simpósio estadual de agroenergia
V reunião técnica de agroenergia - RS

DE NOVO ASSEMBLY OF TUNG TREE SEEDS TRANSCRIPTOME

Vanessa Galli[1,2], Rafael da Silva Messias [1], Sérgio Delmar dos Anjos e Silva [1], Rogério Margis[2]

INTRODUCTION

Tung oil, the major product of tung tree (*Vernicia fordii*) seeds, is considered one of the highest quality oils. The tung seeds accumulate high levels of α-eleostearic acid (approximately 80 %), a trienoic fatty acid with conjugated double bonds (9cis, 11trans, 13trans octadecatrienoic acid), that is widely used in paints, high quality printing, plasticizers, medicine, and in chemical reagents. Moreover, because tung seeds accumulate high content of oil (approximately 50 %), it has being recently considered for use in biodiesel production. However, large-scale production of tung oil through traditional farming is hampered because of the poor agronomic traits of this plant species (PARK et al., 2008; SHANG et al., 2010). Therefore, increasing the yield and adjusting characters of tung oil are the major challenges for industry.

Over the past several years, we have greatly improved our understanding of a plethora of biological processes. Transcriptome sequencing is an efficient methodology for large scale gene discovery, molecular markers development, genomic and transcriptomic assembly, and microarray development (VENTURINI et al., 2013). To provide a comprehensive and accurate foundation for molecular studies of tung tree, herein we present the reference transcriptome dataset of *V. fordii* mature seeds, assembled and annotated from deep RNA-Seq data. This study represents the first large-scale transcriptome annotation of tung tree.

MATERIAL AND METHODS

For the construction of the RNA-Seq, fruits from *V. fordii* plants grown in an open environment at Embrapa Clima Temperado (Pelotas, Brazil) were collected at mature stage (approximately 120 days after flower opening - DAF). Total RNA was isolated using Trizol reagent (Invitrogen), according to the manufactures' protocol. The RNA quality was evaluated by electrophoresis on a 1 % agarose gel and the RNA concentration was determined by absorbance at 260 nm, in a Nanodrop spectrophotometer (Nanodrop Technologies).

[1]Centro de Biotecnologia, PPGBCM, Laboratório de Genomas e População de Plantas, prédio 43431, Universidade Federal do Rio Grande do Sul - UFRGS, P.O. Box 15005, CEP 91501-970, Porto Alegre, RS, Brazil.
[2]Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA, P.O. Box 403, CEP 96010-971, Pelotas, RS, Brazil.

Total RNA (> 10 µg) isolated from mature seeds (120 DAF) was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing and sequencing by Illumina HiSeq2000. After removing low quality reads, RNA-Seq data was *de novo* assembled into contigs using Velvet/Oases package (SCHULZ et al., 2012). We used a minimum contig length of 200, and a multi (i.e. 21, 31, 41, 51 and 61 bp) *k*-mers (substrings of length *k*) based strategy to capture the most diverse assembly with improved specificity and sensitivity, especially for low expressed genes. We then used the USEARCH algorithm (EDGAR, 2010) to obtain the unigenes. The annotation was performed by applying BLASTX program (e-value<$1e^{-6}$) against the non-redundant sequence from NCBI (http://www.ncbi.nlm.nih.gov/), and the Swiss-prot database (http://www.expasy.ch/sprot/).

RESULTS AND DISCUSSION

We have sequenced 43,081,927 ESTs from mature tung seeds by using Illumina sequencing. The ESTs were divided into 21, 31, 41, 51 and 61 *k*-mers in order to improve the specificity and sensitivity of the assembly using Velvet. Therefore, 97,647, 82,023, 69,855, 62,041, and 51,157 transcripts were obtained using 21, 31, 41, 51 or 61 *k*-mers for assembly, respectively. The mean size of those transcripts was 1,108, 1,223, 1,276, 1,191 and 1,266 bp in the 21, 31, 41, 51 and 61 *k*-mers strategy, respectively. The statistics of *V. fordii* transcripts obtained with different *k*-mers using Velvet is shown in Table 1. The use of multi *k*-mers for *de novo* assembly have been successfully performed by several authors (GRUENHEIT et al., 2012; YANG et al., 2012).

Table 1. Statistics of *V. fordii* transcripts obtained with different k-mer using Velvet, ano. Local/RS.

Description	K-mer 21	K-mer 31	K-mer 41	K-mer 51	K-mer 61	Total
Number of transcripts	97,647	82,023	69,855	62,041	51,157	362,723
Median transcript length	577	785	887	823	868	3,940
Mean transcript length	1,108	1,223	1,276	1,191	1,266	6,064
Max transcript length	17,304	19,718	16,856	16,123	16,607	86,608
No. transcript > 1kbp	37,180	35,871	32,372	27,023	23,442	155,888
N50	2,170	2,187	2,146	1,931	2,092	10,526

All contigs were further merged by integrating sequence overlap to determine the number of unique sequences. Therefore, 47,585 unisequences (unigenes) were obtained, creating an initial reference transcriptome. The lengths of the unigenes ranged from 200 to 19,718 nucleotides (nt), with an average size of 1,684 nt. From the unigenes obtained, 64% were more than 1,000 nt,

confirming the quality of the transcriptome assembled. The distribution of the unigenes according to the length is presented in Figure 1, where most of the unigenes have between 1,000 and 2,500 nt.

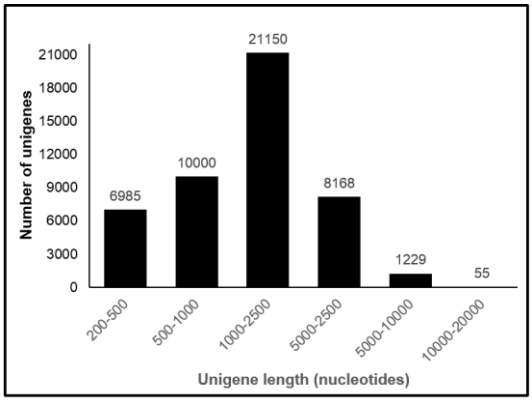

Figure 1. Size distribution of the unigenes contigs from tung transcriptome, ano. Local/RS..

All unigenes were aligned against the non-redundant (NR) protein database of GenBank using BLASTX with an e-value cut-off of $1e^{-6}$. We found matches for 45,824 unigenes (96 %). Most of the best hits were from *Vitis vinifera* (17,784 sequences, 38.8 %), probably because there are a large number of deposited ESTs of this specie in the NR database. The second most frequent specie was *Glycine max* (5,546 sequences, 12.1 %), followed by *Populus trichocarpa* (4,665 sequences, 10.2 %) and *Arabidopsis thaliana* (4,398 sequences, 9.57 %). Only 213 sequences (0.46 %) matched sequences from *V. fordii* deposited in the NR database, confirming the low amount of publicly available sequences from this plant (Figure 2A).

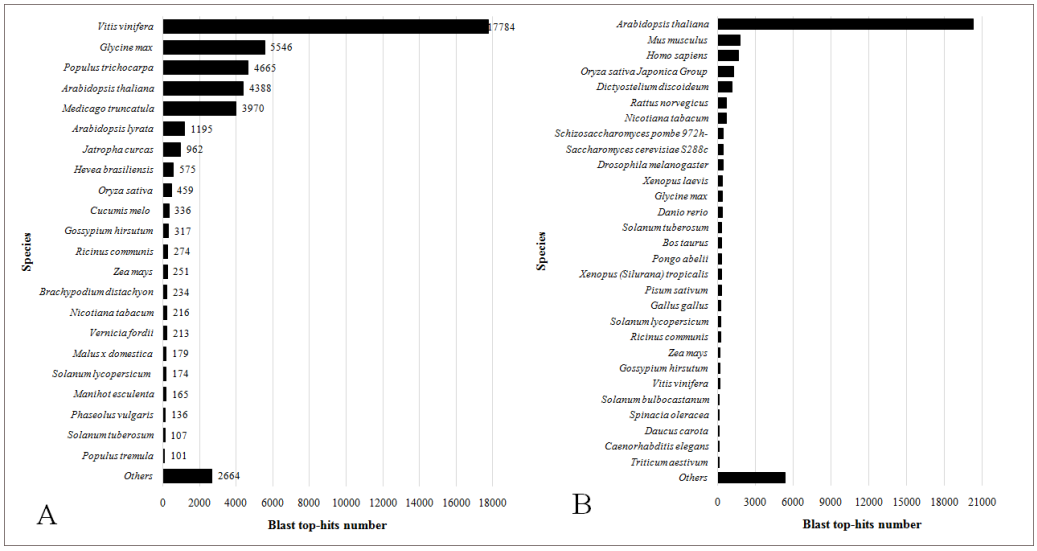

Figure 2. Blast top-hits resulted from the annotation using the GenBank non-redundant protein

database (A) and Swiss-Prot database (B), ano. Local/RS.

We also aligned the unigenes against the Swiss-Prot (SW) database, which resulted in the annotation of 38,785 unigenes (81 %). *A. thaliana* was the most frequent species in this analysis (20,278 sequences, 52.3 %), probably because SW database is enriched with sequences from this specie, as a plant-model (Figure 2B).

CONCLUSIONS

In the present work we provided a collection of ESTs and unigenes derived from mature seeds of tung. We performed the assembly and annotation of the transcriptome. This information generated is of paramount importance to be used in breeding programs and to engineer the entire oil synthesis pathway of tung seeds.

REFERENCES

EDGAR, RC.Search and clustering orders of magnitude faster than BLAST, **Bioinformatics**, v. 26, p. 2460 – 2461, 2010.

GRUENHEIT N, DEUSCH O, ESSER C, BECKER M, VOELCKEL C, LOCKHART P. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants **BMC Genom**, v. 13, p. 92, 2012.

PARK JY, KIM DK, WANG ZM, LU P, PARK SC, LEE JS. Production and characterization of biodiesel from tung oil. **Appl Biochem Biotechnol**, v. 148, p. 109 – 117, 2008.

SCHULZ MH, ZERBINO DR, VINGRON M, BIRNEY E. Oases: robust de novo RNAseq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28, p. 1086 – 1092, 2012.

SHANG Q, JIANG W, LU H, LIANG B. Properties of tung oil biodiesel and its blends with diesel. **Bioresour Technol**, v. 101, p. 826 – 828, 2010.

VENTURINI L, FERRARINI A, ZENONI S, et al. De novo transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. **BMC Genom**, v. 14, p. 41, 2013.

YANG W, QI Y, BI K, FU J. Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: A comparative transcriptomic analysis of two ranid frogs, *Rana chensinensis* and *R. kukunoris*. **BMC Genom**, v. 13, p. 588, 2012.