J.Natn.Sci.Foundation Sri Lanka 2015 43 (4): 319-336
DOI: http://dx.doi.org/10.4038/jnsfsr.v43i4.7967

RESEARCH ARTICLE

Transcriptome analysis for discovering candidate genes involve in embryogenesis in coconut (*Cocos nucifera* L.) through 454 pyrosequencing

H.D. Dharshani Bandupriya[1,2*] and Jim M. Dunwell[3]
[1]*School of Biological Sciences, University of Reading, Reading, RG6 6AS, UK.*
[2]*Tissue Culture Division, Coconut Research Institute, Bandirippuwa Estate, Lunuwila.*
[3]*School of Agriculture, Policy and Development, University of Reading, RG6 6AR, UK.*

Revised: 12 March 2015; Accepted: 22 April 2015

Abstract: Coconut, *Cocos nucifera* L. is a major plantation crop, which ensures income for millions of people in the tropical region. Detailed molecular studies on zygotic embryo development would provide valuable clues for the identification of molecular markers to improve somatic embryogenesis. Since there is no ongoing genome project for this species, coconut expressed sequence tags (EST) would be an interesting technique to identify important coconut embryo specific genes as well as other functional genes in different biochemical pathways. The goal of this study was to analyse the ESTs by examining the transcriptome data of the different embryo tissue types together with one somatic tissue. Here, four cDNA libraries from immature embryo, mature embryo, microspore derived embryo and mature leaves were constructed. cDNA was sequenced by the Roche-454 GS-FLX system and assembled into 32621 putative unigenes and 155017 singletons. Of these unigenes, 18651 had significant sequence similarities to non-redundant protein database, from which 16153 were assigned to one or more gene ontology categories. Homologue genes, which are responsible for embryo development such as chitinase, beta-1,3-glucanase, ATP synthase CF0 subunit, thaumatin-like protein and metallothionein-like protein were identified among the embryo EST collection. Of the unigenes, 6694 were mapped into 139 KEGG pathways including carbohydrate metabolism, energy metabolism, lipid metabolism, amino acid metabolism and nucleotide metabolism. This collection of 454-derived EST data generated from different tissue types provides a significant resource for genome wide studies and gene discovery of coconut, a non-model species.

Keywords: Coconut, embryogenesis, expressed sequence tags, immature embryo, mature embryo, microspore derived embryo.

INTRODUCTION

Coconut (*Cocos nucifera* L.) is a major plantation crop in the tropics, which ensures income for millions of people. Being an out-breeding crop, coconut is highly heterozygous and this is a barrier for improving the desired characters (i.e. higher yield, disease resistance, better quality) by conventional breeding, leaving *in vitro* vegetative propagation as the only tool for crop improvement on a large scale. Although *in vitro* propagation of coconut has been researched for more than three decades, a reproducible *in vitro* clonal propagation method for large scale plant production is yet to be developed. It is clear from the limited success in cloning coconut, that there is a need for better understanding of the process of somatic embryogenesis. Detailed molecular studies on zygotic embryo development would provide valuable clues to improve somatic embryogenesis and is thus regarded as an important model system. However processes related to embryogenesis in coconut are still poorly understood at the molecular level. Despite the fact that coconut serves as an important plantation crop, which is grown in more than 90 countries worldwide, currently there is no ongoing genome study for this species. The coconut genome is approximately 2.1 billion base pairs in size (Sniady *et al*., 2003), significantly larger than currently sequenced plants such as *Arabidopsis*, rice and *Medicago*. The lack of sequence information for coconut has seriously limited the progress of gene identification

*Corresponding author (dbandupriya@yahoo.com)

and characterisation, global transcript profile analysis, generation of molecular markers and probe designing for gene array experiments. This lack of knowledge can be overcome to a certain extent by using modern sequencing technologies.

With the improvement of DNA-sequencing technology, high speed, high-throughput methods popularly known as next generation sequencing technologies have evolved for sequencing (Mochida & Shinozaki, 2010). Next generation sequencing technology is one of the methodologies to generate expressed sequence tags (ESTs). The EST analysis is a rapid and cost effective way to identify expressed genes. Since non-coding and repetitive DNA, which is normally a major portion of a genome are avoided and only expressed genes are sequenced EST analysis is a rapid and cost effective way to generate important genetic information. High-throughput sequencing technologies have been used widely in gene expression studies in various tissue types, developmental stages or under different environmental conditions (Ho *et al.*, 2007; Kyndt *et al.*, 2012; Firon *et al*., 2013). Transcriptomic approaches have been successfully used to catalogue genes that are expressed during embryogenesis in economically important species such as potato (Sharma *et al*., 2008) and rice (Xu *et al*., 2012).

A next generation high-throughput sequencing method based on the Roche 454 Genome Sequencer (GS) FLX platform has emerged (Margulies *et al*., 2005) as a cost effective and most widely used *de novo* EST sequencing. It has been used so far for the successful construction of EST libraries from different plant species including the model plant *Arabidopsis thaliana* (Weber *et al.*, 2007) and other species such as *Palomero Toluqueno* (Vega-Arreguin *et al.,* 2009), *Olea europaea* (Alagna *et al*., 2009), *Castanea dentata* and *Castanea mollissima* (Barakat *et al.*, 2009) and *Eucalyptus grandis* (Novaes *et al*., 2008), as well as fish species such as *Zoarces viviparus* (Kristiansson *et al.*, 2009), coral species *A. millepora* (Meyer *et al.*, 2009), worms (Shin *et al*., 2008) and insects (Hahn *et al*., 2009). The 454 sequencing technology can identify a large number of expressed sequences, which enables the sequencing of large genome species that are inaccessible with traditional sequencing methods. The sequenced cDNA shows direct information on the mature transcripts for coding part of the genome. Thus, EST databases are very useful tools for the discovery of novel genes, investigation of genes of unknown function, comparative genomic studies, gene mapping, and functional studies.

A global analysis of the transcriptomes was performed using the 454 next generation sequencing technology from two zygotic embryo developmental stages and one *in vitro* derived embryo originated from microspores. Transcriptomes of developing leaves were also analysed in order to compare with the embryogenic stages. This comprehensive analysis of transcriptomes from different developmental stages of the embryo can provide genomic resources for the discovery of novel genes associated with embryo development and improve the global view of the potential molecular mechanisms underlying embryonic development.

METHODS AND MATERIALS

Plant growth conditions and tissue collection

Three different embryo tissue types from two zygotic embryo stages and one somatic embryo developed through *in vitro* culture, and one somatic tissue (leaf) were used for the cDNA library construction. The zygotic embryo stages, immature embryos at the age of nine months after pollination (9ME) and mature embryos at the age of 12 months after pollination (12ME) were sampled in 2010 from 25 year old Sri Lanka Tall coconut palms from Bandirippuwa Estate, Lunuwila, Sri Lanka. Microspore derived embryos (MDE) were initiated according to the protocol described by Perera *et al*. (2008). Developing leaves (LEAF) were sampled from an 8 month old *in vitro* germinated embryo culture plant obtained as described by Weerakoon *et al.* (2002). Leaf tissues were taken as a biological control for the embryogenic tissues.

RNA extraction

RNA was extracted from each tissue using the RNeasy plant mini kit (Qiagen, UK) according to the manufacturer's protocol. RNA was treated with DNaseI (RNase free DNase set; Qiagen, UK) according to the manufacturer's instructions to remove DNA contaminations. The RNA concentration was estimated using a ND-1000 spectrophotometer (ThermoScientific, NanoDropTM ND-1000) and the quality was checked in 1.0 % agarose gel.

cDNA synthesis

Total RNA was used for the preparation of double stranded cDNA using SMART approach (Zhu *et al.*, 2001). The oligonucleotides used in the experiment were as follows.

SMART Sfi1A : 5'- AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT ACG GCCrGrGrG-3'

CDS-Sfi1B : 5'- AAG CAG TGG TAT CAA CGC AGA GTG GCC GAG GCG GCCd(T)20–3'
SMART PCR primer : 5'-AAG CAG TGG TAT CAA CGC AGA GT- 3'

The primer annealing mixture (total volume 5 µL) containing 0.3 µg of total RNA; 10 pmol SMART-Sfi1A oligonucleotide and 10 pmol CDS-Sfi1B primer was heated at 72 °C for 2 min and cooled on ice for 2 min. First-strand cDNA synthesis was carried out in a total volume of 10 µL by mixing the annealed primer-RNA mixture from the first step with PowerScript Reverse Transcriptase (BD Biosciences Clontech) containing 1X first-strand buffer [50 mM Tris-HCl (pH 8.3); 75 mM KCl; 6 mM $MgCl_2$], 2 mM DTT and 1 mM of each dNTP, incubated at 42 °C for 2 h in an air incubator and then cooled on ice. First-strand cDNA was diluted 5 times with TE buffer, heated at 72 °C for 7 min and used for long distance PCR (Barnes, 1994) in a 50 µL reaction containing 1 µL diluted first-strand cDNA, 1X Advantage 2 reaction buffer (BD Biosciences Clontech), 200 µM dNTPs, 0.3 µM SMART PCR primer and 1X Advantage 2 Polymerise mix (BD Biosciences Clontech). 25 PCR cycles were performed using the following parameters: 95 °C for 7 s; 65 °C for 20 s; 72 °C for 3 min. The amplified cDNA PCR product was purified using QIAquick PCR Purification Kit (Qiagen, CA), concentrated by ethanol precipitation and adjusted to a final concentration of 50 ngμL^{-1}. For each embryo stage, cDNA was prepared from 3 separate embryos and pooled together. Leaf tissues from 2 separate plants were used to prepare cDNA, which were pooled together. A total yield of 3 µg of cDNA was prepared for each tissue. 454 sequencing was conducted by the Centre for Genomic Research, University of Liverpool, UK using the standard 454 amplicon sequencing protocol for pyrosequencing using a 454-GS FLX genome sequencer (454 Life Sciences, Roche).

De novo sequence assembly

DNA sequencing of 4 libraries was performed using a 454-GS FLX genome sequencer (454 Life Sciences, Roche) and the sequence data processing was performed with the GS FLX software v2.0.01 (454 Life Sciences, Roche) by using a series of normalisation and quality filtering techniques for the screening of weak and low quality sequences (Chen *et al.,* 2011). Assembly of these high quality sequences into consensus unigenes was achieved at default parameters by the 454 Newbler assembler software package provided with the 454 GS FLX System.

Homology searches and functional annotation

The unigene sequences in each library were blasted manually with the non-redundant protein database at the NCBI using the default settings of BLASTX programme. A powerful and free data mining software BLAST2GO (*http://www.BLAST2go.de/*) (Conesa *et al.*, 2005) was used to annotate the unique sequences of each library. During BLAST2GO annotation, similarity searches were performed against the nr database (non-redundant protein sequence database with entries from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq). The BLAST software used was BLAST 2.2.24. The sequences were searched using BLASTX with an E-value cut-off of $1E^{-6}$. The top hit BLAST results with an E-value equal to or less than 10^{-6} were considered as significant matches and the functional categories of these transcripts were further identified and annotated with gene ontology (GO) terms.

Pathway assignment with KEGG

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) mapping was used to assign biochemical pathways (Ogata *et al.*, 1999). Enzyme commission (EC) numbers were assigned to the unique transcripts, which had BLASTX scores with a cut-off value of $E \leq 10^{-6}$ as determined by the protein database search. The sequences were mapped to KEGG pathways according to the EC distribution in the pathway databases.

RESULTS

Expressed sequence tags (EST) generation

Four cDNA libraries constructed by SMART technology were sequenced using 454-GS FLX platform, which produced a total of 223.7 Mb from 979428 high quality sequences. A summary of these four EST datasets is given in Table 1. Initial quality filtering with the default settings yielded 63.8 Mb from 245864 high quality sequence reads for leaf, 48.4 Mb from 218376 for 9ME, 41.8 Mb from 207624 HQ sequences for 12ME and 69.7 Mb from 307564 for MDE. After removing the adapter sequences and too short sequences (less than 50 bp) the available sequences for assembly in leaf were 204677, which accounted for 83.24 % of the HQ sequences. In the 9ME library 160737 (73.61 %) reads were available after quality filtering for the assembly of unigenes. Similarly, removal of too short sequences in the other two libraries 12ME and MDE produced 155872 (75.07 %) and 221238 (71.93 %) sequences, respectively

Table 1: Summary of coconut ESTs from four libraries

	Leaf	9ME	12ME	MDE	Total
Number of HQ reads	245864	218376	207624	307564	979428
Total bases of HQ reads	63818922	48384097	41804406	69714046	223721471
Average length of sequence read	260	222	201	227	228
Reads used in assembly	204677	160737	155872	221238	742524
Reads assembled as unigenes	160590	129422	119535	178456	588003
Number of unigenes	11,653	6,297	5,332	9,339	32621
Average unigene length ± SD	522 ± 362	433 ± 220	418 ± 230	469 ± 265	
Number of singletons	44583	31315	36337	42782	155017
Range unigene length	95 - 4517	76 - 2337	93 - 2148	96 - 3418	
Number of unigenes less than 500bp	7627	4876	4134	6817	
Number of unigenes higher than 1000bp	1158	199	165	449	
Unigenes with BLASTX matches (E value cut off 10^{-6})	7530	3203	2870	5248	
Annotated sequences	6466	2749	2448	4490	

for the unigene assembly. In each of the libraries, a large number of singletons were recorded. Out of the sequences, 44583 in leaf, 31315 in 9ME, 36337 in 12ME and 42782 in MDE were singletons. The assembly of quality filtered sequence reads using Newbler software provided with the Roche GS FLX sequencer led to the construction of 11653, 6297, 5332 and 9339 unigenes from leaf, 9ME, 12ME and MDE, respectively with average lengths 522 ± 362, 433 ± 220, 417 ± 230 and 469 ± 265 bp. The variation of length distribution of the unigene is shown in Figure 1. It demonstrates that majority of the unigenes from all four libraries fall between 200 − 500 bp.

According to the BLASTX results, approximately 50 % of the sequences in each library showed significant sequence similarities to proteins in the NCBI databases at the cut-off value of $E \leq 10^{-6}$. Three thousand and three sequences (51 %) of 9ME showed that they encode putative amino acid sequences with significant similarities to the sequences deposited in the protein database. Out of the 5332 sequences of 12ME, 2870 (54 %) showed sequence similarity to known proteins. Similarly, 5248 (56 %) MDE sequences had BLAST hits to known proteins in the non-redundant protein database. The highest percentage of BLAST hits were observed for leaf in which 7530

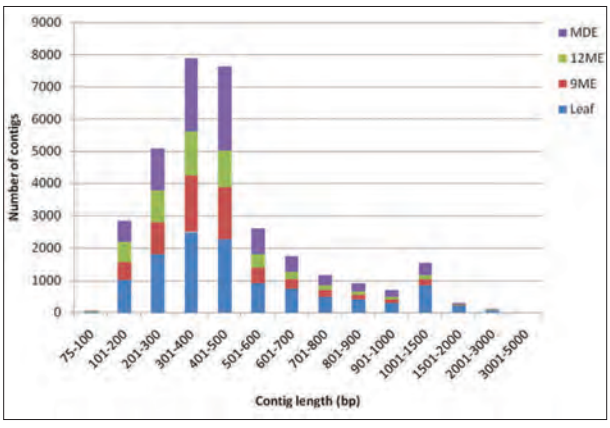

Figure 1: Length distribution of coconut contigs/unigenes in four EST libraries

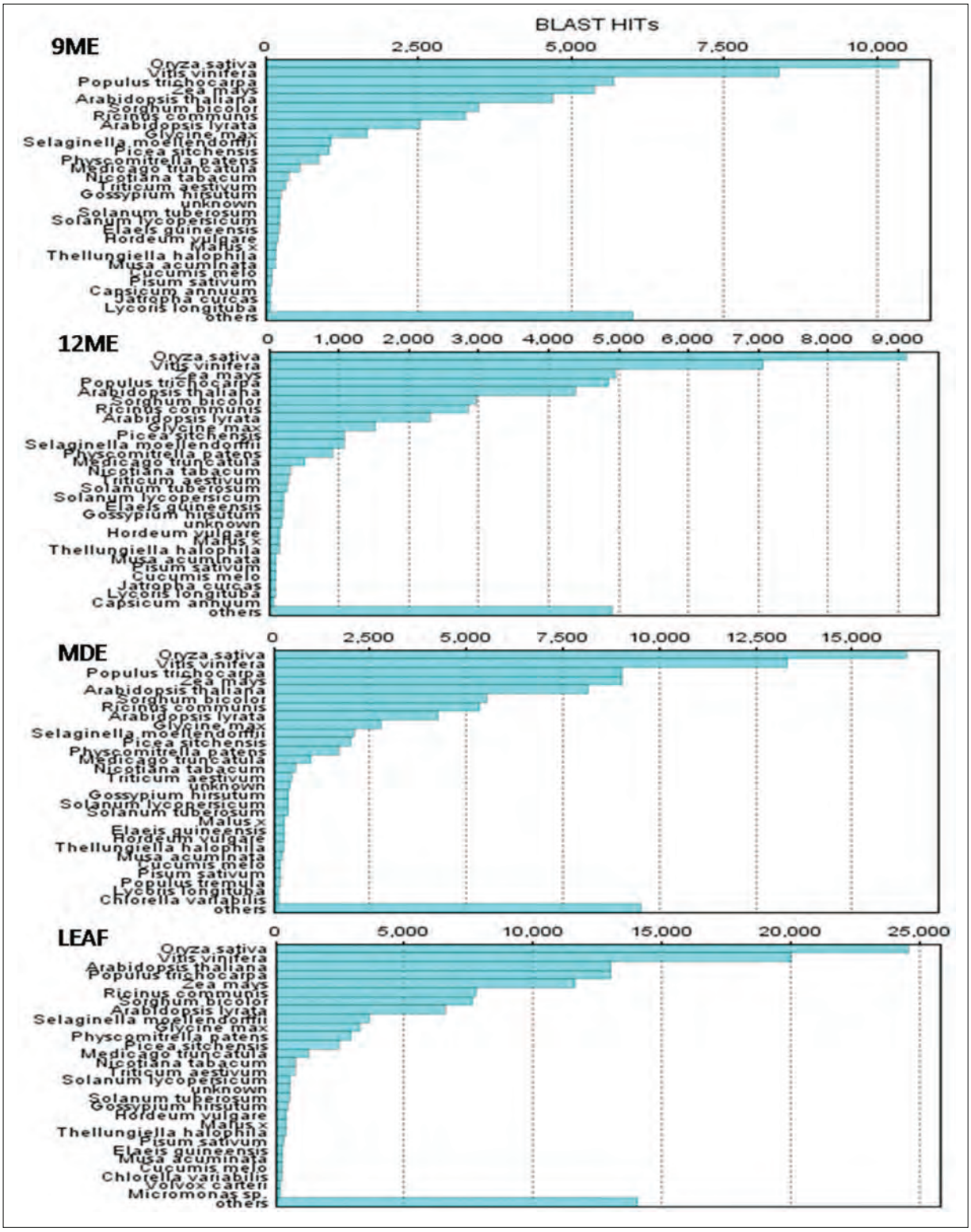

Figure 2: Species distribution chart of BLASTX similarity matches in different coconut EST libraries

(65 %) of the 11653 sequences showed positive hits with databases. The species distribution chart of BLAST hits showed that the majority of sequences of all four tissue types have protein similarity with *Oryza sativa* followed by *Vitis vinifera.* The next three top hits for the species distribution were *Populus trichocarpa, Zea mays* and *Arabidopsis thaliana* with more than 10000 hits per species even though the order differs slightly among the four libraries (Figure 2). Of the annotated unigenes 462 were shared by all four libraries. Numbers of tissue specific unigenes were 365 in 9ME, 335 in 12ME, 1012 in MDE and 1384 in LEAF (Figure 3).

Putative identity of gene abundance in different libraries

The EST abundance in the clustered unigenes in each library was considered to determine the expression of transcripts as the libraries were not normalised. The most highly expressed top 20 genes in each library are given in Table 2. This table was generated by comparing the BLAST2GO data and considered only the genes, which showed significant matches to the protein database. One of the common features in embryo tissue libraries was the occurrence of higher number of ESTs, which do not have significant sequence similarities to any known proteins in databases when counted for the top 20 highly abundant ESTs (30 in 9ME, 25 in 12ME and 17 in MDE). However a proportion of these ESTs (6/30 in 9ME, 11/25 in 12ME and 7/17 in MDE), which did not show similarities to the protein databases had significant matches to the EST database when *Elaeis guineensis* was selected as the organism at default settings. It was noted that compared to the embryo tissue libraries, the leaf library had only two ESTs with no matches to the protein database when searched for the top 20 abundant ESTs. The most widely expressed gene in 9ME encoded a thaumatin-like protein, which was represented by 599 ESTs (unigene 4994). In 12ME, the most highly expressed gene (276 ESTs; unigene 4845) was coded for class IV chitinase. Interestingly an AP2 gene, which has an amino acid sequence similar to that of the oil palm AP2 gene [oil palm homologue (Morcilla *et al*., 2007) of *CnANT* (Bandupriya *et al*., 2013; 2014)] was among the predominant transcripts of 12ME. The most abundant transcript detected in MDE was identified as a ribosomal protein [60s ribosomal protein l31 (unigene 8147, 461 ESTs), 60s ribosomal protein l28 (unigene 1912, 409 ESTs)] followed by glutamine synthetase (unigene 3945, 406 ESTs). The most abundant transcript (unigene 10691) in leaf library, which had 654 reads was annotated as Transketolase 1 (EC: 2.2.1.1), which is one of the enzymes associated with the KEGG pathway of carbon fixation in photosynthetic organisms. In addition, an EST encoding glyceraldehyde-3-phosphate dehydrogenase (EC: 1.2.1.12), another enzyme in the same pathway was among the 20 most abundant ESTs in leaf library. As expected, there were other ESTs encoding annotated proteins such as haeme binding protein 2, photosystem ip 700 apoprotein a2 and chloroplast oxygen-evolving enhancer protein 1, which have important functions in photosynthesis among abundant ESTs in the leaf library. On the contrary, these putative abundance genes could not be identified in the embryo EST libraries. Nonetheless, the top 20 transcripts in the LEAF library were significantly different from the embryo EST library.

Similarities between embryo tissue libraries (9ME, 12ME and MDE) were noticed by comparing some of the genes, which are encoded by the highest abundant transcripts among them. Chitinase was the only putative gene present in all three embryo specific libraries among the top 20 ESTs. However, there were six other genes, which occurred within the top 20 ESTs in either of these two embryo libraries. These genes encoded beta glucanase (represented in 9ME & MDE); ATP synthase CF0 subunit I, alcohol dehydrogenase1 (represented in 9ME & 12ME), reversibly glycosylated polypeptide and 60S ribosomal protein L31 (represented in 12ME & MDE). However, transcripts for the putative elongation factor 1 gene involved in the housekeeping functions of cells were found abundantly in all four libraries. Homology search of abundantly expressed putative unigenes in embryo transcriptome libraries, which are likely to play a role during embryogenesis showed unique domains similarities (Figure 4) to the similar genes in the GenBank.

Gene ontology annotation

In each of the four libraries, approximately 86 % of the unigenes, which showed positive BLAST hits with known proteins could be annotated into one of the three GO categories. Figure 5 shows the functional classification of coconut unigenes in each library into different GO categories within the molecular function main category. GO level 3 was used for the annotation and construction of the pie charts. The total number of GO accessions assigned for molecular function for LEAF, 9ME, 12ME and MDE libraries were 8810, 4745, 4154 and 7462, respectively. A large proportion of GO assigned sequences (78 % of LEAF, 75 % of 9ME, 77 % of 12ME and 75 % of MDE) in this category fell into eight GO categories, namely, nucleotide binding, ion

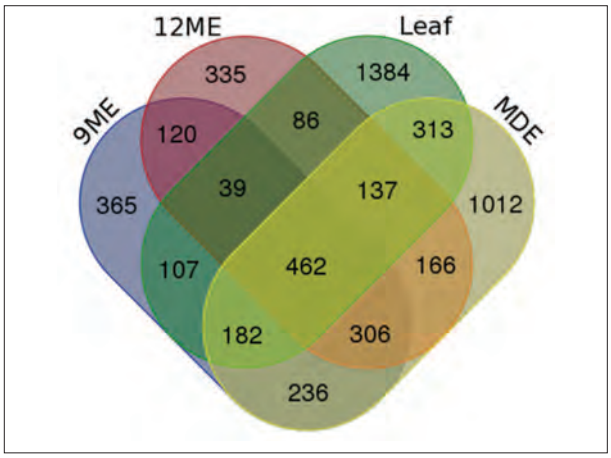

Figure 3: Venn chart showing unique and shared unigenes found in four coconut transcriptomes libraries

binding, protein binding, nucleic acid binding, nucleoside binding, transferase activity, hydrolase activity and oxidoreductase activity, which are components of the two major GO categories of binding and catalytic activity. Of these, assignments to the nucleotide binding ontology made up the majority in LEAF and 12ME libraries, while tranferase activity ontology contained the majority of the 9ME and MDE library GO terms. Basically, there were 16 other common molecular function GO terms for all four libraries excluding previously described eight categories, which finally gave 24 common categories. However, these molecular functions were represented at low levels. Yet, there were some sub categories, which were only found in one particular library or shared by two libraries. For example, oxygen binding GO molecular function was found only in the 9ME library. Moreover, there were two unique GO categories (enzyme inhibitor activity and

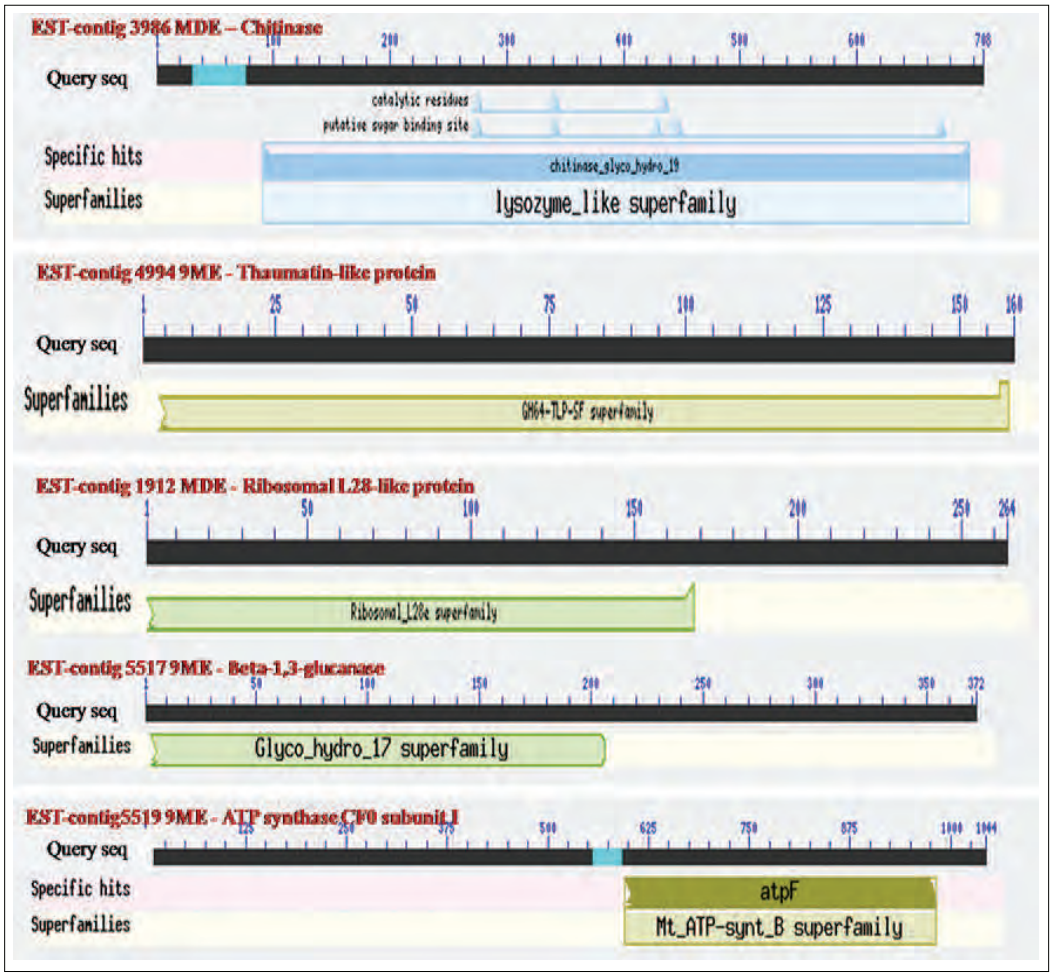

Figure 4: Unique domain similarities in the homology search of abundantly expressed putative unigenes in embryo transcriptome libraries. Query seq is the respective unigene sequence which is given immediately above

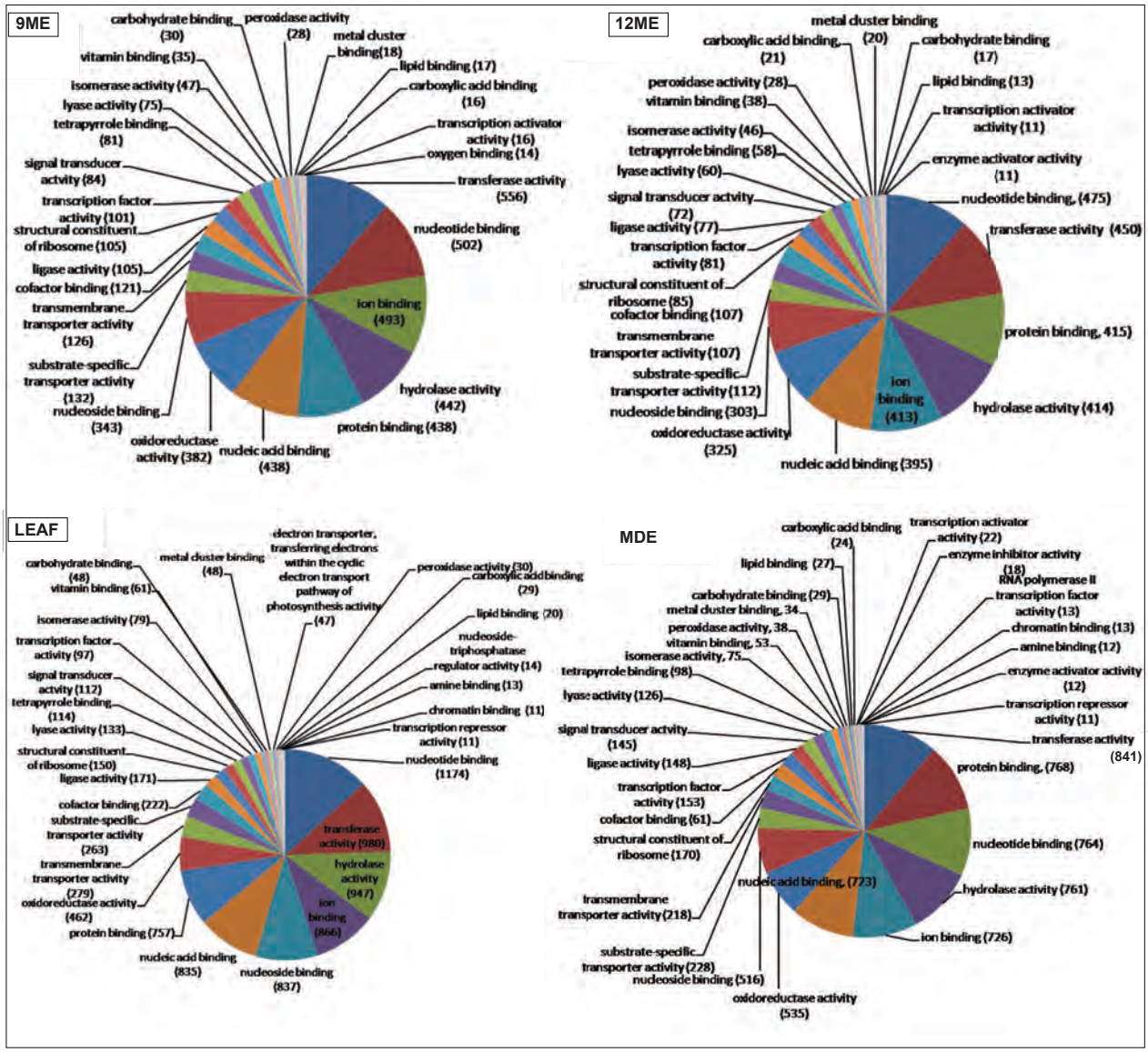

Figure 5: GO term distribution of ESTs in different tissue libraries within the moleular function category. The GO Level 3 pie chart was constructed with the default settings.

RNA polymerase II transcription factor activity) in the MDE library and two other unique GO terms (nucleoside-triphosphatase regulator activity and electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity) in the LEAF library. In addition to that, three GO categories; amine binding, transcription repressor activity and chromatin binding, were found in both MDE and LEAF libraries while the enzyme activator activity term was shared by 12ME and MDE libraries. Another important observation was that the transcription activator activity category was present in all three embryo tissue libraries while it was absent in the LEAF library.

Assignment of GO accessions for biological process at level 2 is shown in Figure 6. This gave a total of 5390, 4692, 8681 and 7744 GO terms for 9ME, 12ME, MDE and LEAF libraries, respectively. There were 15 GO categories shared by all four libraries. In the LEAF library, the unique category nitrogen utilisation was observed in addition to the above categories and also immune system process category, which was present

in all three embryo tissue libraries was absent in LEAF library. Within these categories, the majority of GO assigned sequences were classified into two major functions; metabolic process and cellular process. The percentages of sequences assigned for each of the above categories were quite similar in embryo tissue libraries (metabolic process: ~29 % and cellular process: 28 %). However, the number of ESTs present in these two categories (metabolic process: 37 % and cellular process: 33 %) in the LEAF library was remarkably high. In embryo tissue libraries, the two categories described above were followed by biological regulation (10 %), response to stimulus (8 %), localisation (5 %), developmental process (4 %), multicelluar organism process (4 %), cellular compartment (3 %), signalling (2 %), reproduction (2 %), multi-organism process (2 % in 9ME and MDE) and cellular component biogenesis (2 % in 12ME). It was interesting to see the distribution of ESTs in the above categories at similar proportions in all embryo tissue libraries. However, the order of distribution of ESTs within these categories and the percentage distribution were different in LEAF library when compared to the embryo tissue libraries. Also, it was worth noting the much lower representation of the GO annotations for reproduction category in the LEAF library compared to other embryo tissue libraries.

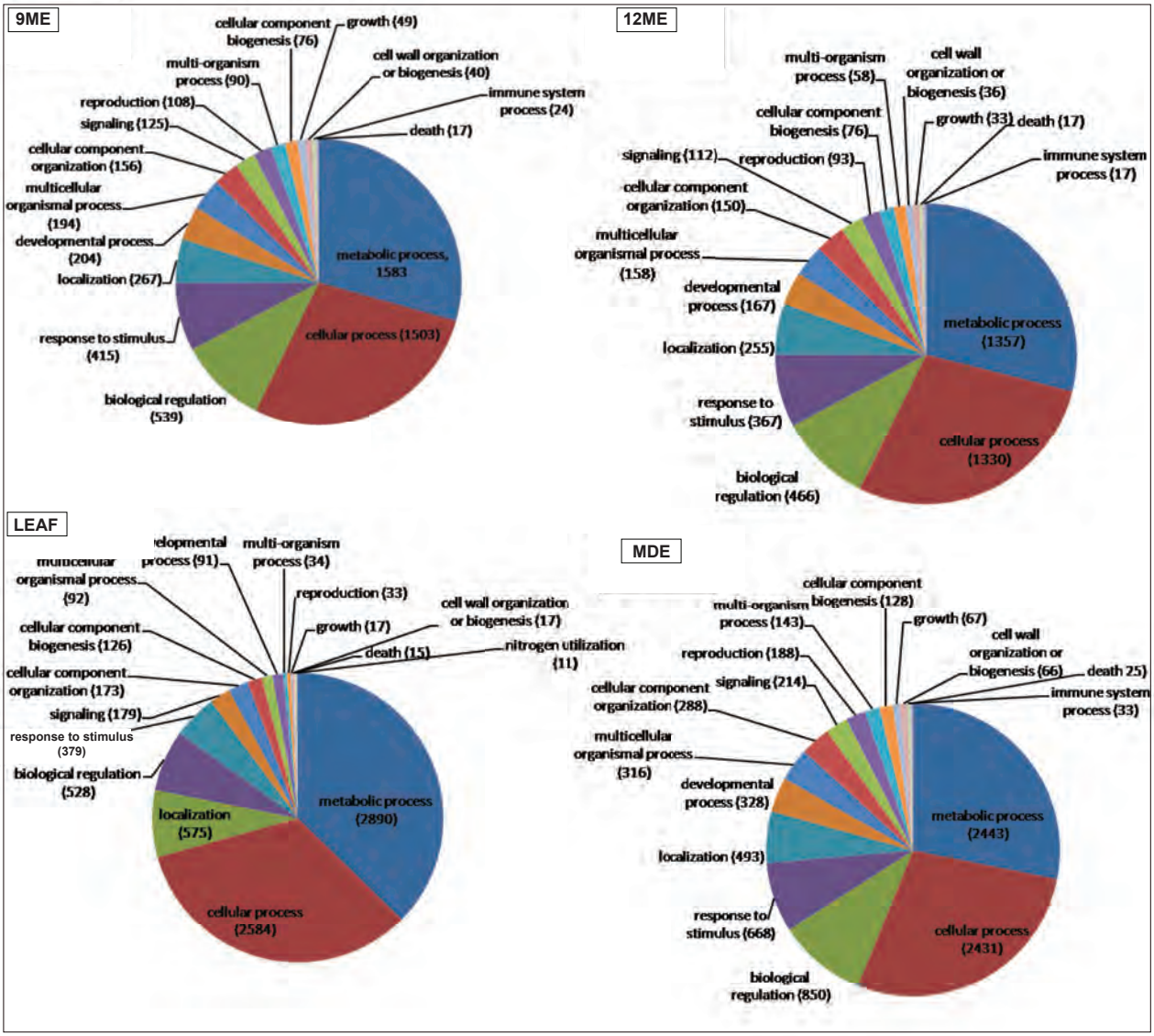

Figure 6: Functional distribution of coconut unigenes associated with biological process GO terms in 9ME, 12ME, MDE and LEAF libraries at GO level 2

Considering the cellular component GO catogary, 9ME and 12ME libraries consisted of 18 cellular component GO categories in which 2166 and 2094 GO terms were distributed, respectively. In the MDE library, 3152 GO terms were assigned into 20 cellular component categories. 2557 sequences were categorised into 19 cellular components in the LEAF library (Figure 7). The comparison between the libraries revealed that the percentage occurrence of unigenes in different cellular components show some similarities among certain libraries. For example, it was noted that the most represented categories in embryo tissue libraries were cytoplasmic membrane-bounded

Table 2: Putative identity of the 20 most abundant sequences in different EST libraries

Unigene number	No. of EST	Putative identity	Species	GI Number	E value	EC
LEAF						
10691	654	Transketolase 1	*Capsicum annuum*	3559814	0	2.2.1.1
2281	512	Ribosomal protein l3	*Ricinus communis*	37625023	0	3.6.5.3
11302	438	yt521-b-like family	*Oryza sativa*	115455327	0	-
8176	354	hypothetical protein	*Oryza sativa*	125545646	7.80E-28	-
10284	313	heme binding protein 2	*Oryza sativa*	115435220	1.13E-17	-
11604	308	glycine dehydrogenase	*Flaveria anomala*	438003	2.19E-127	1.4.4.2
9545	293	polyamine oxidase	*Zea mays*	218184397	8.02E-70	2.7.7.49, 1.5.3.11
8985	287	cationic amino acid transporter	*Populus trichocarpa*	222862313	0.00E+00	-
11136	286	vacuolar h+ translocating inorganic pyrophosphatase	*Oryza sativa*	115466734	2.43E-107	3.6.1.1
9330	286	elongation factor	*Elaeis guineensis*	192910732	1.78E-42	-
9470	280	serine decarboxylase	*Oryza sativa*	125539802	1.78E-126	4.1.1.0
647	277	Photosystem ip 700 apoprotein a2	*Elaeis olefera*	156598249	0.00E+00	-
615	269	RNA polymerase beta subunit	*Elaeis olefera*	156598158	0.00E+00	2.7.7.6
10994	266	glyceraldehyde-3-phosphate dehydrogenase	*Glycine max*	217071898	7.62E-176	1.2.1.12, 1.2.1.13
10719	266	multidrug pheromone mdr abc transporter family	*Sorghum bicolor*	212276142	6.70E-56	-
9402	266	MYC transcription factor	*Vitis vinifera*	4321762	5.90E-136	-
11089	247	galactinol synthase	*Vitis vinifera*	157358428	2.30E-55	2.4.1.123
505	245	lhy protein	*Oryza sativa*	147856747	3.80E-126	-
9676	234	chloroplast oxygen-evolving enhancer protein 1	*Leymus chinensis*		1.00E-07	-
246	231	inorganic pyrophosphatase	*Nicotiana tabacum*		1.30E-77	3.6.1.1
9ME						
4994	599	thaumatin-like protein	*Vitis vinifera*	33329390	2.40E-17	-
5201	405	chalcone synthase	*Elaeis oleifera*	154354073	2.00E-62	2.3.1.74
5467	338	beta-1,3-glucanase	*Elaeis guineensis*	192910882	8.50E-28	3.2.1.39, 3.2.1.73
5954	249	beta-1,4-glucanase	*Dendrobium* hybrid cultivar	148509076	6.10E-76	3.2.1.4
5213	228	N-rich protein	*Glycine max*	57898928	1.60E-10	-
5178	224	vegetative storage protein PNI288	*Zea mays*	195620590	3.70E-10	-
5521	220	Hypothetical protein	*Vitis vinifera*	223531402	1.40E-17	-
283	218	galactosyltransferase	*Ricinus communis*	223545424	4.00E-18	2.4.1.134
5510	215	Sterol methyltransferase	*Arabidopsis thaliana*	15240691	4.00E-39	2.1.1.41
5284	190	3-n-debenzoyltaxol n-benzoyltransferase-like	*Oryza sativa*	9558426	3.00E-126	2.3.1.0
5522	191	DNA binding	*Oryza sativa*	115476528	2.00E-53	-
5490	188	elongation factor 1 b alpha subunit	*Elaeis guineensis*	192910732	5.00E-15	-
5875	188	gcpe protein	*Hevea brasiliensis*	164605000	1.00E-69	1.17.4.3
5153	173	phosphoribosylanthranilate transferase	*Oryza sativa*	115461410	1.50E-32	-
5243	171	ZIM motif family protein expressed	*Oryza sativa*	108708686	2.70E-13	-
5638	171	class IV chitinase	*Vitis vinifera*	29608460	2.90E-23	3.2.1.14
312	170	alcohol dehydrogenase1	*Coix lacryma-jobi*	217069784	2.60E-154	-
4473	170	NADH:flavin oxidoreductase	*Arabidopsis thaliana*	5059115	2.10E-18	1.3.1.42
5519	167	ATP synthase CF0 subunit I	*Elaeis oleifera*	156598114	8.80E-65	3.6.5.3, 3.6.3.14
6283	158	protein binding protein, putative	*Ricinus communis*	255588826	2.70E-12	-

continued -

- continued from page 328

Unigene number	No. of EST	Putative identity	Species	GI Number	E value	EC
12ME						
4845	276	class IV chitinase	*Pisum sativum*	1705807	1.40E-33	3.2.1.14
5259	249	unnamed protein	*Vitis vinifera*	157356287	4.13E-87	2.4.1.186
3106	235	Elongation factor	*Populus trichocarpa*	7489318	2.0E-5	-
4343	213	auxin-repressed kda protein	*Zea mays*	195612466	1.15E-11	-
41	211	AINTEGUMENTA-like gene	*Elaeis guineensis*	56567285	1.66E-166	-
226	187	24-sterol C-methyltransferase	*Gossypium hirsutum*	73761691	8.76E-31	2.1.1.41
4331	172	pyrophosphate-dependent phosphofructo-1-kinase	*Elaeis oleifera*	55296628	7.54E-48	2.7.1.11; 2.7.1.90
4292	162	udp arabinose mutase	*Pisum sativum*	2130521	3.81E-28	2.4.1.186; 2.4.1.12
4412	161	alcohol dehydrogenase I	*Coix lacryma-jobi*	217069784	5.53E-154	-
4291	156	hexokinase 2	*Nicotiana tabacum*	45387409	2.76E-10	2.7.1.1
5088	153	phosphofructokinase, putative	*Ricinus communis*	223544315	1.0E-16	
4917	151	alcohol dehydrogenases	*Populus trichocarpa*	224060281	2.26E-52	1.1.1.1
4616	150	sequence-specific dna binding transcription factor	*Oryza sativa*	223542458	2.40E-58	-
159	147	atp synthase cf0 subunit I	*Medicago truncatula*	153012207	2.94E-12	3.6.3.14
4288	147	reversibly glycosylated polypeptide	*Oryza sativa*	4158232	1.14E-35	2.4.1.186; 2.4.1.12
227	146	sterol 24-c-methyltransferase	*Oryza sativa*	3560531	1.11E-62	2.1.1.0
5243	143	actin depolymerizing	*Elaeis guineensis*	7330254	5.91E-42	-
4334	137	cycloartenol c-24 methyltransferace	*Dioscorea zingiberensis*	30881481	1.60E-10	2.1.1.0
4713	132	hexokinase 3	*Nicotiana sylvestris*	50512102	9.80E-11	2.7.1.2
4708	131	chitinase	*Nicotiana tabacum*	116323	2.50E-14	-
MDE						
8147	461	60s ribosomal protein l31	*Populus trichocarpa*	224106051	2.84E-10	3.6.5.3
1912	409	60s ribosomal protein l28	*Elaeis guineensis*	192908658	4.60E-37	3.6.5.3
3945	406	glutamine synthetase	*Ricinus communis*	1934758	3.49E-21	6.3.1.2
9071	382	reversibly glycosylated polypeptide	*Vitis vinifera*	223546230	2.94E-170	2.4.1.186
7749	364	dihydrolipoamide dehydrogenase precursor	*Capsicum annuum*	44662784	2.89E-47	1.8.1.4
8836	349	DnaJ protein	*Daucus carota*	10945669	1.64E-138	-
8151	329	Unknown protein	*Zea mays*	195605062	1.64E-10	-
8153	318	leucoanthocyanidin dioxygenase /Antocynin synthase	*Anthurium andraeanum*	118183646	1.77E-52	1.13.11.0; 1.14.11.19
9090	305	endo- -beta-glucanase	*Oryza sativa*	75144679	5.60E-118	3.2.1.4
8878	300	phospholipase d	*Oryza sativa*	108935871	6.03E-115	3.1.4.4
8112	290	metallothionein-like protein type 2	*Typha latifolia*	13491970	4.26E-19	3.6.5.3
9251	285	elongation factor 1	*Elaeis guineensis*	192910732	2.81E-52	-
656	279	Unknown protein	*Vitis vinifera*	195604292	1.21E-106	-
8770	273	class 1 chitinase	*Elaeis guineensis*	223545207	1.97E-150	3.2.1.14
731	272	glutamine synthetase	*Hevea brasiliensis*	2213877	3.88E-81	6.3.1.2
3986	260	chitinase	*Musa acuminata*	17932712	2.49E-123	3.2.1.14
8652	256	rna polymerase beta subunit	*Yucca schidigera*	69216898	2.50E-84	2.7.7.6
9059	256	putative stress protein	*Ricinus communis*	255564142	9.30E-46	-
8927	252	60S ribosomal protein L31	*Oryza sativa*	42407684	8.20E-10	3.6.5.3
8073	243	tubulin alpha chain	*Trichinella spiralis*	134142167	5.20E-12	-

vesicle (18 % of 9ME, 16 % of 12ME and 19 % of MDE) and plasma membrane (10, 11, 13 % of 9ME, 12ME and MDE, respectively) followed by protein complex (9, 11 % and DE, respectively) and integral to membrane (9 % of 9ME & 12ME and 11 % of MDE), while 27 % of LEAF unigenes were classified as cytoplasmic membrane-bounded vesicle followed by integral to membrane (16 %) and plasma membrane (10 %). The category protein complex, which represented a major component in embryo tissue libraries, was absent in the LEAF library.

Assignment of KEGG biochemical pathways

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway assignment attempt predicted a total

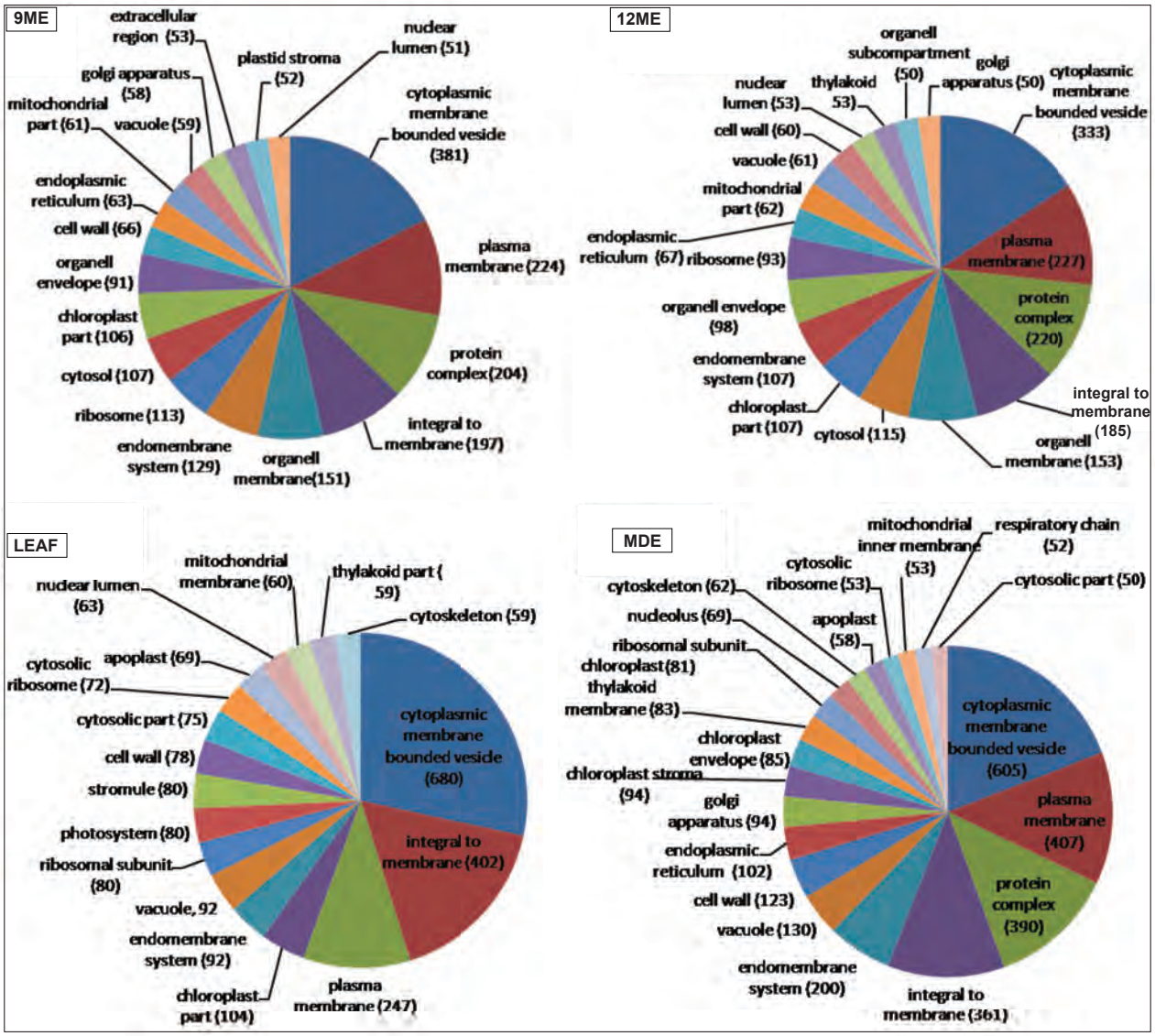

Figure 7: Distribution of coconut EST GO terms associated with cellular component in 9ME, 12ME, MDE and LEAF libraries

of 139 pathways for coconut ESTs. Of the sequences in 9ME library, 1225 sequences (19 % of the total unigenes from 9ME) having 1557 enzyme commission (EC) numbers were assigned into 131 metabolic pathways. 1010 unigenes (19 % of the total unigenes of 12ME library) had a match to 1333 enzyme codes, which were mapped into 125 pathways in 12ME library. In MDE, 128 pathways were identified for 2330 enzyme codes distributed among 1800 unigenes (19 % of the MDE unigenes). The LEAF library sequences showed the highest percentage (23 %) of unigenes (2659) having enzyme codes (3592) distributed among 125 KEGG pathways. According to this analysis it was clear that the predicted pathways represent most of the plant metabolic and/or biochemical pathways for biosynthesis, utilisation, assimilation, degradation, and detoxification. Among the fourteen major pathway categories identified for the coconut ESTs, a higher number of ESTs was represented in carbohydrate metabolism, energy metabolism, lipid metabolism, amino acid metabolism and nucleotide metabolism. In that case, carbohydrate metabolism was the top ranked KEGG category in all four libraries (22, 30, 36 and 31 % in LEAF, 9ME, 12ME and MDE, respectively). The KEGG map analysis of different EST libraries pointed out that LEAF library ESTs and embryo tissue library ESTs show different sequence abundance

in KEGG pathways. The second largest group of ESTs represented amino acid metabolism (23 % in 9ME and 21 % in 12ME and MDE) in these embryo tissue libraries while the second highest EST bearing pathway in LEAF was energy metabolism. Interestingly, the biosynthesis of secondary metabolites, which was represented only by 4 % of the KEGG ESTs of LEAF library took the fourth position of the most abundant KEGG pathway unigenes in embryo tissue libraries by occupying 14, 8 and 11 % of the KEGG pathway assigned ESTs in 9ME, 12ME and MDE, respectively.

Several enzymes involved in major plant metabolic pathways including purine metabolism enzymes, glycolysis/gluconeogenesis, starch and sucrose metabolism, oxidative phosphorylation, methane metabolism and carbon fixation in photosynthetic organisms were identified. Apart from these pathways, enzymes mapped into the important pathways including citrate cycle (TCA), pentose phosphate pathway, pentose and glucuronate interconversions, fructose and mannose metabolism and galactose metabolism were identified among the coconut ESTs. Moreover, genes encoding in several secondary metabolite biosynthesis pathways including phenylpropanoid biosynthesis, terpenoid backbone biosynthesis, flavonoid biosynthesis and isoquinoline alkaloid biosynthesis were mapped into the unigenes derived from 454 cDNA libraries. Although majority of the KEGG pathways were well represented by the ESTs of all four libraries, still there were some pathways, which could be identified only in embryo tissue libraries but not in the LEAF library. The higher number of such pathways, which were identified only in embryo tissue libraries was in the subclass of glycan biosynthesis and metabolism. Out of the twenty eight enzymes identified as embryo tissue specific, 17 were from different pathways under glycan biosynthesis and metabolism. However, of the 28 enzymes only 8 were represented in all three embryo tissue libraries while the rest were either specific to one library (13) or represented in two of the three libraries (7).

DISCUSSION

In the present study, four coconut transcriptome populations from different tissue types were sequenced by 454 GS-FLX system to produce a total of 979448 sequence reads, which cover 2.2 x 10^8 base pairs of data representing a substantial sequence resource for coconut. The analysis of sequence results showed that the reads generated from 454 sequencing can be efficiently assembled through the generation of 32621 quality unigenes and used as a resource to characterise functional categories of this non model organism. However, the actual number of total ESTs could be less than the number given here as the different unigenes could represent different portions of the same gene. Average read lengths of 260, 222, 201 and 227 were obtained for LEAF, 9ME, 12ME and MDE libraries, respectively, which is within the capacity of 454 GS-FLX and consistent with the results obtained in previous studies (Novaes *et al*., 2008; Meyer *et al*., 2009; Wang *et al*., 2009; Zeng *et al*., 2010). However, after releasing the 454 GS-FLX titanium series, the average read length has been substantially increased (Calduch-Giner *et al.*, 2013) in this technology. After assembly, the unigenes of the coconut libraries were on average lengths ranging from 418 bp in 12ME library to 522 bp in LEAF library and this is reported to be larger than the unigene lengths assembled using the 454 GS-FLX instrument in some of the previous studies [e.g. 246 bp (Zeng *et al.*, 2010); 197 bp (Vera *et al.*, 2008); 247 bp (Novaes *et al*., 2008)].

The presence of unknown or unclassified ESTs in large proportions is always associated with transcriptome analysis studies (Low *et al*., 2008; Bettencourt *et al.*, 2010; Sun *et al*., 2010). Similarly in the present study, the percentage of unigenes in all four libraries without any significant hit in the nr protein database ranged between 35 % of LEAF unigenes to 49 % of MDE. In certain studies, the use of low stringency levels has resulted in a higher number of similarity hits (Bettencourt *et al*., 2010; Sun *et al*., 2010). The fact that a few ESTs in embryo tissue libraries compared to LEAF library gave matches to the database may suggest that a proportion of these ESTs might have encoded new genes with specific functions during embryo development.

In this study, EST data were used to compare and identify the genes, which are expressed abundantly in different libraries, assuming that a higher number of reads in a particular unigene represent higher a number of mRNA molecules encoding that gene in a given EST library as described in previous studies (Ho *et al*., 2007; Costa *et al*., 2010). The ESTs identified in different libraries showed that all four libraries are informative. According to the comparative expression analysis, it was demonstrated that the embryo tissue libraries share certain degree of similarities while they were different from Leaf abundant ESTs. As might be expected, several genes such as transketolase 1, glyceraldehyde-3-phosphate dehydrogenase, pyrophosphatase, haeme binding protein 2, photosystem ip 700 apoprotein a2 and chloroplast oxygen-evolving enhancer protein 1 related to carbon fixation

in photosynthetic organisms, were among the highly expressed genes in coconut leaf data. This provides evidence for the effectiveness of 454 sequencing approach for the identification of transcripts in a particular organ.

The presence of chitinase among the top 20 most highly expressed transcripts in all three embryo tissue libraries gave evidence for its involvement during embryogenesis. Chitinases affect the early phases of embryo development and it could rescue somatic embryo development in mutant lines (De Jong *et al*., 1992), play a crucial role in embryogenesis by exhibiting higher expression mainly in transcriptome analysis in different species [e.g. *Cyclamen persicum* (Rensing *et al.*, 2005); oil palm (Ho *et al.*, 2007)] and hydrolyse arabinogalactan proteins (van Hengel *et al.*, 2001), which are structurally complex macromolecules showing a positive relationship with somatic embryogenesis (Majewska-Sawka & Nothnagel, 2000). The abundance levels of chitinase in the embryo libraries were suggestive of its involvement in embryo development and should be considered as a candidate gene for further studies.

Of the embryo tissue ESTs, there were a few more ESTs, which showed increased level of transcripts encoded for genes described during somatic and zygotic embryogenesis of oil palm. Beta-1,3-glucanase, ATP synthase CF0 subunit (Ho *et al.*, 2007), thaumatin like protein and metallothionein-like protein (Ho *et al.*, 2007) are some of the genes previously characterised during oil palm embryogenesis. Since the oil palm whole genome sequencing project has come to an end (Singh *et al*., 2013), the data gathered from this study will have substantial value for the comparative studies between coconut and oil palm in the future. The possible involvement of highly expressed transcripts in embryo tissue libraries encoded for glutamine synthetase, which plays a central role in nitrogen metabolism (Miflin & Habash, 2002) and two other components; reversibly glycosylated polypeptide and beta-1,4-endoglucanase, which are cell wall associated proteins (Dhugga *et al.*, 1997), could be explained by the rapidly dividing nature of cells in these tissues. The involvement of cell wall in signal transduction, formation of tensions influencing cell division planes and rebuilding of the cell wall are characteristics during embryogenesis (Malinowski & Filipecki, 2002). A number of proteins have been identified as abundantly present in plant seeds inhibiting the growth of phytopathogenic fungi (Velazhahan *et al.*, 2001). Grouping of 1,3-glucanases (Leah *et al*., 1991) and thaumatin-like proteins (Roberts & Selitrennikoff, 1990; Vigers *et al.*, 1992) among them suggested the possible involvement of these genes against pathogenic fungi once the embryo started germination. Also, the presence of transcripts encoding C-24 sterol methyltransferase in embryo library (9ME), an enzyme involved in sterol biosynthesis, which has been previously reported as important during *Arabidopsis* zygotic embryogenesis, (Schrick *et al*., 2002) suggest determining changes in sterol composition, which is related to cell polarity required for auxin efflux.

To obtain an overview of the gene functions of different EST libraries of coconut, the annotated EST unigenes were categorised into different GO terms using BLAST2GO. These analyses showed that the EST libraries cover a wide range of biological functions. A majority of the GO terms in molecular function ontology were comprised with proteins involved in binding or catalytic activity, which is comparable with the observations made previously in both embryogenesis related EST data (Rensing *et al*., 2005) and developing seeds (Costa *et al*., 2010). However, no major differences were observed in highly represented GO term associated categories between the leaf library and the other embryo tissue library, suggesting that a major portion of transcripts of different tissues are involving in similar functions. Nonetheless, specific allocation of certain GO categories into a particular library might have a relationship with the function of that particular tissue at that time. For example oxygen binding sub category, which is represented only in 9ME has been previously reported in the oil palm initiation library but not in the proliferation library (Lin *et al.*, 2009), suggesting the GO terms associated with early stages of embryogenesis. When the biological process ontology is considered, cellular process and metabolic process were the main subcategories bearing higher number of GO terms in all four libraries. In embryogenic tissues, this can be explained by having cells at dividing stages involving cell cycle thus using energy for cell maintenance as explained for a ciliate species (Lokanathan *et al*., 2010). However regardless of the organism, a majority of GO terms have been grouped into these two subcategories of biological process from different plant and animal species. Noteworthy differences could be observed between the LEAF and embryo tissue libraries when considering cellular component ontology. Differences in cellular structure between the two types of tissues may have contributed towards this. The leaf tissues used in this analysis were obtained from *in vitro* raised plantlets, which had a continuous supply of exogenous nutrients. This might have some impact on the gene expression in leaf tissues as they were not totally self

maintaining tissues. It is noteworthy to mention that the comparison of GO term distribution among different tissue types including embryogenic and vegetative in oil palm has not revealed significant differences (Ho *et al.,* 2007).

Since its release (Ogata *et al*., 1999), the KEGG database data has been used widely as a reference tool for the interpretation of large scale EST data sets (Alagna *et al*., 2009; Lokanathan *et al*., 2010; Sun *et al.,* 2010). Annotations analysis of the KEGG pathways on coconut EST data provided better understanding of the basic physiology of coconut embryo and leaf (vegetative) tissues. Genes in the major KEGG metabolism pathway categories; carbohydrate, energy, lipid and amino acid, were stored abundantly showing that the embryos are utilising resources for the rapid development. Glycolysis/gluconeogenesis, TCA cycle, carbon fixation and starch and sucrose metabolism pathways, which play key roles in material utilisation and energy production were represented more in the embryo libraries than in the LEAF library providing more clues on the physiology of embryos that have intense and rapid cell division. These results are consistant with the transcriptomic functional analysis data of a hybrid rice (Ge *et al*., 2008) suggesting a similar molecular biology underlying monocot embryogenesis. The KEGG analysis also permitted to identify some secondary metabolite biosynthesis pathways in embryo libraries such as phenylpropanoid biosynthesis, tropane, piperidine and pyridine alkaloid biosynthesis, flavonoid biosynthesis and terpenoid backbone biosynthesis represented by higher number of ESTs. It has been previously documented that phenolic acids, intermediates of phenylpropanoid metabolism, function in several ways during plant embryogenesis (mainly somatic embryogenesis) (Cvikrova *et al*., 2003). The functions of phenolic acids include auxin polar transport during embryogenesis (Jacobs & Rubery, 1988) and the involvement of cell wall structure alterations (Cvikrova *et al*., 1991). The flavonoid biosynthesis genes were up-regulated during somatic embryogenesis of *Medicago truncatula* (Mantiri *et al*., 2008) and have also been related to stress protection (Winkel-Shirley, 2002). As revealed by the higher abundant ESTs in each library chalcone synthase (CHS), which catalyses the first step in flavonoid biosynthesis was one of the members in 9ME library in coconut in this study. It has been reported that CHS is transcriptionally activated by GA (Weiss *et al.*, 1990) and sucrose (Tsukaya *et al.*, 1991); essential components during embryo development. Also, CHS has been studied as a stress inducible gene in cell culture systems (Tsukaya *et al.*, 1991).

In addition to the above pathways, some KEGG pathways, a majority from the sub categories of glycan biosynthesis and metabolism, were determined only in the embryo tissue libraries. ESTs encoding these enzymes might be a valuable resource for further studies. Although much data are not available on glycan biosynthesis in plants, there are reports in this regard during embryogenesis of animal species. For example cell surface glycans have been used to visualise the development of the enveloping layer during the early stages of zebrafish embryogenesis (Baskin *et al*., 2010). A number of other tissue dependent pathways may account for biologically associate differences among the tissues. Further characterisation of the genes represented in these pathways may enhance the understanding of the functions of these genes during embryogenesis. Nonetheless, more candidate genes involved in biosynthesis pathways were identified through this study although they were of low abundance.

In summary, we applied the 454 Next Generation Sequencing technology and the *de novo* assembly to generate 20968 unigenes from three embryo developmental stages and 11653 unigenes from leaf tissues. The types and quantities of genes expressed in coconut embryos, their functions, and metabolic pathways were revealed. In this study, a list of putative transcripts such as chitinase, beta-1,3-glucanase, ATP synthase CF0 subunit, thaumatin-like protein and metallothionein-like protein, which may be involved in various biological processes during embryogenesis were identified to provide a valuable resource for further studies. Moreover, further investigation of novel transcripts, which have no significant homology in public databases will pave the way to identify the potential candidates involved in embryogenesis.

Acknowledgement

We are thankful to Dr L.K. Weerakoon, Former Head, Tissue Culture Division, Coconut Research Institute, Lunuwila, Sri Lanka for her assistance in sending coconut samples, and to Dr Andrew Meade and Dr Anushka Wicramasuriya for their kind assistance during data analysis. Authors are also thankful to Mr. Prasad Sanjeewa for the extended assistance in improving the quality of the pictures. H.D. Dharshani Bandupriya wishes to thank the Association of Commonwealth Universities in the UK for the award of a Commonwealth Scholarship, during the tenure of which this work was carried out.

REFERENCES

1. Alagna F., D'Agostino N., Torchia L., Servili M., Rao R., Pietrella M., Giuliano G., Chiusano M.L., Baldoni L. & Perrotta G. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* **10**: 399. DOI: http://dx.doi.org/10.1186/1471-2164-10-399
2. Bandupriya H.D.D., Gibbings J.G. & Dunwell J.M. (2013). Isolation and characterization of an *AINTEGUMENTA-like* gene in different coconut (*Cocos nucifera* L.) varieties from Sri Lanka. *Tree Genetics and Genomes* **9**(3): 813 − 827. DOI: http://dx.doi.org/10.1007/s11295-013-0600-5
3. Bandupriya H.D.D., Gibbings J.G. & Dunwell J.M. (2013). Overexpression of coconut AINTEGUMENTA-like gene, CnANT, promotes *in vitro* regeneration in transgenic *Arabidopsis. Plant Cell Tissue and Organ Culture* **116**(1): 67 – 79. DOI: http://dx.doi.org/10.1007/s11240-013-0383-2
4. Barakat A., DiLoreto D.S., Zhang Y., Smith C., Baier K., Powell W.A, Wheeler N., Sederoff R. & Carlson J.E. (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC* Plant Biology **9**: 51. DOI: http://dx.doi.org/10.1186/1471-2229-9-51
5. Barnes W.M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America* **91**(6): 2216 – 2220. DOI: http://dx.doi.org/10.1073/pnas.91.6.2216
6. Baskin J.M., Dehnert K.W., Laughlin S.T., Amacher S.L. & Bertozzi C.R. (2010). Visualizing enveloping layer glycans during zebrafish early embryogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **107**(23): 10360 – 10365. DOI: http://dx.doi.org/10.1073/pnas.0912081107
7. Bettencourt R., Pinheiro M., Egas C., Gomes P., Afonso M., Shank T. & Santos R.S. (2010). High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus. BMC Genomics* **11**: 559. DOI: http://dx.doi.org/10.1186/1471-2164-11-559
8. Calduch-Giner J.A., Bermejo-Nogales A., Benedito-Palos L., Estensoro I., Ballester-Lozano G., Sitjà-Bobadilla A. & Pérez-Sánchez J. (2013). Deep sequencing for de novo construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. *BMC Genomics* **14**(9): 178. DOI: http://dx.doi.org/10.1186/1471-2164-14-178
9. Chen S. *et al*. (12 authors) (2011). 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng. Plant Cell Reports* **30**: 1593 – 1601. DOI: http://dx.doi.org/10.1007/s00299-011-1070-6
10. Conesa A., Gotz S., Garcia-Gomez J.M., Terol J., Talon M. & Robles M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674 – 3676. DOI: http://dx.doi.org/10.1093/bioinformatics/bti610
11. Costa G.G.L. *et al*. (12 authors) (2010). Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L. *BMC Genomics* **11**: 462. DOI: http://dx.doi.org/10.1186/1471-2164-11-462
12. Cvikrova M., Mala J., Hrubcova M., Eder J., Zon J. & Machackova I. (2003). Effect of inhibition of biosynthesis of phenylpropanoids on sessile oak somatic embryogenesis. *Plant Physiology and Biochemistry* **41**: 251 – 259.
13. Cvikrova M., Meravy L., Machackova I. & Eder J. (1991). Phenylalanine ammonia-lyase, phenolic-acids and ethylene in alfalfa (*Medicago sativa* L.) cell-cultures in relation to their embryogenic ability. *Plant Cell Reports* **10**(5): 251 – 255. DOI: http://dx.doi.org/10.1007/BF00232569
14. De Jong A.J., Cordewener L., Lo Schiavo F., Terzi M., Vandekerckhove J., Vankammen A. & de Vries S.C. (1992). A carrot somatic embryo mutant is rescuded by chitinase. *Plant Cell* **4**(4): 425 – 433. DOI: http://dx.doi.org/10.1105/tpc.4.4.425
15. Dhugga K.S., Tiwari S.C. & Ray P.M. (1997). A reversibly glycosylated polypeptide (RGP1) possibly involved in plant cell wall synthesis: purification, gene cloning, and trans-Golgi localization. *Proceedings of the National Academy of Sciences of the United States of America* **94**(14): 7679 – 7684. DOI: http://dx.doi.org/10.1073/pnas.94.14.7679
16. Firon N. *et al*. (11 authors) (2013). Transcriptional profiling of sweet potato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC Genomics* **14**: 460. DOI: http://dx.doi.org/10.1186/1471-2164-14-460
17. Ge X.M., Chen W., Song S.H., Wang W.W., Hu S.N. & Yu J. (2008). Transcriptomic profiling of mature embryo from an elite super-hybrid rice LYP9 and its parental lines. *BMC Plant Biology* **8**: 114. DOI: http://dx.doi.org/10.1186/1471-2229-8-114
18. Hahn D.A., Ragland G.J., Shoemaker D.D. & Denlinger D.L. (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis. BMC Genomics* **10**: 234. DOI: http://dx.doi.org/10.1186/1471-2164-10-234
19. Ho C-L. *et al*. (14 authors) (2007). Analysis and functional anotation of expressed sequence tags (ESTs) from multiple

tissues of oil palm (*Elaeis guineensis* Jacq.). *BMC Genomics* **8**: 381.
DOI: http://dx.doi.org/10.1186/1471-2164-8-381

20. Jacobs M. & Rubery P.H. (1988). Naturally-occurring auxin transport regulators. *Science* **241**: 346 – 349.
DOI: http://dx.doi.org/10.1126/science.241.4863.346
21. Kyndt T., Denil S., Haegeman A., Trooskens G., Bauters L., Van Criekinge W., De Meyer T. & Gheysen G. (2012). Transcriptional reprogramming by root knot and migratory nematode infection in rice. *New Phytologist* **196**: 887 – 900.
DOI: http://dx.doi.org/10.1111/j.1469-8137.2012.04311.x
22. Kristiansson E., Asker N., Forlin L. & Larsson D.G.J. (2009). Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics* **10**: 345.
DOI: http://dx.doi.org/10.1186/1471-2164-10-345
23. Leah R., Tommerup H., Svendsen I. & Mundy J. (1991). Biochemical and molecular characterization of 3 barley seed proteins with antifungal properties. *Journal of Biological Chemistry* **266**: 1564 – 1573.
24. Lin H.C., Morcillo F., Dussert S., Tranchant-Dubreuil C., Tregear J.W. & Tranbarger T.J. (2009). Transcriptome analysis during somatic embryogenesis of the tropical monocot *Elaeis guineensis*: evidence for conserved gene functions in early development. *Plant Molecular Biology* **70**(1): 173 – 192.
DOI: http://dx.doi.org/10.1007/s11103-009-9464-3
25. Lokanathan Y., Mohd-Adnan A., Wan K.L. & Nathan S. (2010). Transcriptome analysis of the *Cryptocaryon irritans* tomont stage identifies potential genes for the detection and control of cryptocaryonosis. *BMC Genomics* **11**: 76.
DOI: http://dx.doi.org/10.1186/1471-2164-11-76
26. Low E.L., Alias H., Boon S., Shariff E.M., Tan C.A., Ooi L.C.L., Cheah S., Raha A., Wan K. & Singh R. (2008). Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: identifying genes associated with callogenesis and embryogenesis. *BMC Plant Biology* **8**: 62.
DOI: http://dx.doi.org/10.1186/1471-2229-8-62
27. Majewska-Sawka A. & Nothnagel E.A. (2000). The multiple roles of arabinogalactan proteins in plant development. *Plant Physiology* **122**: 3 – 9.
28. Malinowski R. & Filipecki M. (2002). The role of cell wall in plant embryogenesis. *Cellular and Molecular Biology Letters* **7**: 1137 – 1151.
29. Mantiri F.R., Kurdyukov S., Lohar D.P., Sharopova N., Saeed N.A., Wang X.D., VandenBosch K.A. & Rose R.J. (2008). The transcription factor MtSERF1 of the ERF subfamily identified by transcriptional profiling is required for somatic embryogenesis induced by auxin plus cytokinin in *Medicago truncatula. Plant Physiology* **146**(4): 1622 – 1636.
DOI: http://dx.doi.org/10.1104/pp.107.110379
30. Margulies M. *et al.* (56 authors) (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376 – 380.
DOI: http://dx.doi.org/10.1038/nature03959
31. Meyer E., Aglyamova G.V., Wang S., Buchanan-Carter J., Abrego D., Colbourne J.K., Willis B. & Matz M.V. (2009). Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**: 219.
DOI: http://dx.doi.org/10.1186/1471-2164-10-219
32. Miflin B.J. & Habash D.Z. (2002). The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops. *Journal of Experimental Botany* **53**: 979 – 987.
DOI: http://dx.doi.org/10.1093/jexbot/53.370.979
33. Mochida K. & Shinozaki K. (2010). Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiology* **51**(4): 497 – 523.
DOI: http://dx.doi.org/10.1093/pcp/pcq027
34. Novaes E., Drost D.R., Farmerie W.G., Pappas G.J., Grattapaglia D., Sederoff R.R. & Kirst M. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.
DOI: http://dx.doi.org/10.1186/1471-2164-9-312
35. Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. & Kanehisa M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**: 29 – 34.
DOI: http://dx.doi.org/10.1093/nar/27.1.29
36. Perera P.I.P., Hocher V., Verdeil J-L., Bandupriya H.D.D., Yakandawala D.M.Y. & Weerakoon L.K. (2008). Androgenic potential of coconut (*Cocos nucifera* L.). *Plant Cell Tissue and Organ Culture* **92**: 293 – 302.
DOI: http://dx.doi.org/10.1007/s11240-008-9337-5
37. Rensing S.A., Lang D., Schumann E., Reski R. & Hohe A. (2005). EST sequencing from embryogenic *Cyclamen persicum* cell cultures identifies a high proportion of transcripts homologous to plant genes involved in somatic embryogenesis. *Journal of Plant Growth Regulation* **24**: 102 – 115.
DOI: http://dx.doi.org/10.1007/s00344-005-0033-y
38. Roberts W.K. & Selitrennikoff C.P. (1990). Zeamatin, an antifungal protein from maize with membrane-permeabilizing activity. *Journal of General Microbiology* **136**: 1771 – 1778.
DOI: http://dx.doi.org/10.1099/00221287-136-9-1771
39. Schrick K., Mayer U., Martin G., Bellini C., Kuhnt C., Schmidt J. & Jurgens G. (2002). Interactions between sterol biosynthesis genes in embryonic development of *Arabidopsis. The Plant Journal* **31**: 61 – 73.
DOI: http://dx.doi.org/10.1046/j.1365-313X.2002.01333.x
40. Sharma S.K., Millam S., Hedley P.E., McNicol J. & Bryan G.J. (2008). Molecular regulation of somatic embryogenesis

in potato: an auxin led perspective. *Plant Molecular Biology* **68**(1): 185 – 201.
DOI: http://dx.doi.org/10.1007/s11103-008-9360-2

41. Shin H., Hirst M., Bainbridge M.N., Magrini V., Mardis E., Moerman D.G., Marra M.A., Baillie D.L. & Jones S.J.M. (2008). Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biology* **6**: 30.
DOI: http://dx.doi.org/10.1186/1741-7007-6-30
42. Singh R. *et al.* (28 authors) (2013). Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* **500**: 335 – 339.
DOI: http://dx.doi.org/10.1038/nature12309
43. Sniady V., Becker D., Herrán A., Ritter E. & Rohde W. (2003). A rapid way of physical mapping in coconut and oil palm. Available at *http://www.tropentag.de/2003/abstracts/full/282*.
44. Sun C., Li Y., Wu Q., Luo H., Sun Y., Song J., Lui E.M.K. & Chen S. (2010). *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* **11**: 262.
DOI: http://dx.doi.org/10.1186/1471-2164-11-262
45. Tsukaya H., Ohshima T., Naito S., Chino M. & Komeda Y. (1991). Sugar-dependent expression of the chs-a gene for chalcone synthase from petunia in transgenic *Arabidopsis. Plant Physiology* **97**(4): 1414 − 1421.
DOI: http://dx.doi.org/10.1104/pp.97.4.1414
46. van Hengel A.J., Tadesse Z., Immerzeel P., Schols H., van Kammen A. & de Vries S.C. (2001). N-acetylglucosamine and glucosamine-containing arabinogalactan proteins control somatic embryogenesis. *Plant Physiology* **125**(4): 1880 – 1890.
DOI: http://dx.doi.org/10.1104/pp.125.4.1880
47. Velazhahan R., Radhajeyalakshmi R., Thangavelu R. & Muthukrishnan S. (2001). An antifungal protein purified from pearl millet seeds shows sequence homology to lipid transfer proteins. *Biologia Plantarum* **44**: 417 – 421.
DOI: http://dx.doi.org/10.1023/A:1012463315579
48. Vega-Arreguin J.C., Ibarra-Laclette E., Jimenez-Moraila B., Martinez O., Vielle-Calzada J.P., Herrera-Estrella L. & Herrera-Estrella A. (2009). Deep sampling of the *Palomero* maize transcriptome by a high throughput strategy of pyrosequencing. *BMC Genomics* **10**: 299.
DOI: http://dx.doi.org/10.1186/1471-2164-10-299
49. Vera J.C., Wheat C.W., Fescemyer H.W., Frilander M.J., Crawford D.L., Hanski I. & Marden J.H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**: 1636 – 1647.
DOI: http://dx.doi.org/10.1111/j.1365-294X.2008.03666.x
50. Vigers A.J., Wiedemann S., Roberts W.K., Legrand M., Selitrennikoff C.P. & Fritig B. (1992). Thaumatin-like pathogenesis-related proteins are antifungal. *Plant Science* **83**: 155 – 161.
51. Wang W., Wang Y.J., Zhang Q., Qi Y. & Guo D.J. (2009). Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10**: 465.
DOI: http://dx.doi.org/10.1186/1471-2164-10-465
52. Weerakoon L.K., Vidhanaarachchi V.R.M., Fernando S.C., Fernando A. & Gamage C.K.A. (2002). Increasing the efficiency of embryo culture technology to promote coconut germplasm collecting and exchange in Sri Lanka. *Coconut Embryo in vitro Culture Part II* (eds. F. Engelmann, P. Batugal & J.T. Oliver). IPGRI-APO, Serdang, Malaysia.
53. Weiss D., Vantunen A.J., Halevy A.H., Mol J.N.M. & Gerats A.G.M. (1990). Stamens and gibberellic-acid in the regulation of flavonoid gene-expression in the corolla of *Petunia-hybrida. Plant Physiology* **94**: 511 – 515.
DOI: http://dx.doi.org/10.1104/pp.94.2.511
54. Weber A.P.M., Weber K.L., Carr K., Wilkerson C. & Ohlrogge J.B. (2007). Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144**: 32 – 42.
DOI: http://dx.doi.org/10.1104/pp.107.096677
55. Winkel-Shirley B. (2002). Biosynthesis of flavonoids and effects of stress. *Current Opinion in Plant Biology* **5**: 218 – 223.
56. Xu H., Gao Y. & Wang J. (2012). Transcriptomic analysis of rice (*Oryza sativa*) developing embryos using the RNA-Seq technique. *PLoS One* **7**: e30646.
DOI: http://dx.doi.org/10.1371/journal.pone.0030646
57. Zeng S.H., Xiao G., Guo J., Fei Z.J., Xu Y.Q., Roe B.A. & Wang Y. (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* **11**: 94.
DOI: http://dx.doi.org/10.1186/1471-2164-11-94
58. Zhu Y.Y., Machleder E.M., Chenchik A., Li R. & Siebert P.D. (2001). Reverse transcriptase template switching: a SMART approach for full length cDNA library construction. *Biotechniques* **30**: 892 – 897.