Physiol Mol Biol Plants
DOI 10.1007/s12298-017-0429-8

RESEARCH ARTICLE

Genome-wide transcriptome profiling of *Carica papaya* L. embryogenic callus

Nur Diyana Jamaluddin[1] · Normah Mohd Noor[1] · Hoe-Han Goh[1]

Received: 7 December 2016 / Revised: 24 February 2017 / Accepted: 6 March 2017

Abstract Genome-wide transcriptome profiling is a powerful tool to study global gene expression patterns in plant development. We report the first transcriptome profile analysis of papaya embryogenic callus to improve our understanding on genes associated with somatic embryogenesis. By using 3′ mRNA-sequencing, we generated 6,190,687 processed reads and 47.0% were aligned to papaya genome reference, in which 21,170 (75.4%) of 27,082 annotated genes were found to be expressed but only 41% was expressed at functionally high levels. The top 10% of genes with high transcript abundance were significantly enriched in biological processes related to cell proliferation, stress response, and metabolism. Genes functioning in somatic embryogenesis such as SERK and LEA, hormone-related genes, stress-related genes, and genes involved in secondary metabolite biosynthesis pathways were highly expressed. Transcription factors such as NAC, WRKY, MYB, WUSCHEL, Agamous-like MADS-box protein and bHLH important in somatic embryos of other plants species were found to be expressed in papaya embryogenic callus. Abundant expression of enolase and ADH is consistent with proteome study of papaya somatic embryo. Our study highlights that some genes related to secondary metabolite biosynthesis, especially phenylpropanoid biosynthesis, were highly expressed in papaya embryogenic callus, which might have implication for cell factory applications. The discovery of all genes expressed in papaya embryogenic callus provides an important information into early biological processes during the induction of embryogenesis and useful for future research in other plant species.

Keywords 3′ mRNA sequencing · Embryogenic callus · Papaya · RNA-seq · Transcriptome

Abbreviations

ADH	Alcohol dehydrogenase
BAP	Benzylaminopurine
bHLH	Basic/HELIX–LOOP–HELIX
CPM	Count per million
GO	Gene ontology
GST	Glutathione *S*-transferase
KEGG	Kyoto encyclopaedia of genes and genomes
LEA	Late embryogenesis abundant
NAA	α-Naphthaleneacetic acid
PAL	Phenylalanine ammonia lyase
SERK	Somatic embryogenesis receptor-like kinase

Introduction

Papaya (*Carica papaya* L.) is an important fruit crop in tropical and subtropical regions well known for its nutritional benefits and medicinal properties (Elgadir et al. 2014). Papaya is easy to grow and able to produce flower and fruit throughout the year which makes it a

The collection of sequences generated in this study is available under NCBI BioProject accession PRJNA323966 and Sequence Read Archive (SRA) database (Accession Numbers SRR4087172 and SRR4087196).

Electronic supplementary material The online version of this article (doi:10.1007/s12298-017-0429-8) contains supplementary material, which is available to authorized users.

✉ Hoe-Han Goh
gohhh@ukm.edu.my

[1] Institute of Systems Biology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Published online: 20 March 2017

suitable model fruit tree (da Silva et al. 2007). The availability of completed 372 Mbp papaya genome with various bioinformatics tools and resources accelerates the identification of genes involved in important biological processes (Ming et al. 2008). Molecular studies in papaya have been facilitated by the establishment of in vitro culture and transformation system, including numerous studies in callus induction and regeneration from various explants (Ascencio-Cabral et al. 2008; Bhattacharya and Khuspe 2001; Chen and Chen 1992; Chen et al. 1987; Fitch et al. 1993; Litz and Conover 1981, 1982; Sun et al. 2011; Yu et al. 2000).

Somatic embryogenesis through tissue culture is a popular approach for the production of cultivars with desirable agricultural traits. Somatic embryogenesis-related genes has been extensively characterised in many plant species including Arabidopsis (Gliwicka et al. 2013; Wickramasuriya and Dunwell 2015), maize (Salvo et al. 2014), longan (Lai and Lin 2013), rice (Chen et al. 2011; Xu et al. 2012), soybean (Thibaud-Nissen et al. 2003), potato (Sharma et al. 2008) and oil palm (Lin et al. 2009). A number of genes playing important roles in somatic embryogenesis have been reported such as late embryogenesis abundant (LEA) protein (Wickramasuriya and Dunwell 2015), somatic embryogenesis receptor-like kinase (SERK) (Salvo et al. 2014), WUSCHEL (Gliwicka et al. 2013), AGAMOUS (Gliwicka et al. 2013) and MYB transcription factor (Xu et al. 2012).

To date, there is no transcriptome study on papaya embryogenic callus. Transcriptome profiling of papaya has only been reported in the root (Porter et al. 2008), flower (Urasaki et al. 2012) and fruit (Fabi et al. 2010, 2014). Recently, a proteomic study in papaya callus showed that enolase, esterase and ADH to be potential biomarkers for somatic embryo in papaya (de Moura et al. 2014). Our study aims to understand the molecular mechanism of papaya callus formation by investigating the gene expression profile of embryogenic callus by using RNA-seq with 3′-mRNA sequencing technology. We acquired the expression profile of embryogenic callus to identify embryogenesis-related genes and highlight our findings that genes in secondary metabolite biosynthesis pathways were abundantly present. This study provides the first transcriptome profile which gives insights into the molecular genetics of embryogenic papaya callus.

Materials and methods

Induction of embryogenic callus

Immature green fruits of papaya var. 'sekaki' were harvested from experimental plot at Universiti Kebangsaan Malaysia. Seeds were surface sterilised for 5 min using 20% sodium hypochlorite containing tween 20 and rinsed three times with sterile distilled water. Immature zygotic embryos were cultured on callus induction medium, maintained in a controlled growth chamber (CU-22L, Percival Scientific) at 25 ± 1 °C in the dark. The basal media comprised Murashige and Skoog (MS) medium (Murashige and Skoog 1962), 3 g/L gelrite and 30 g/L sucrose, with pH adjusted to 5.7–5.8. Autoclaved growth media were supplemented with established (Sun et al. 2011) concentrations of plant growth regulators: 1.5 mg/L 2,4-dichlorophenoxy acid (2,4-D), 0.5 mg/L α-naphthaleneacetic acid (NAA), 2.25 mg/L benzylaminopurine (BAP) and 1 mg/L kinetin for callus induction. Induction of somatic embryos was performed using regeneration media consist of 0.02 mg/L NAA and 0.1 mg/L BAP at 25 ± 1 °C in the light. All growth media and hormones were purchased from Duchefa Biochemie.

RNA isolation

Approximately 300 mg of two independent 4-week old embryogenic callus samples at proliferation stage was aseptically harvested and frozen in liquid nitrogen. RNA was extracted using TRIzol (Invitrogen) according to manufacturer's instruction. RNA quality and quantity were determined by using NanoDrop instrument and agarose gel electrophoresis to fulfil all the requirements for QuantSeq library preparation.

Library construction and QuantSeq 3′ mRNA sequencing

Total RNA samples (500 ng) were cleaned using DNAse I kit according to the Rapid out removal DNA kit instruction (Thermoscientific) and converted into cDNA by using QuantSeq 3′ mRNA-seq reverse (REV) Library Prep Kit (Lexogen) according to manufacturer's instruction (Moll et al. 2014) to generate compatible library for Illumina sequencing. Briefly, library generation was initiated by oligodT priming for first strand cDNA which generated one fragment per transcript. The second strand cDNA was subsequently synthesised using random primers. Illumina-specific linker sequences were introduced by the primer with barcoding indices for different samples. The quality of cDNA libraries was determined using a High Sensitivity DNA Assay 2100 Bioanalyzer (Agilent) for quality control analysis. Sequencing of the callus of papaya cDNA library with 100 bp single end reads was performed using Illumina HiSeq 2500 system at Australian Genome Research Facility (AGRF) according to standard protocols.

Transcriptome analysis

Raw sequencing reads from two independent samples (deposited to NCBI SRA database with the accession numbers SRR4087172 and SRR4087196) were processed individually to remove low quality sequences (QV < 10 of 4-base sliding window) and unknown sequences with 'N' using Trimmomatic (Bolger et al. 2014) to retain reads with minimum length of at least 20 bases long. To quantify transcript abundance, the processed reads were mapped to papaya genome reference version Cpapaya_113 (http://www.plantgdb.org/CpGDB/) which was modified to include 500 bp extensions from both ends of the CDS sequences and with the 'N' removed. The mapping was performed using Bowtie2 (Langmead and Salzberg 2012) with stringent "end-to-end" alignment and all other parameters were set to default values according to recommended data analysis workflow by Lexogen (http://www.lexogen.com/quantseq-data-analysis). The reads mapped to the indexed reference were quantified using eXpress (Roberts and Pachter 2013) based on Bowtie2 alignment to estimate gene abundance in count per million (CPM). Genes with average CPM values greater than zero (CPM > 0) were considered expressed and assigned into low, mid and high abundance categories for further analysis.

Functional annotation, gene ontology classification and enrichment analysis

Functional annotation information of the reference papaya gene sequences was obtained from PLAZA 3.0 (http://bioinformatics.psb.ugent.be/plaza/) and further annotated using Trinotate analysis pipeline (http://trinotate.github.io/). Gene ontology (GO) analysis was conducted based on combined GO annotations using WEGO database (Ye et al. 2006) (http://wego.genomics.org.cn/cgi-bin/wego/index.pl). Cytoscape software with Biological Network Gene Ontology (BiNGO) plugin (Maere et al. 2005) were used to determine which GO categories are overrepresented. The first step in BiNGO analysis is to identify homologous gene sequences in *Arabidopsis thaliana* through local BLAST to retrieve GO term associate with BLAST hits. Hypergeometric test with Benjamini and Hochberg false discovery rate (FDR) were performed using default parameters for adjusted *P* value. To identify significantly enriched pathways, we used a web-based tool KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/anno_iden.php) which incorporates several metabolic pathway databases, including KEGG PATHWAY, BioCyc and Panther (Xie et al. 2011). KEGG pathway annotation was performed using KAAS Ver. 2.0 (http://www.genome.jp/kaas-bin/kaas_main, updated April 1, 2015) (Moriya et al. 2007) based on all organisms under plant category (ath, aly, cit, tcc, gmx, fve, csv, vvi, sly, osa, olu, ota, mis, cme, gel), GHOSTX as search program, and SBH (single-directional best hit) as assignment mode.

Results

Establishment of embryogenic callus

Embryogenic calli of *C. papaya* were successfully generated from immature zygotic embryos after 4 weeks of culture on callus induction medium (Fig. 1a). These calli were friable, loose and yellowish in colour and exhibited high somatic embryogenic potential. Some of the callus cells developed into somatic embryos (Fig. 1b) and eventually formed plantlets (Fig. 1c). Two independent 4-week-old embryogenic callus samples with similar morphology were chosen for transcriptome profiling analysis.

Transcriptome analysis of embryogenic callus in *C. papaya*

Transcriptome profiling of papaya embryogenic callus was performed through QuantSeq 3′ mRNA sequencing with Illumina HiSeq 2500 platform generating 100 bp single-end reads (Table 1). A total of 6,190,687 (56.8%) clean processed reads were obtained from 10,891,013 raw reads after filtering and trimming for read alignment to the papaya reference sequences, which achieved 47.0% total alignment rate.

Transcript abundance for each of the 28,072 total annotated genes was estimated from the alignment result to generate a count matrix normalised by the total count to obtain average count per million (CPM) values from the two samples (Supplementary Table S1). The two callus samples showed high Pearson's correlation coefficient (r) of 0.963. Hence, further analysis was based on mean CPM values calculated from the two samples. A total of 6902 (24.6%) genes with CPM values of zero were considered not expressed in the callus (Table 2). The CPM values of 21,170 (75.4%) expressed genes were assigned into low, mid, high, and very high transcript abundance based on the distribution of CPM values. Low abundance category includes 9656 (34.4%) genes with minimal expression (CPM < 5), which suggests that these genes might not be functional in the callus. The highest number of genes were in the mid transcript abundance category (CPM > 5–100) totalling 10,434 (37.2%). Forty-seven (0.1%) genes were very highly expressed at CPM > 1000.

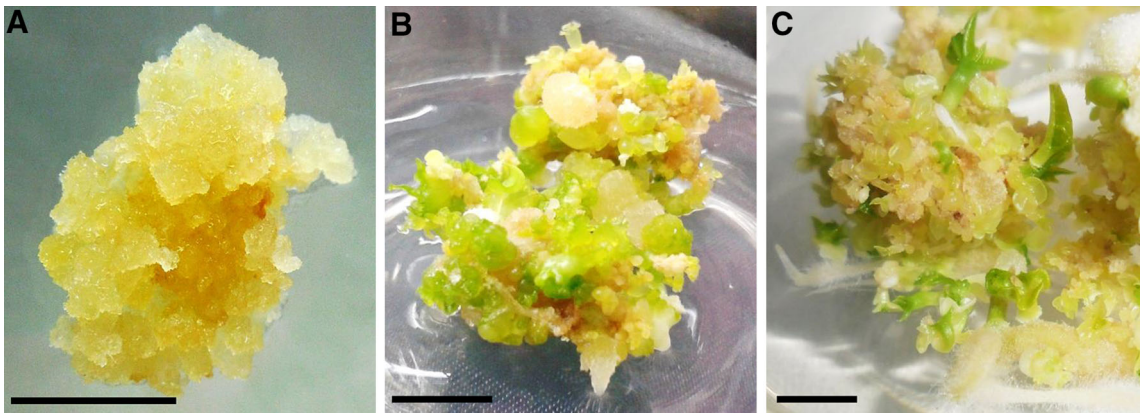

Fig. 1 Development of papaya embryogenic callus. **a** Callus induction from immature zygotic embryos after 4 weeks, **b** formation of somatic embryos after 8 weeks on regeneration medium, **c** formation of plantlets from somatic embryos after 3 months. *Bars* represent 1 cm

Table 1 Statistics of sequencing and read alignment

Attribute	Number of reads
Raw reads	10,891,013
Processed reads	6,190,687
Total aligned	2,912,241

Table 2 Average transcript abundance in embryogenic callus of *C. papaya*

Normalised transcript count (CPM)	Transcript abundance	Number of genes
0	–	6902
>0–5	Low	9656
>5–100	Mid	10,434
>100–1000	High	1033
>1000	Very high	47
	Total	28,072

Functional annotation, gene classification and enrichment analysis

Based on the functional annotation information of the reference papaya gene sequences obtained from PLAZA 3.0, only 15,463 (55%) genes have descriptions. We further improve this annotation information by performing BLASTx analysis with Swiss-Prot using Trinotate analysis pipeline which found hits on 1712 of unannotated genes. As a result, a total of 17,175 (61.2%) genes were functionally annotated for further analysis (Supplementary Table S1).

Gene ontology (GO) analysis was performed for three categories of gene expression, namely, no expression (CPM = 0), low expression (CPM > 0–5) and high expression (CPM > 5). Overall, there was no clear difference between the expressed (CPM > 0) and non-expressed genes in papaya embryogenic callus (Fig. 2). However, virion, virion part, transcription regulator activity, and cell killing were among the GO terms found to be proportionally greater for non-expressed genes. Conversely, genes expressed at low and high levels were proportionally more abundant in 12 out of 14 subcategories of cellular component classification, 13 out of 14 subcategories of molecular function classification, and all 23 subcategories of biological process classification. All GO terms were found to be proportionally greater for highly expressed genes compared with genes with low expression, except for nutrient reservoir activity, protein tag, and cell killing. Notably, synapse, auxiliary transport protein, metallochaperone activity and protein tag were the GO terms unique to expressed genes. This showed that most genes involved in different biological functions were expressed in papaya embryogenic callus, but genes involved in less important processes were expressed at low levels.

Gene ontology (GO) enrichment analysis was performed using BiNGO based on top 10% of genes (N = 281) with high transcript abundance. BiNGO allows the visualisation of overrepresented GO function and the relationship between GO as a network. The output consists of color-coded nodes which represent GO terms while edges represent the relationship between GO terms. In high abundance transcripts, cellular process, metabolic process, and response to stimulus were the most significantly enriched GO terms in biological process (Fig. 3). GO terms associated with gene expression, biosynthetic process, response to stress were also significantly enriched. These significantly enriched GO terms represent the active biological processes in the papaya embryogenic callus.

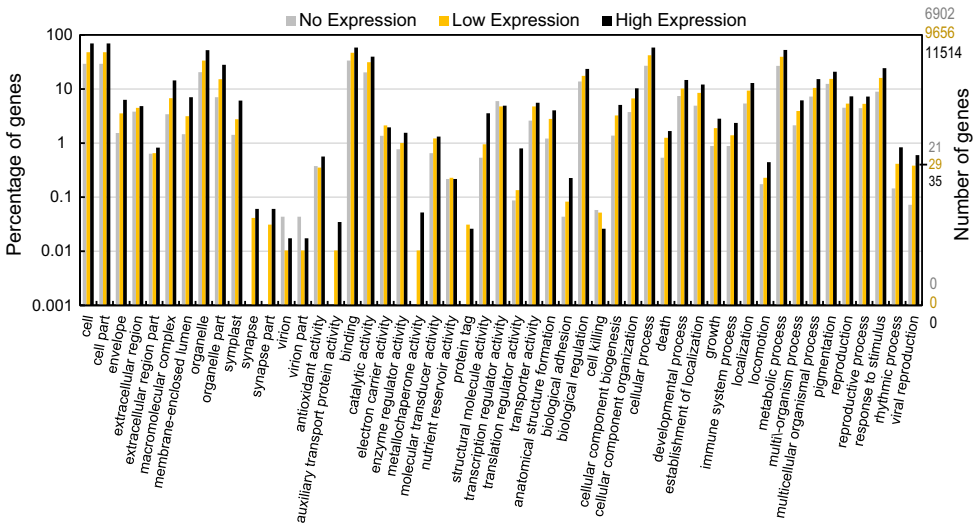

Fig. 2 Gene ontology classification of genes based on expression levels in papaya embryogenic callus. No expression (CPM = 0), low expression (CPM > 0–5), and high expression (CPM > 5). The *X* axis is the definition of GO term and the *Y* axis is the percentage or number of gene mapped by the GO term

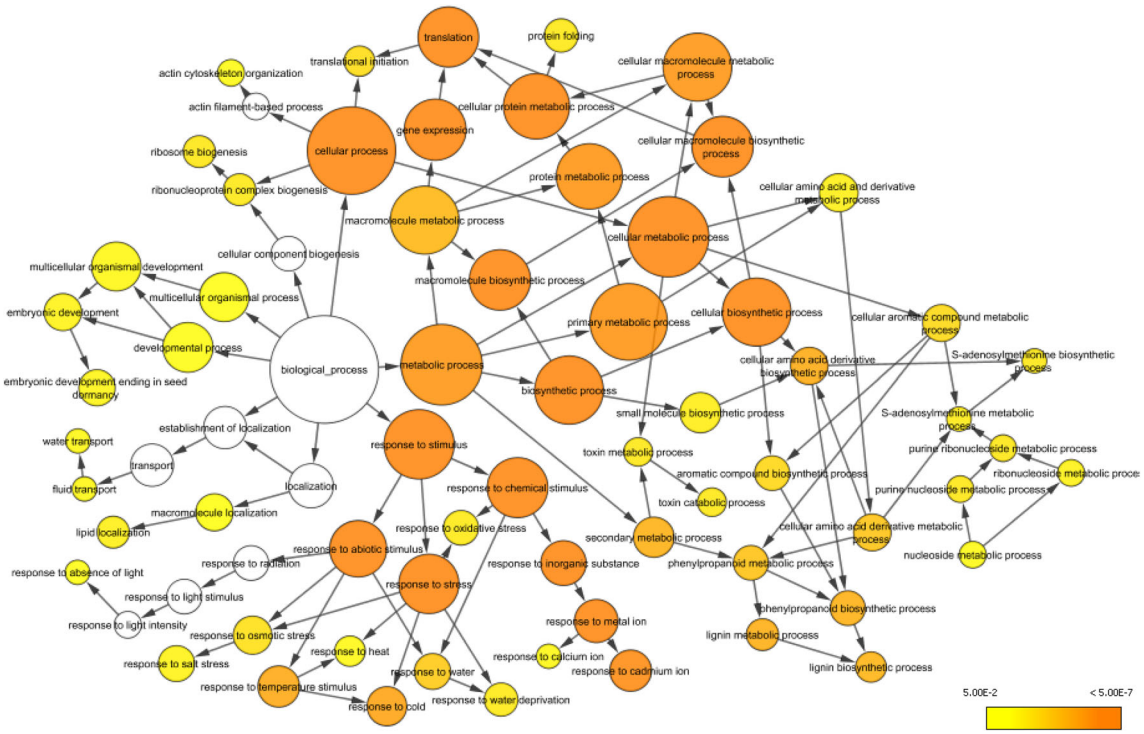

Fig. 3 Overrepresented GO terms (biological process) in top 10% of genes with high abundance transcript. The significance level of the overrepresented GO term is shown in the heatmap. White means not statistically significant at FDR < 0.05

KEGG pathway annotation and enrichment analysis

A total of 347 KEGG pathways were annotated with 28,072 genes in papaya, in which 21,170 expressed genes and 6902 of non-expressed genes were mapped to 345 and 288 KEGG pathways respectively (Supplement Table S2). Most genes in key pathways of metabolism and processes important for cell growth and proliferation were found to be expressed based on the proportion of expressed to non-expressed genes relative to the total mapped entry (Fig. 4). For example, biosynthesis of amino acids, metabolic pathways, cell cycle, plant hormone signal transduction and glycolysis. Notably, all genes involved in riboflavin and thiamine metabolism were active in papaya embryogenic callus. The metabolism of phenylalanine, alanine, aspartate and glutamate appeared to be more active than other amino acids; and likewise for the biosynthesis of arginine, valine, leucine and isoleucine. Interestingly, many genes in the biosynthesis of secondary metabolites, especially phenylpropanoid, flavonoid, and carotenoid

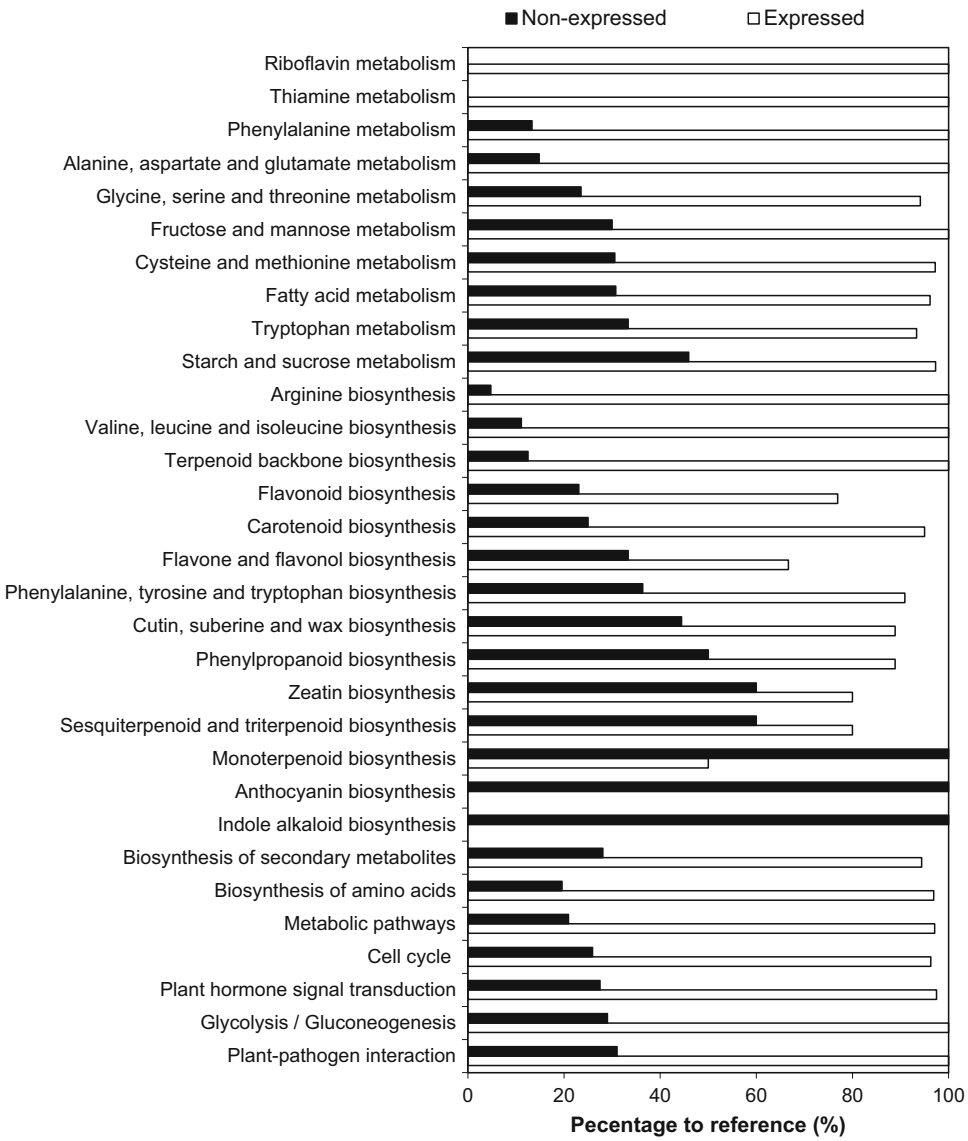

Fig. 4 KEGG pathway annotation of expressed and non-expressed genes in papaya embryogenic callus

were expressed in papaya embryogenic callus. None of the genes involved in the biosynthesis of anthocyanin and indole alkaloid were expressed.

KEGG pathway enrichment analysis of top 10% of genes with high transcript abundance showed that phenylpropanoid biosynthesis was significantly enriched (Table 3). Noteworthy is that many of these genes are involved in phenylpropanoid biosynthesis (Fig. 5), which include phenylalanine ammonia-lyase (PAL), caffeic acid 3-*O*-methyltransferase, trans-cinnamate 4-monooxygenase, cinnamyl alcohol dehydrogenase, cinnamoyl-CoA reductase, and peroxidase (Table 4). Other highly expressed genes of interest include RNA helicase, histone, translation initiation factor, elongation factor and ribosomal proteins for cell proliferation; calmodulin and 14–3–3-like protein for cell signalling; ion transporter and storage proteins for cellular homeostasis; whereas late embryogenesis abundant (LEA) proteins, heat shock protein (HSP), dehydrin, glutathione *S*-transferase, and blue copper protein are related to stress responses during somatic embryogenesis in the dark (Table 4). Genes involved in primary metabolic processes can also be found in abundance, including adenosylhomocysteinase, *S*-adenosylmethionine synthase and asparagine synthetase in amino acid metabolism, and glyceraldehyde-3-phosphate dehydrogenase in glycolysis (Table 4). This is consistent with the result from GO enrichment analysis.

Discussion

Papaya embryogenic calli which can regenerate into plantlets (Fig. 1) were successfully induced by previously established culture media (Sun et al. 2011) and two

Table 3 Pathway enrichment analysis of top 10% of genes with high transcript abundance

Pathway	Database	*P* value	Corrected *P* value
Ribosome	KEGG PATHWAY	6.4E−35	1.1E−32
Phenylpropanoid biosynthesis	KEGG PATHWAY	6.4E−07	1.1E−05
RNA transport	KEGG PATHWAY	7.2E−06	1.1E−04
Methionine degradation I	BioCyc	2.4E−05	3.4E−04
S-adenosyl-L-methionine cycle II	BioCyc	0.0001	0.0008
Glutathione metabolism	KEGG PATHWAY	0.0001	0.0010
Biosynthesis of secondary metabolites	KEGG PATHWAY	0.0002	0.0021
Glutathione-mediated detoxification II	BioCyc	0.0005	0.0048

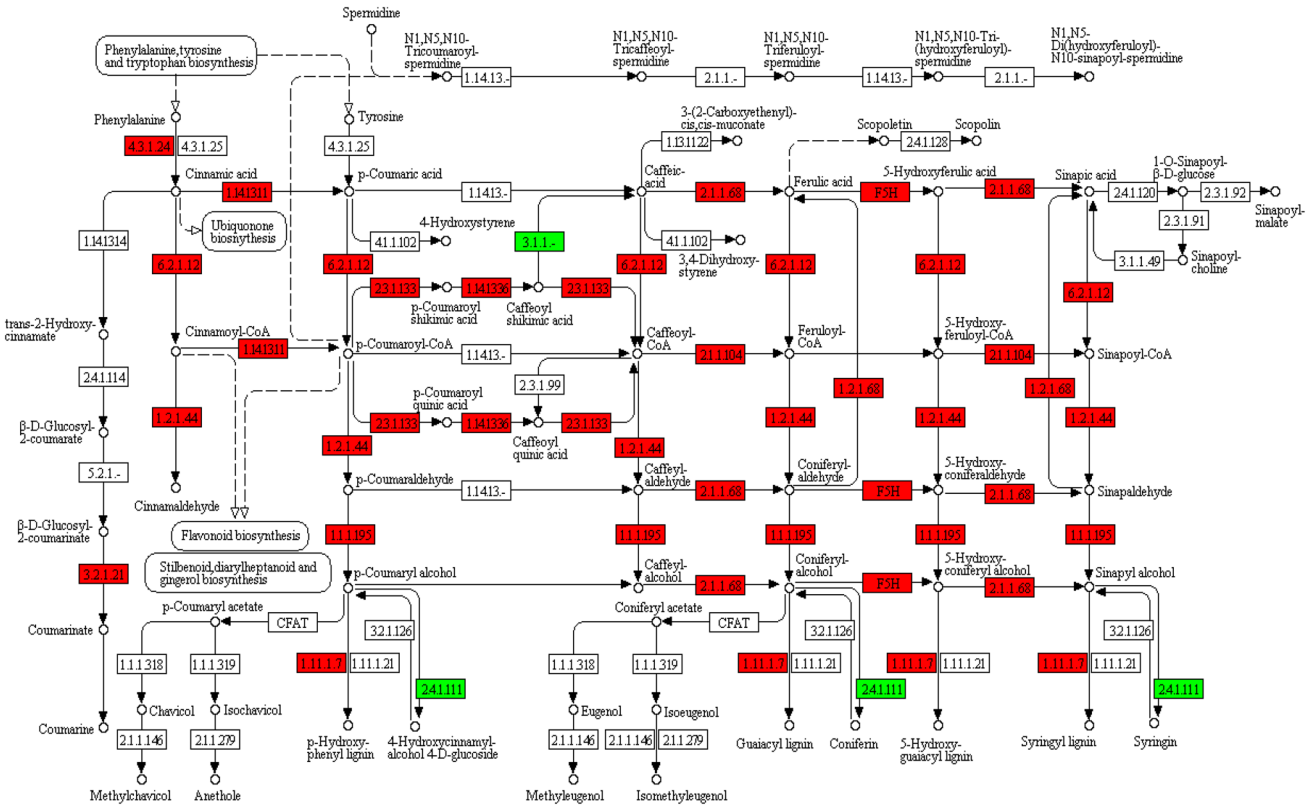

Fig. 5 Expression of phenylpropanoid biosynthesis genes in papaya embryogenic callus. Expressed (*red*) and non-expressed genes (*green*) were mapped. Mapped genes include phenylalanine ammonia-lyase [4.3.1.24], caffeic acid 3-*O*-methyltransferase [2.1.1.68], trans-cinnamate 4-monooxygenase [1.14.13.11], cinnamyl alcohol dehydrogenase [1.1.1.195], cinnamoyl-CoA reductase [1.2.1.44], peroxidase [1.11.1.7], caffeoyl-CoA *O*-methyltransferase [2.1.1.104], coumaroylquinate 3′-monooxygenase [1.14.13.36], shikimate *O*-hydroxycinnamoyltransferase [2.3.1.133], 4-coumarate-CoA ligase [6.2.1.12], ferulate-5-hydroxylase (F5H), coniferyl-aldehyde dehydrogenase [1.2.1.68], beta-glucosidase [3.2.1.21], caffeoylshikimate esterase [3.1.1.−], and coniferyl-alcohol glucosyltransferase [2.4.1.111] (colour figure online)

independent 4-week-old callus samples were harvested to study their transcriptome profiles. We applied QuantSeq 3′ mRNA sequencing approach using Illumina sequencing technology. The approach generated one fragment per transcript at the 3′ end with high strand specificity (>99.9%), allowing more precise and unambiguous gene expression quantification at much lower read number requirement compared to conventional mRNA-seq (Moll et al. 2014). Since we targeted the 3′ untranslated region containing only a few splice junctions, the analysis workflow was straight-forward by mapping to the modified papaya reference sequences with extended ends. We achieved 47.0% of mapping reads which is lower than that of obtained in technical note (Moll et al. 2014) perhaps due to limitations in the draft papaya genome annotation (Ming et al. 2008) compared to other model organisms.

From gene expression analysis, transcripts of majority genes (75.4%) can be detected, but 34.4% of genes were

Table 4 List of selected genes which were highly expressed in papaya embryogenic callus

Gene ID	CPM	Description
Cell proliferation		
CP00003G02560	1371.9	RNA helicase RhlB
CP00055G00740	1293.4	Histone H3
CP00097G00540	1032.0	Translation initiation factor SUI1
CP00034G02040	1008.7	Elongation factor 1-beta 2
CPCG00010	910.8	30S ribosomal protein S12
CP00019G02040	894.9	Acidic ribosomal protein P0
CP00037G01810	866.3	Cis–trans isomerase (protein folding)
CP00033G00130	754.2	60S ribosomal protein L27a
CP00042G00900	749.0	Actin, cytoplasmic
CP00013G01230	736.6	Elongation factor 1-delta 2
CP00078G00310	723.4	Profilin-1
CP00001G03500	715.6	40S ribosomal protein S15a
Cell signalling and regulation		
CP00033G01600	2576.1	Calmodulin-5/6/7/8
CP00179G00160	1293.7	14–3–3-like protein
CP00373G00030	996.7	Auxin-repressed 12.5 kDa protein
CP00003G00420	856.6	Conjugating enzyme E2
CP00097G00560	831.6	Polyubiquitin-C
CP00005G00830	730.2	Bifunctional enolase 2/transcriptional activator
CP00021G01680	728.8	Polyadenylate-binding protein, cytoplasmic and nuclear
CP02820G00010	585.6	F-box protein
CP00481G00010	530.8	Ethylene responsive transcription factor RAP2-12
Cellular homeostasis		
CP00006G00940	369,425.1	Probable cadmium/zinc-transporting ATPase HMA1, chloroplastic
CP00010G02210	22,864.0	2S seed storage protein 5
CP00082G00010	2218.0	Copper transporter 5
CP02639G00010	1755.5	Non-specific lipid-transfer protein 2
CP04842G00010	1226.9	Annexin-like protein RJ4
CP00132G00700	1208.5	Germin-like protein subfamily 1 member 7
CP00092G00940	608.5	Aquaporin TIP1
Somatic embryogenesis		
CP00009G02460	1916.3	LEA5/indole-3-acetic acid-induced protein ARG2
CP00009G02370	715.4	Protein LEA14
CP00106G00730	227.7	BAK1/SERK2
CP00016G01970	126.0	LEA5
CP00066G01110	117.5	BAK1/SERK1
Stress response		
CP00260G00030	5483.2	DnaK/HSP72
CP00033G01180	4829.9	Protein DnaJ
CP00012G01110	1829.6	Heat shock protein HSP90
CP00034G01930	1654.5	Cysteine proteinase inhibitor
CP00026G02260	858.7	Dehydrin Xero 1
CP00190G00040	817.1	Thaumatin-like protein 1
CP00077G01140	800.7	Probable glutathione *S*-transferase MSR-1
CP00046G00200	759.2	Blue copper protein (absence of light)
CP30593G00010	554.7	Ascorbate peroxidase 1, cytosolic
Primary metabolism		
CP00157G00450	1379.9	Adenosylhomocysteinase
CP00146G00370	855.8	*S*-adenosylmethionine synthase

Table 4 continued

Gene ID	CPM	Description
CP00004G00810	570.8	Glyceraldehyde 3-phosphate dehydrogenase
CP00002G04090	509.4	Asparagine synthetase
Secondary metabolism		
CP00152G00140	1302.5	Lignin-forming anionic peroxidase
CP00003G01650	1016.5	Caffeic acid 3-*O*-methyltransferase
CP00107G00140	1007.7	Amorpha-4,11-diene 12-monooxygenase
CP00092G01180	986.4	Phenylalanine ammonia-lyase
CP00002G02720	511.0	Trans-cinnamate 4-monooxygenase
CP00106G00860	484.0	Cinnamyl alcohol dehydrogenase
CP00010G00740	473.3	Cinnamoyl-CoA reductase

found only at low level, and only 47 (0.1%) genes were very highly expressed (CPM > 1000) (Table 2). This reflects the high sensitivity of 3′ mRNA sequencing approach in detecting rare transcripts compared with the conventional RNA-seq approaches, which produce multiple fragments from a single transcript (Moll et al. 2014). Therefore, it is estimated that only around 41% of annotated genes (CPM > 5) were functionally active in 4-week-old papaya embryogenic calli. More callus samples at different stages of development are needed to observe the changes in gene expression over time.

There was no clear difference in the proportion of GO terms of expressed genes (with low and high expression) compared with non-expressed genes, apart from certain GO categories uniquely present for expressed genes (Fig. 2). However, the differences were obvious when considering KEGG pathway annotation based on the proportion of expressed and non-expressed genes to total annotated genes (Fig. 4). Genes involved in pathways which are important for cell growth and proliferation, such as metabolic pathways, cell cycle, plant hormone signal transduction and glycolysis were mostly expressed. The importance of riboflavin and thiamine metabolism in papaya embryogenic callus was implicated by the expression of all genes involved in the pathways (Fig. 4). This supports that vitamin supplements can help in embryogenic callus induction (Asano et al. 1996). Zeatin (cytokinin) is one of plant growth regulators that promotes cell proliferation, which has been reported to increase during cell division of tobacco suspension culture (Redig et al. 1996). However, not all genes involved in zeatin biosynthesis were expressed (Fig. 4).

Interestingly, many of the genes in the biosynthesis of secondary metabolites were found to be highly expressed. The secondary metabolites could act as antioxidants to counteract the oxidative stresses in tissue culture condition. This is consistent with significant GO enrichment of highly expressed genes in response to stimuli related to chemical or abiotic stress (Fig. 3). Furthermore, phenylpropanoid biosynthesis was also significantly enriched (Table 3), in which PAL was found to be highly abundant. PAL catalyses the first step in phenylpropanoid biosynthesis pathway which produces important secondary metabolites, such as flavonoids and lignin. The number of genes expressed in phenylpropanoid biosynthesis pathway (Fig. 5) and the abundance of PAL in papaya embryogenic callus suggest that secondary metabolites could be abundant even during the early stages of callus formation. Enrichment of phenylpropanoid pathway was also reported in strawberry embryogenic callus (Gao et al. 2015).

On the other hand, the expression of genes in carotenoid biosynthesis, including geranylgeranyl diphosphate reductase (CP00423G00040 and CP00193G00080), phytoene desaturase (CP00157G00030), beta-carotene 3-hydroxylase (CP00107G01070), zeta-carotene isomerase (CP00008G02970), and carotene epsilon-monooxygenase (CP00005G01260) (Supplementary Table S1) is consistent with the characteristic yellowish colour of healthy papaya embryogenic callus (Fig. 1a). It is well-known that papaya fruits are rich in carotenoids, our study revealed that those genes involved were expressed even in 4-week-old embryogenic calli.

Many hormone-related genes were expressed in papaya embryogenic callus (Table 4). The highly expressed indole-3-acetic acid-induced protein is consistent with previous finding in Arabidopsis callus (Gliwicka et al. 2013). In addition, the expression of auxin-repressed 12.5 kDa protein and brassinosteroid insensitive 1-asociated receptor kinase 1 (BAK1) in high abundance supports that somatic embryogenesis involves the coordination of different phytohormones and auxin is important in callus induction. Furthermore, ethylene-responsive transcription factor (ERF) expression suggests the involvement of ethylene in papaya somatic embryogenesis as reported previously in oil palm (Piyatrakul et al. 2012). ERF was also detected in friable embryogenic callus of cassava (Ma et al. 2014).

Somatic embryogenesis receptor-like kinase (SERK) genes were shown to be markers of somatic embryogenesis, such as carrot (Schmidt et al. 1997) and maize (Baudino et al.

2001). In papaya, SERK1 and BAK1/SERK2 genes were expressed at high abundance level. Embryo defective gene (EMB, CP00667G00010) reported in longan embryogenic callus was detected only at low level in papaya (Supplementary Table S1). Recently, enolase (CP00005G00830), esterase (CP00025G01800) and ADH3 (CP00049G00680) proteins were proposed to be important for the maturation of papaya embryogenic callus (de Moura et al. 2014) which were also reported in maize (Everett et al. 1985; Fransz et al. 1989) and Arabidopsis (Andriotis et al. 2010). These genes were found to be abundant in our study (Table 4, Supplementary Table S1).

We also identified transcription factor (TF) families that may play important roles in embryogenic callus of papaya. For example, NAM/ATAF1/CUC2 (NAC) (CP00064G01120, CP00594G00010 and CP00099G00620), WRKY (CP00002G03210, CP00768G00010 and CP00055G01020), MYB (CP00006G00950 and CP00005G03080), WUSCHEL (CP00065G01530), Agamous-like MADS-box protein (CP00043G00690) and Basic/HELIX-LOOP-HELIX (bHLH) (CP00097G00750) were previously found in callus of Arabidopsis, strawberry and maize (Gao et al. 2015; Gliwicka et al. 2013; Salvo et al. 2014; Wickramasuriya and Dunwell 2015).

In vitro formation of callus involves many processes at molecular and cellular levels which can be revealed by GO enrichment analysis. Stress-related proteins, heat shock protein (HSP) and WRKY TF (Pandey and Somssich 2009) which are known for their involvement in stress responses were found to be abundant. In papaya callus, GO terms related to stress response were significantly enriched indicating stress-related genes were highly expressed and consistent with previous studies on embryogenic callus (de Moura et al. 2014; Gliwicka et al. 2013; Wickramasuriya and Dunwell 2015; Xu et al. 2012). Furthermore, glutathione *S*-transferase (GST) gene family (CP00077G01140, CP00009G01950, CP00001G 03680 and CP00056G00490) which is involved in plant defence and oxidative stress was also observed at high expression in papaya callus. This is consistent with previous proteomic study (de Moura et al. 2014). The diverse roles of GST in plant development has been reviewed (Fehér et al. 2003; Galland et al. 2007), including dedifferentiation and reactivation of cell division. GST was reported in the somatic embryogenesis of chicory (Galland et al. 2007), grape (Marsoni et al. 2008) and soybean (Thibaud-Nissen et al. 2003). This supports that somatic embryogenesis is closely linked to stresses.

Conclusion

Most of the earlier studies on papaya callus focused on the optimisation of culture media composition and describing embryogenesis based on morphological changes. This is the first report describing genome-wide expression of genes in papaya embryogenic callus, providing a snapshot of the transcripts present in 4-week-old embryogenic calli. In summary, consistent with earlier studies in different plants, majority of genes were expressed in embryogenic callus especially those involved in cell growth and proliferation, such as metabolic pathways, cell cycle, plant hormone signal transduction and glycolysis. The expression of many stress response genes in papaya callus supports that stress could be an inducer of somatic embryogenesis. Secondary metabolite biosynthesis appeared to be active in embryogenic callus of papaya based on the proportion of expressed genes. This further suggests that papaya embryogenic callus can be useful in biopharming, as a cell factory to produce valuable phytochemicals, medicinal and bioactive compounds in vitro. Information on the molecular genetics of complex biological events in papaya embryogenic callus provides insights for future study in somatic embryogenesis of other fruit crops.

Acknowledgements We thank Kok-Keong Loke for helping with the RNA-seq analysis by generating the modified papaya genome reference for read alignment. This research was supported by the Malaysian Ministry of Science, Technology and Innovation (MOSTI) Sciencefund Grant 02-01-02-SF0907 and Universiti Kebangsaan Malaysia Research University Grant (GGPM-2011-053).

Authors Contributions NDJ and HHG conceived and designed the experiments. NDJ performed the experiments. NDJ and HHG analysed the data. NDJ, NMN and HHG wrote the paper.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests. There is no restriction on publication of the data or information described in this manuscript.

Ethical approval This study was conducted according to compliance with ethical standards. This study does not involve the use of any human, animal and endangered or protected plant species as materials.

References

Andriotis VM, Kruger NJ, Pike MJ, Smith AM (2010) Plastidial glycolysis in developing Arabidopsis embryos. New Phytol 185:649–662. doi:10.1111/j.1469-8137.2009.03113.x

Asano Y, Katsumoto H, Inokuma C, Kaneko S, Ito Y, Fujiie A (1996) Cytokinin and thiamine requirements and stimulative effects of riboflavin and α-ketoglutaric acid on embryogenic callus induction from the seeds of Zoysia japonica steud. J Plant Physiol 149:413–417. doi:10.1016/S0176-1617(96)80142-8

Ascencio-Cabral A, Gutiérrez-Pulido H, Rodríguez-Garay B, Gutiérrez-Mora A (2008) Plant regeneration of *Carica papaya* L. through somatic embryogenesis in response to light quality, gelling agent and phloridzin. Sci Hortic 118:155–160

Baudino S et al (2001) Molecular characterisation of two novel maize LRR receptor-like kinases, which belong to the SERK gene family. Planta 213:1–10

Bhattacharya J, Khuspe S (2001) In vitro and in vivo germination of papaya (*Carica papaya* L.) seeds. Sci Hortic 91:39–49

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi:10.1093/bioinformatics/btu170

Chen M, Chen C (1992) Plant regeneration from Carica protoplasts. Plant Cell Rep 11:404–407

Chen M, Wang P, Maeda E (1987) Somatic embryogenesis and plant regeneration in *Carica papaya* L. tissue culture derived from root explants. Plant Cell Rep 6:348–351

Chen C-J, Liu Q, Zhang Y-C, Qu L-H, Chen Y-Q, Gautheret D (2011) Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic callus. RNA Biol 8:538–547

da Silva JAT, Rashid Z, Nhut DT, Sivakumar D, Gera A, Souza MT Jr, Tennant PF (2007) Papaya (*Carica papaya* L.) biology and biotechnology. Tree For Sci Biotechnol 1:47–73

de Moura Vale E et al (2014) Comparative proteomic analysis of somatic embryo maturation in *Carica papaya* L. Proteom Sci 12:1

Elgadir MA, Salama M, Adam A (2014) *Carica papaya* as a source of natural medicine and its utilization on selected pharmacetical applications. Int J Pharm Pharm Sci 6:868–871

Everett N, Wach M, Ashworth D (1985) Biochemical markers of embryogenesis in tissue cultures of the maize inbred B73. Plant Sci 41:133–140

Fabi JP, Mendes LRBC, Lajolo FM, do Nascimento JRO (2010) Transcript profiling of papaya fruit reveals differentially expressed genes associated with fruit ripening. Plant Sci 179:225–233

Fabi JP, Broetto SG, da Silva SLGL, Zhong S, Lajolo FM, do Nascimento JRO (2014) Analysis of papaya cell wall-related genes during fruit ripening indicates a central role of polygalacturonases during pulp softening. PLoS ONE 9:e105685

Fehér A, Pasternak TP, Dudits D (2003) Transition of somatic plant cells to an embryogenic state. Plant Cell Tissue Org Cult 74:201–228. doi:10.1023/a:1024033216561

Fitch MM, Manshardt RM, Gonsalves D, Slightom JL (1993) Transgenic papaya plants from Agrobacterium-mediated transformation of somatic embryos. Plant Cell Rep 12:245–249

Fransz P, De Ruijter N, Schel J (1989) Isozymes as biochemical and cytochemical markers in embryogenic callus cultures of maize (*Zea mays* L.). Plant Cell Rep 8:67–70

Galland R, Blervacq A-S, Blassiau C, Smagghe B, Decottignies J-P, Hilbert J-L (2007) Glutathione-*S*-transferase is detected during somatic embryogenesis in chicory. Plant Signal Behav 2:343–348

Gao L, Zhang J, Hou Y, Yao Y, Ji Q (2015) RNA-seq screening of differentially-expressed genes during somatic embryogenesis in *Fragaria* x *ananassa* Duch. 'Benihopp'. J Hortic Sci Biotechnol 90:671–681

Gliwicka M, Nowak K, Balazadeh S, Mueller-Roeber B, Gaj MD (2013) Extensive modulation of the transcription factor transcriptome during somatic embryogenesis in *Arabidopsis thaliana*. PLoS ONE 8:e69261

Lai Z, Lin Y (2013) Analysis of the global transcriptome of longan (*Dimocarpus longan* Lour.) embryogenic callus using Illumina paired-end sequencing. BMC Genom 14:1

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923

Lin H-C, Morcillo F, Dussert S, Tranchant-Dubreuil C, Tregear JW, Tranbarger TJ (2009) Transcriptome analysis during somatic embryogenesis of the tropical monocot *Elaeis guineensis*: evidence for conserved gene functions in early development. Plant Mol Biol 70:173–192

Litz RE, Conover RA (1981) In vitro polyembryony in *Carica papaya* L. ovules. Z Pflanzenphysiol 104:285–288

Litz R, Conover R (1982) In vitro somatic embryogenesis and plant regeneration from *Carica papaya* L. ovular callus. Plant Sci Lett 26:153–158

Ma Q, Zhou W, Zhang P (2014) Transition from somatic embryo to friable embryogenic callus in cassava: dynamic changes in cellular structure, physiological status, and gene expression profiles. Front Plant Sci 6:824

Maere S, Heymans K, Kuiper M (2005) BiNGO: A cytoscape plugin to assess overrepresentation of geneontology categories in biological networks. Bioinformatics 21:3448–3449. doi:10.1093/bioinformatics/bti551

Marsoni M, Bracale M, Espen L, Prinsi B, Negri AS, Vannini C (2008) Proteomic analysis of somatic embryogenesis in *Vitis vinifera*. Plant Cell Rep 27:347–356. doi:10.1007/s00299-007-0438-0

Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Moll P, Ante M, Seitz A, Reda T (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods. doi:10.1038/nmeth.f.376

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185. doi:10.1093/nar/gkm321

Murashige T, Skoog F (1962) A Revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol Plant 15:473–497. doi:10.1111/j.1399-3054.1962.tb08052.x

Pandey SP, Somssich IE (2009) The role of WRKY transcription factors in plant immunity. Plant Physiol 150:1648–1655

Piyatrakul P et al (2012) Some ethylene biosynthesis and AP2/ERF genes reveal a specific pattern of expression during somatic embryogenesis in *Hevea brasiliensis*. BMC Plant Biol 12:1

Porter BW, Aizawa KS, Zhu YJ, Christopher DA (2008) Differentially expressed and new non-protein-coding genes from a *Carica papaya* root transcriptome survey. Plant Sci 174:38–50. doi:10.1016/j.plantsci.2007.09.013

Redig P, Shaul O, Inzé D, Van Montagu M, Van Onckelen H (1996) Levels of endogenous cytokinins, indole-3-acetic acid and abscisic acid during the cell cycle of synchronized tobacco BY-2 cells. FEBS Lett 391:175–180

Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth 10:71–73. doi:10.1038/nmeth.2251

Salvo SA, Hirsch CN, Buell CR, Kaeppler SM, Kaeppler HF (2014) Whole transcriptome profiling of maize during early somatic embryogenesis reveals altered expression of stress factors and embryogenesis-related genes. PLoS ONE 9:e111407

Schmidt E, Guzzo F, Toonen M, De Vries S (1997) A leucine-rich repeat containing receptor-like kinase marks somatic plant cells competent to form embryos. Development 124:2049–2062

Sharma SK, Millam S, Hedley PE, McNicol J, Bryan GJ (2008) Molecular regulation of somatic embryogenesis in potato: an auxin led perspective. Plant Mol Biol 68:185–201

Sun D-Q, Lu X-H, Liang G-L, Guo Q-G, Mo Y-W, Xie J-H (2011) Production of triploid plants of papaya by endosperm culture. Plant Cell Tissue Org Cult 104:23–29

Thibaud-Nissen F, Shealy RT, Khanna A, Vodkin LO (2003) Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol 132:118–136

Urasaki N et al (2012) Digital transcriptome analysis of putative sex-determination genes in (*Carica papaya*). PLoS ONE 7:e40904. doi:10.1371/journal.pone.0040904

Wickramasuriya AM, Dunwell JM (2015) Global scale transcriptome analysis of Arabidopsis embryogenesis in vitro. BMC Genom 16:1

Xie C et al (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. doi:10.1093/nar/gkr483

Xu K, Liu J, Fan M, Xin W, Hu Y, Xu C (2012) A genome-wide transcriptome profiling reveals the early molecular events during callus initiation in Arabidopsis multiple organs. Genomics 100:116–124

Ye J et al (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. doi:10.1093/nar/gkl031

Yu T-A, Yeh S-D, Cheng Y-H, Yang J-S (2000) Efficient rooting for establishment of papaya plantlets by micropropagation. Plant Cell Tissue Org Cult 61:29–35