (C.L.); marcos.ramos@cebal.pt (A.M.R.)
[2] MED-Mediterranean Institute for Agriculture, Environment and Development, Évora, Portugal
[3] cE3c–Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[4] INIAV Instituto Nacional de Investigação Agrária e Veterinária, I.P. Laboratório Nacional de Referência de Saúde Animal. 2780-157 Oeiras, Portugal
* Correspondence: ana.usie@cebal.pt (A.U.); liliana.marum@cebal.pt (L.M.)
† Presented at the 2nd International Electronic Conference on Plant Sciences—10th Anniversary of Journal Plants, 1–15 December 2021; Available online: https://iecps2021.sciforum.net/.
‡ Authors contributed equally to this work.
§ Current affiliation.

Abstract: Stone pine (*Pinus Pinea* L.) is an emblematic tree distributed around the whole Mediterranean basin. The species is well known for the economics of its timber, resins and edible seeds, the stone pine nuts commercialized in food industry. Despite its relevance, the genomic information available for the species is scarce, and until now no reference genome is available. The main purpose of this study was to characterize the stone pine transcriptome of seven different tissues, by performing a *de novo* transcriptome assembly. A total of 55,328 genes were predicted and functionally annotated based on SWISS-PROT and nr-NCBI databases and InterProScan signatures.

Keywords: Stone pine; transcriptome; *de novo* assembly; RNA-Seq;

Citation: Usié, A.; Mendes, B.; Antunes, M.; Leão, C.; Marum, L.; Ramos, A.M. Transcriptome characterization of different tissues of stone pine (Pinus pinea L.): de novo assembly. Biol. Life Sci. Forum **2021**, 1, x. https://doi.org/10.3390/xxxxx

Academic Editor: Carmen Arena

Published: 30 November 2021

1. Introduction

Stone pine (*Pinus Pinea* L.) is a Mediterranean species distributed in coastal areas from the western of the Iberian Peninsula to Turkey [1]. Stone pine is a valuable species for its pine nuts or pine kernels, which are a delicious and highly nutritious edible seeds being a good source of fat, proteins and vitamins, among other phytochemical characteristics [2–4]. In addition, the species is also well known for the economics of its timber and resins. Between 2010 and 2015 the Portuguese stone pine area increased by 20,700 ha reaching 193,600 ha in 2015 [5], being the second largest area of stone pine of the world.

Advances in sequencing and assembly technologies have allowed a rapid progress in the characterization of the angiosperms genomes, while for gymnosperms organisms, such as conifers, the same does not happen due to the complexity and higher size of their genomes. For instance, the mean size of genomes at the *Pinus* genus and subgenus are 28.3 Gbs and 26.4 Gbs, respectively [6]. The recent advances in the third generation of high-throughput sequencing technologies and their cost reduction, allowed the sequencing of two pine genomes from the genus Pinus, *P. labertiana* (GCA_001447015.2) and *P. taeda* [7].

Despite the scarce genomic information available for the stone pine, the characterization of the transcriptome even for species with no reference genome available can be performed using RNA-Seq. Transcriptome differences between different plant tissues have been well studied so far, providing a comprehensive characterization of the species

transcriptome. Here, in order to explore the transcriptome differences between different tissues of stone pine, a transcriptome characterization of needles, xylem, stem bark, terminal bud, first and second year pine cone, and pine nut, was performed by a *de novo* transcriptome assembly. This study provides for the first time transcriptome resources of seven different tissues of the stone pine, being a valuable resource for further studies.

2. Materials and Methods

2.1. Sample Preparation, RNA Extraction and Sequencing

Samples of different tissues (needles, xylem, stem bark, terminal bud, first and second year pine cone and pine nut) were collected from five trees located in Coruche (Portugal). Samples were immediately frozen in liquid nitrogen and stored at −80 °C until being processed. The RNA extraction was performed according to Le Provost [8] with minor modifications. The extracted RNA was sequenced in two different Illumina platforms, NextSeq 550 and HiSeq 4000, producing paired-end (PE) reads of 75bp and 100bp in length, respectively.

2.2. Sequencing Data, Transcriptome Assembly and Annotation

The raw reads were pre-processed with Trimmomatic v.0.38 [9] keeping reads with a minimum quality of 20, over a screen size window of 10% of the read length, and minimum length of 80% of the read length. Then, the *de novo* transcriptome assembly was performed using Mira v.4.0.2 [10] discarding contigs shorter than 200bp.

Gene prediction and transcriptome annotation was performed using TransDecoder v.5.5 [11], following its guidelines. BlastP was used to functionally annotate the predicted genes, identifying homologous genes from SWISS-PROT and nr-NCBI plants databases, and InterProScan was used to obtain protein domains, gene ontology (GO) terms and KEGG pathways [12–15].

2.3. Tissue-Specific Characterization

In order to characterize the transcriptome at the tissue level the pre-processed reads of all individuals were mapped against the assembled transcriptome with STAR v.2.7.3a [16], using the two-pass mode according to the manual guidelines. The unique mapped reads were retained and used to estimate the RNA abundance of the predicted genes by StringTie (parameter used -e) [17]. The tissue-specific characterization was performed taking into account only genes with an abundance ≥5 in at least one of the biological replicates in at least one tissue, considering those genes as genes expressed. Then, BinGO plugin from Cytoscape was used to identify GOs overrepresented over the set of genes expressed in each tissue, performing a (BH) multiple testing correction with a p-value ≤ 0.05 [18].

3. Results

3.1. Transcriptome Assembly and Annotation

A *de* novo transcriptome assembly was generated from the sampled tissues. Sequencing of cDNA of all samples from both sequencing platforms generated a total of 2,026,716,380 PE reads. After trimming low-quality bases and removing low quality reads with Trimmomatic, 1,898,376,282 high-quality reads were kept, representing the 93.7% of the raw reads (Table 1). The transcriptome assembly of stone pine resulted in 165,179 contigs equal or greater than 200 bp, which represented an accumulative assembly size of 81,310 Mb (Table 2). A total of 55,328 candidate genes were identified by Transdecoder from which 41,839 found at least one homologous hit against the SWISS-PROT database. The remaining predicted genes with no hits were further blasted against the nr-NCBI plants where 8322 genes found at least one homology hit.

Functional categories in terms of GOs and associated KEGG pathways were identified by InterProScan. A total of 28,258 (51.07%) predicted genes were assigned with at least one GO term, covering 2079 different GO terms (BP–biological processes: 41.75%;

MF–molecular function: 45.46%; CC–cellular components: 12.94%). Moreover, 4134 predicted genes were successfully assigned to at least one KEGG pathway of the 124 identified, codifying 482 different enzymes.

Table 1. Number of reads from the RNA-Seq data of different tissues of *Pinus Pinea*, before and after quality control (QC).

Tissue	Nº Samples	Nº Raw Reads	Nº Reads after QC	% Reads after QC
BGI- Illumina Platform HiSeq 4000				
Needle	5	146,326,868	138,159,558	94.4
Xylem	5	138,531,098	133,130,580	96.1
Stem bark	5	143,731,678	135,937,996	94.6
Terminal bud	5	135,949,880	130,136,482	95.7
1st year pine cone	5	135,310,934	125,938,522	93.1
2nd year pine cone	5	69,963,354	65,575,696	93.7
Pine nut	5	144,025,338	135,919,536	94.4
Total	**35**	**913,839,150**	**864,798,370**	**94.6**
BIOCANT-Illumina Platform NextSeq 550				
Needle	2	131,986,590	121,104,596	91.8
Xylem	3	157,461,310	149,028,612	94.6
Stem bark	4	160,306,566	140,699,320	87.8
Terminal bud	2	150,995,220	141,263,030	93.6
1st year pine cone	3	203,695,358	190,859,772	93.7
2nd year pine cone	3	154,578,072	145,230,246	94,0
Pine nut	3	153,854,114	145,392,336	94.5
Total	**20**	**1,112,877,230**	**1,033,577,912**	**92.9**

Table 2. General assembly metrics for the stone pine transcriptome.

Metric	Value
Total number of contigs	165,179
Nº of contigs ≥200 bp	165,179
Nº of contigs ≥500 bp	45,648
Nº of contigs ≥1000 bp	13,912
Nº of contigs ≥2000 bp	4043
Nº of contigs ≥4000 bp	467
Nº of contigs ≥6000 bp	58
Nº of contigs ≥8000 bp	13
Total length of contigs	813,10,033 bp
Largest contig	11,938 bp
GC %	45.32
N50	567

3.2. Tissue-Specific Characterization

After removing genes with low abundances the universe of expressed genes considered for the transcriptome characterization was 54,627, from which 30,137 genes were co-expressed in all tissues. In addition, 5738 genes were found exclusively expressed in pine nut, where in the other tissues the number of exclusively expressed genes was much lower (needles: 1087; stem bark: 212; xylem: 210; terminal bud: 143; first year pine cone: 74; second year pine cone: 21). By performing the pairwise comparisons of genes expressed per tissue, the pine nut tissue is the one with less genes expressed in common with the other tissues. For instance, 5862 genes were co-expressed in all tissues, but pine nut. On the other hand, the two pine cone tissues, first year and second year, are the ones with more similarity (higher Jaccard index).

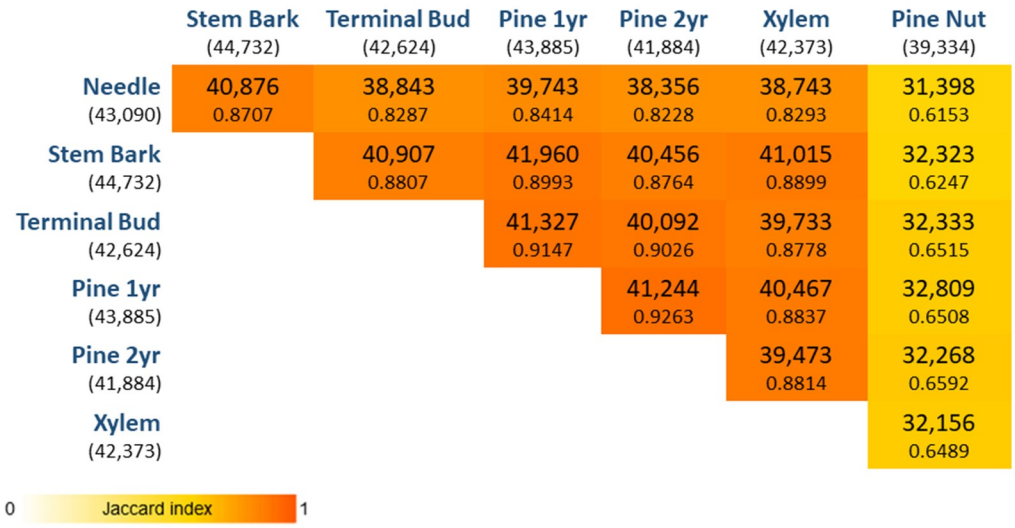

Figure 1. Pairwise comparison per tissue. In brackets are represented the total number of genes expressed per tissue. Within each square is represented the number of genes in common between tissues and below that number, the corresponding Jaccard index. The higher the index value, the more similar the two tissues compared.

In order to understand the enrichment occurrence of overrepresented GO terms in each tissue, the proportion of genes expressed in each tissue was compared with the expressed genes overall the transcriptome assembly. The analysis showed that 1170 GOs were found overrepresented among all tissues (BP: 683; MF: 284; CC: 203). When looking for exclusive overrepresented GO terms per tissue a total of 409 GO terms overrepresented were found (Table 3).

Table 3. Exclusive overrepresented GO terms per tissue classified by categories. BP: Biological processes; MF: Molecular functions; CC: Cellular components.

Tissue	BP	MF	CC	Total
Needles	73	31	10	114
Stem bark	2	2	0	4
Terminal bud	7	3	1	11
First year pine cone	0	3	0	3
Second year pine cone	2	3	0	5
Pine nut	181	14	36	231
Xylem	30	6	5	41
TOTAL	**295**	**62**	**52**	**409**

In terms of exclusively overrepresented GOs identified in each tissue was observed that in pine nut tissue, most of them were related with seed maturation, initiation of transcription, translation and stored nutrient mobilization, cell expansion, root development and cell division among others while in needles were related with photosynthesis and energy metabolism. Additionally, metabolic processes associated with coenzymes and cofactors were related with exclusively overrepresented GOs in stem bark tissue and in terminal bud with translation of elongation factors, cell structure and cell wall organization. Finally, catalytic activities were related with exclusively overrepresented GOs in second year pine cone tissue.

Regarding KEGG pathways, 482 different enzymes were codified by 4057 genes expressed among the whole transcriptome (needles: 3769; stem bark: 3474; terminal bud: 3067; first year pine: 3114; pine nut: 2909; second year pine: 2900; xylem: 2400). The most

representative KEGG pathways per tissue were represented in Figure 2. Clear differences were observed between needles and stem bark tissues in comparison with the other tissues. Both contained a higher number of genes expressed associated directly with energy metabolism such as glycolysis; gluconeogenesis and diverse sugar metabolism (galactose, fructose and mannose). The highest difference was observed in both of these tissues against the others in "Glyoxylate and dicarboxylate" and "Carbon fixation in photosynthetic organisms" metabolisms, which usually are more active in photosynthetic tissues.

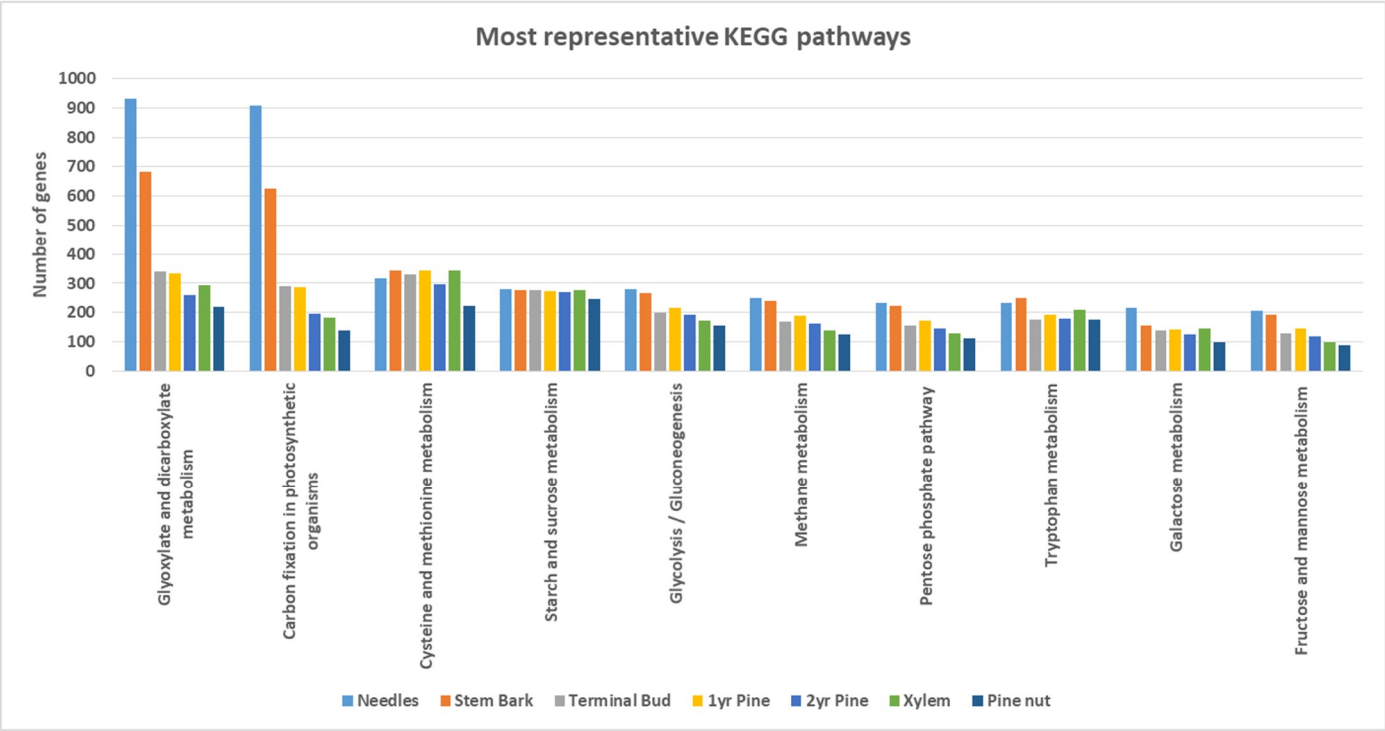

Figure 2. The most representative KEGG pathways associated with all predicted genes.

4. Conclusions

This is the first time that a large scale RNA-seq dataset was generated from seven different tissues of stone pine providing a complete transcriptome characterization. The data produced will be a useful resource for future studies in the species. The transcriptome assembly generated resulted in a total of 55,328 genes identified from which 50,161 were functionally annotated. More studies are on-going in order to assess differences in gene expression between tissues in stone pine.

Author Contributions: This study was conceived by A.M.R. Collection and identification of filed material was performed by A.U., B.M., M.A., C.L. and A.M.R. R.N.A. extraction was performed by C.L. Bioinformatics data analyses were conducted by A.U., B.M. and M.A. Biological interpretation of the results was conducted by A.U., B.M. and L.M. The manuscript was written by A.U. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-financed by Program Alentejo 2020, through the European Fund for Regional Development under the scope SelectPinea- Development of genetic markers for relevant traits in stone pine (ALT20-03-0145-FEDER-000041). Contrato–Programa to L. Marum (CEEC-INST/00131/2018) and UIDB/05183/2020 were funded by FCT.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Quézel, P.; Médail, F. Ecologie et biogéographie des forêts du bassin méditerranéen. 2017.

2. Nergiz, C.; Dönmez, I. Chemical composition and nutritive value of *Pinus pinea* L. seeds. *Food Chem.* **2004**, *86*, 365–368, doi:10.1016/J.FOODCHEM.2003.09.009.
3. Bolling, B.W.; Chen, C.Y.; McKay, D.L.; Blumberg, J.B. Tree nut phytochemicals: Composition, antioxidant capacity, bioactivity, impact factors. A systematic review of almonds, Brazils, cashews, hazelnuts, macadamias, pecans, pine nuts, pistachios and walnuts. *Nutr. Res. Rev.* **2011**, *24*, 244–275, doi:10.1017/S095442241100014X.
4. Evaristo, I.; Batista, D.; Correia, I.; Correia, P.; Costa, R. Chemical profiling of Portuguese Pinus pinea L. nuts. *J. Sci. Food Agric.* **2010**, *90*, 1041–1049, doi:10.1002/JSFA.3914.
5. ICNF, 2019. Available online: http://www2.icnf.pt/portal/florestas/ifn/resource/doc/ifn/ifn6/IFN6_Relatorio_completo-2019-11-28.pdf (accessed on).
6. Grotkopp, E.; Marcel, R.; Sanderson, M.J.; Rost, T.L. Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. *Evolution* **2004**, *58*, 1705–1729, doi:10.1111/J.0014-3820.2004.TB00456.X.
7. Zimin, A.V. et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience* **2017**, *6*, 1–4, doi:10.1093/GIGASCIENCE/GIW016.
8. Le Provost, G.; Herrera, R.; Paiva, J.A.P.; Chaumeil, P.; Salin, F.; Plomion, C. A micromethod for high throughput RNA extraction in forest trees. *Biol. Res.* **2007**, *40*, 291–297, doi:10.4067/S0716-97602007000400003.
9. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120, doi:10.1093/BIOINFORMATICS/BTU170.
10. Chevreux, B.; Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *German Conf. Bioinform.* **1999**, *99*, 45–56.
11. Haas, B.; Papanicolaou, A.J.G.S. TransDecoder (find coding regions within transcripts). *Google Sch*. **2016**.
12. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421, doi:10.1186/1471-2105-10-421.
13. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter,M.C.; Estreicher, A.; Gasteiger, E.; Martin, M.J.; Michoud, K.; O'Donovan, C.; Phan, I.; et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365–370, doi:10.1093/NAR/GKG095.
14. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46*, D8–D13, doi:10.1093/NAR/GKX1095.
15. Jones, P.; Binns, D.; Chang, H.Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240, doi:10.1093/BIOINFORMATICS/BTU031.
16. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21, doi:10.1093/BIOINFORMATICS/BTS635.
17. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295, doi:10.1038/nbt.3122.
18. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **2005**, *21*, 3448–3449, doi:10.1093/BIOINFORMATICS/BTI551.