Plant Omics Journal POJ

POJ 10(05):219-231 (2017) ISSN:1836-3644
doi: 10.21475/poj.10.05.17.pne721

Mechanistic insights into longan (*Dimocarpous longan Lour.*) transcriptome for physiological characterization for defensive genes and differential gene expression analysis with longan embryogenic callus transcriptome

Manisha Goyal, Jitender Singh*, Pankaj Kumar, Anil Sirohi

Bioinformatics Infrastructure facility, College of Biotechnology, S. V. Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India-250110

***Corresponding author: jeets80@gmail.com**

Abstract

Longan is one of several subtropical and climateric fruit tree of South East Asia having economic and medicinal importance. To understand the mechanism behind altered expression of genetic variants towards various complex diseases, transcriptome analysis is pivotal in present biological scenario. Availability of transcriptomic data of longan facilitates its further functional genomics analysis, differential gene expression analysis and progressive research on the plant varieties of similar family belonging to Indian subcontinent. In the present study, a total number of 90,762,826 mixed paired end sequence reads from 7 different tissues of longan transcriptome (ID: SRR2864836) and 64,876,528 reads for longan embryogenic callus (EC) transcriptome (ID: SRR412534) were downloaded from SRA database, NCBI using SRA Tool Kit. A total number of 9,304 and 10,679 contigs (including scaffolded regions) were assembled from the longan and longan-EC clean reads. Cloud blast search was performed against multiple protein and nucleotide databases at NCBI to explore the mechanistic insights, of the longan transcriptome. A total 9,304 contigs/ putative transcripts were classified into 10 major biological processes, 30 GO functional groups, 08 various cellular components and 139 KEGG pathways. In blast analysis against Nr database, database of *Arabidopsis thaliana, Zea mays and Oryza sativa* 33 contigs were reported for possessing disease resistant proteins. Additionally differential gene expression analysis between longan and longan-EC transcripts showed a total of 34 genes to be highly upregulated and 26 downregulated genes. Thus our study paves way for further research to improvise and utilize its economic and medicinal characteristics of longan as reference in context to Indian plants viz. *Litchi chinensis, Arytera divericata, Schleichera oleosa (Lour.)* etc.

Keywords: Longan, Transcriptome, Defensive genes, Contigs, Gene Ontology, Functional annotation, Differential gene expression etc.
Abbreviations: BLAST_Basic Local Alignment Search Tool; GO_Gene Ontology; KEGG_Kyoto Encyclopaedia for Genes and Genomes; DGE_Differential Gene Expression; EC_Embryogenic callus.

Introduction

Longan (*Dimocarpus longan* Lour.) is an exotic subtropical/Tropical perennial and non climateric crop and widely distributed in South East Asia (Jiang et al., 2002). This evergreen tree belongs to family Sapindaceae (Huang et al., 1995; Tindall et al., 1994). Although indigenously belongs to South East Asia and China but more than 20 countries are now expanding its cultivation worldwide (Yaounde, 2011; Menzel and Waite, 2005). China is the biggest producer of longan in terms of both cultivation area (470000 ha) and total production (610000 tons) (Yaounde, 2011). For more than a century, it has been used in Chinese medicinal practices for health benefits, such as to enhance blood metabolism, increase immunity, to treat insomnia and to improve learning and memory retaining capacity (Yang & He et al., 2008; Park et al., 2010). Besides all of the above mentioned eminent advantage, longan exhibits great anti-aging impact, anti cancer activities and antioxidant property (Yang et al., 2008). Additionally it possesses the attributes of nourishing blood, rich source of vitamin C, reduces stress and fatigue and invigorating effects on heart and spleen (Yang et al., 2011). Considering its exceptional economic as well as medicinal significance, studies have been published on regulation of longan embryo development (Lai and Kin, 2013) and many genes encoding proteins have been cloned. The molecular resources of longan were bounded because its genetic expression profile was unavailable. The availability of pooled transcriptomic data of different tissues of longan leads us to determine and analyse its mechanistic insights comprehensively as well as to identify its defence related potentiality against various diseases, viruses, bacteria, stresses etc. Furthermore differently expressed genes in longan correlating with longan-EC were identified. Moreover, the genomic and proteomic information retrieved from the different expressed genes of longan can serve as a highly useful resource to provide more depth of longan and other important belonging to Sapindaceae for their functional genomic analysis. Herein attempts were focused on perceiving the genomic and proteomic aspects of longan using in-silico approaches viz. a viz. de-novo assembly, global transcriptomic functional annotation and digital gene expression profiling. To assist our global analysis of longan and Differential Gene Expression (DGE) analysis using longan-EC transcriptome, sequencing reads were downloaded from SRA (Sequence Reads Archive) database, NCBI (National Centre for Biotechnology Information) and were developed through next-generation sequencing (NGS)

technology ILLUMINA (Illumina HiSeq 2000) (Zhang et al., 2016). Simultaneously, de-novo assembly of these sequence reads using CLC Genomics Workbench was performed and numbers of contigs were generated. Subsequently Blast2GO Pro software was used to analyze and annotate the data set of assembled putative longan transcripts in order to determine its whole transcriptome functionality. Additionally different biological pathways were also identified using Kyoto Encyclopaedia for Genes and Genomes (KEGG) database for facilitating future medicinal perspectives of longan tissues. The study also depict information about novel disease resistant genes identified using NR db, databases of Arabidopsis thaliana, Oryza sativa and Zea mays which can be further utilized fruitfully for the crops and plants of Indian subcontinent. Furthermore differentially expressed genes in longan were explored in resemblance with longan-EC genes using RNA-seq analysis module incorporated in CLC Genomics workbench. Thus, our Study delineates the identification of various measures to increase the yield of crops and plants with enhanced economic factor. Consequently, an accelerated effort to signify the genomic and proteomic information as well as differently expressed genes of longan for productive utilization of its various nutritional characteristics and disease resistance capacity is the need of hour.

Results

A total no. of 90,762,826 paired sequences of longan (ID: SRR2864836) possessing average length of 100 bp (base pairs) were taken into consideration. Raw SRA reads were explored to overall relative GC content found in all the reads (Fig. 1). Consequently quality distribution of the raw paired sequences were calculated as average PHRED score (Fig. 2) to measure the quality of nucleobases generated by automated DNA sequencing (Ewing et al., 1998; Ewing and Green, 1998). Moreover, enriched pentamers were highlighted to get information about over-representation analysis in the above said data set (Fig. 3). The over representation of 4 identified pentamers (5mers) CACCA, TGGTG, CCACC, GGTGG was calculated as the ratio of the observed and expected 5mer frequency. The expected frequency was calculated as product of the empirical nucleotide probabilities that make up the 5mer. Following the quality control analysis, adapter trimming was performed (Table 1). Total number of 90,762,746 sequences was trimmed in average length of 99.3. Post trimming results are summarised in table 1. De-novo assembly of processed longan reads was carried out to generate contig sequences with CLC genomics workbench and contigs were formed in two groups including scaffolded regions and excluding scaffolded regions (Table 2). Longan transcripts including scaffolded regions were taken into account for further annotation and analysis. A bulk of 9304 contigs were generated, among them 4,923 were of length 1 kilo base (kb), 3,588 were of 2 kb, 639 were of length 3 kb and rest of all were in ranges between 4- 14 kb. Interestingly out of all the assembled contigs only two possessed the length of 18 kb and 26 kb (Fig. 4). Consequently, contigs were transferred to Blast2GO Pro suite for further mapping and annotation. Similarly 64,876,258 clean reads of longan-EC (ID: SRR412534) were assembled into 10,679 contigs. Assembly of longan-EC reads was performed for DGE analysis between longan and longan-EC.

Sequence alignment via cloud blast for longan contigs

Blast analysis of 9304 longan contigs revealed that, 173 contigs did not show any significance blast hit against NR db and 1497 contig sequences were not expressed significant similarity against SwissProt db. Plant species expressing considerable similarity with contig sequences are shown in fig. 5 against NR db and against SwissProt db in fig. 6. *Nicotiana tabacum*, *Citrus sinensis* and *Citrus clementina* plant species showed maximum hits in SwissProt db with total blast hits similarity 60% (Fig. 5) and *Citrus sinensis*, *Citrus clementina*, *Theobroma cacao* and *Jatropha curcas* showed dominant blast hits against NR db with 92% similarity (Fig. 6).

Protein Domains and Families in longan transcriptome

Interpro Scan result was explored for protein domains, families and repeats. Protein domains distributed among contigs with identifiers includes P-loop containing nucleoside triphosphate hydrolase ((IPR027417) at the highest peak followed by Protein kinase-like domain (IPR011009), Zinc finger (IPR013083), RING/FYVE/PHD-type, Heamodomain-like (IPR009057), Leucine-rich repeat domain (IPR032675), L domain-like, AAA+ ATPase domain (IPR003593), Glycoside hydrolase superfamily (IPR017853) etc. Many other important domains were represented by few contigs like Heat Shock Protein (Hsp90); N terminal domain (IPR020575), Activator of Hsp90 ATPase; N-terminal domain (IPR015310), Nuclear protein 96 (IPR021967), DNA recombination and repair protein RecA, C-terminal domain (IPR023400), DNA replication factor RFC1, C-terminal domain (IPR013725), Trigger factor, Ribosome-binding, bacterial (IPR008881) etc (Fig. 7a).

Protein families identified among transcripts were Cytochrome p450 (IPR001128) followed by UDP glucosyltransferase (IPR002213), Small GTPase superfamily (IPR001806), Protein phosphatase 2C family (IPR015655), Sugar Transporter (IPR005828), Kinesin like protein (IPR027640), FHY3/FAR1 family (IPR031052) and Peptidase S10- serine carboxy peptidase (IPR001563) and so on (Fig.7b). Repeated regions were captured as they are formed to play important role in biological processes. Graphically distributed major identified repeats (Fig. 7c) are WD40 Repeat (IPR001680), Pentatricopepetide repeat (IPR02885), Leucine rich repeat (IPR001611), Parallel beta-helix repeat (IPR006626), HEAT type2 Repeat (IPR021133), Pumilio RNA binding Repeat (IPR001313) and Hexapeptide repeat etc.

Gene Ontology (GO) classification of longan transcripts

In mapping and annotation, GO db search provides us evidence code qualifiers with each retrieved GO term which suggests the quality of functional assignment of each contig. Evidence code distribution with contigs as well as with their blast hits are summarized graphically (Suppl. fig. 1 and fig. 2). The list of evidence code qualifiers and their abbreviations are provided in suppl. table 2.

Gene ontology mapping and annotation are represented in three categories with which putative transcripts are associated viz.; biological process, cellular component and molecular function. Out of the three main categories contigs associated with Biological process (6,101, 65.6%) were found to be dominant followed by molecular function (5942, 63.9%) and cellular component (4909, 52.8%) (Fig. 8).

Table 1. Adapter trimming report of whole dataset of pooled seven tissues of longan transcriptome.

Trim	Input Reads	No Trim	Trimmed	Discarded
Trim on Quality	90,762,826	82,364,286	8,398,540	0
Ambiguity Trim	90,762,826	90,757,669	5,127	30
Filter on Length	90,762,826	90,762,746	0	50

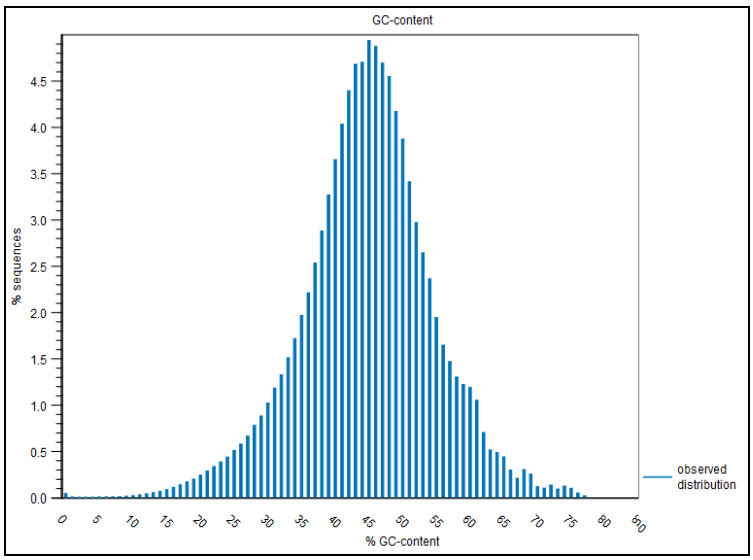

Fig 1. Relative GC-content found in longan transcript sequences (reads).

Table 2. Length measurement of longan contigs. N75 is a length of contigs required to cover 75% of total transcriptome (similarly for N50 and N25).

	Contig length (including scaffold)	Contig length (excluding scaffold)
N75	1,305	1,196
N50	1,671	1,548
N25	2,200	2,056
Minimum	1,000	120
Maximum	25,748	25,748
Average	1,647	1,284
Count	9,304	11,924

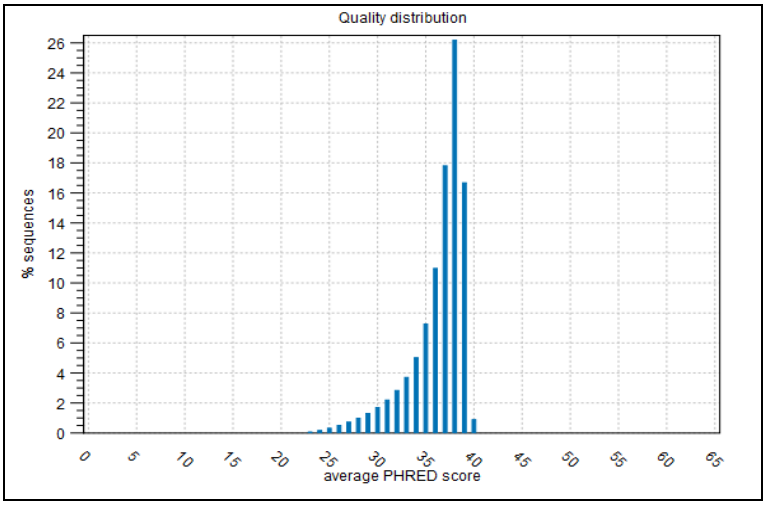

Fig 2. Quality distribution of all the paired reads of longan pooled transcriptome.

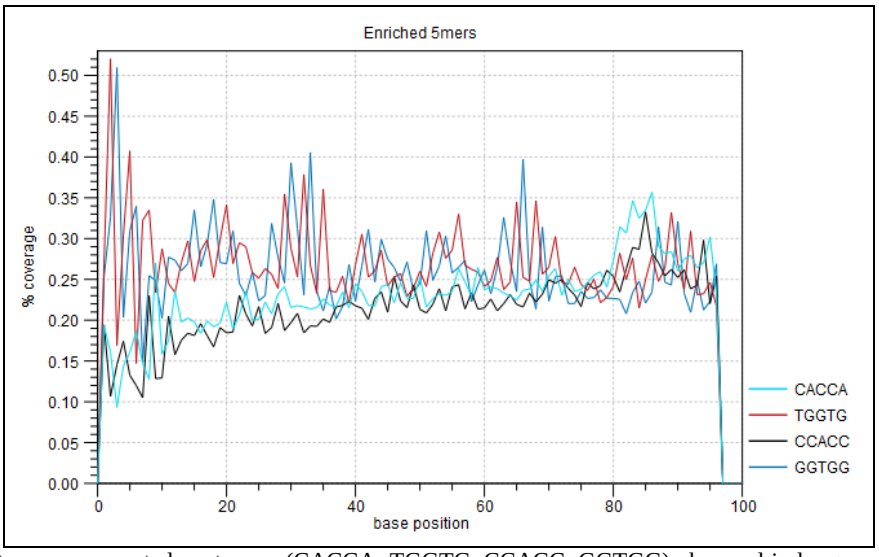

Fig 3. The five most over-represented pentamers (CACCA, TGGTG, CCACC, GGTGG) observed in longan putative transcripts of all the 7 tissues under consideration.

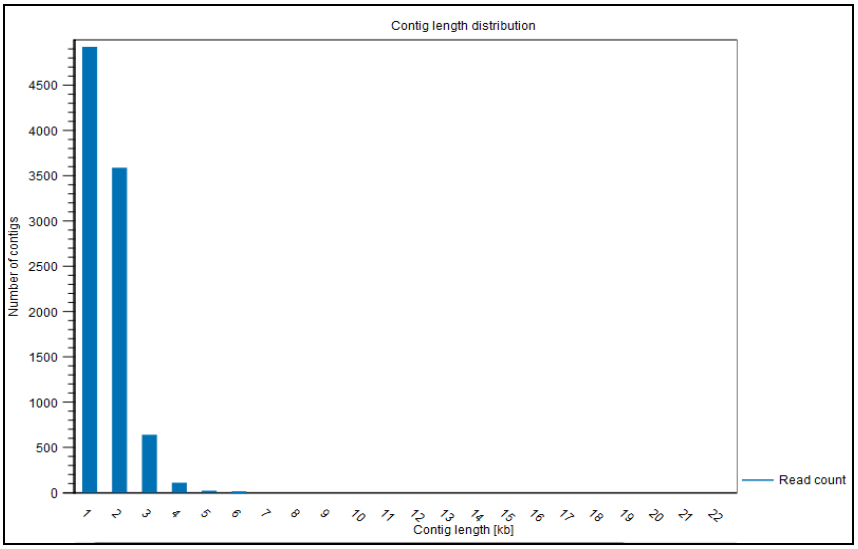

Fig 4. Length distribution of longan contigs produced by clean and high quality reads.

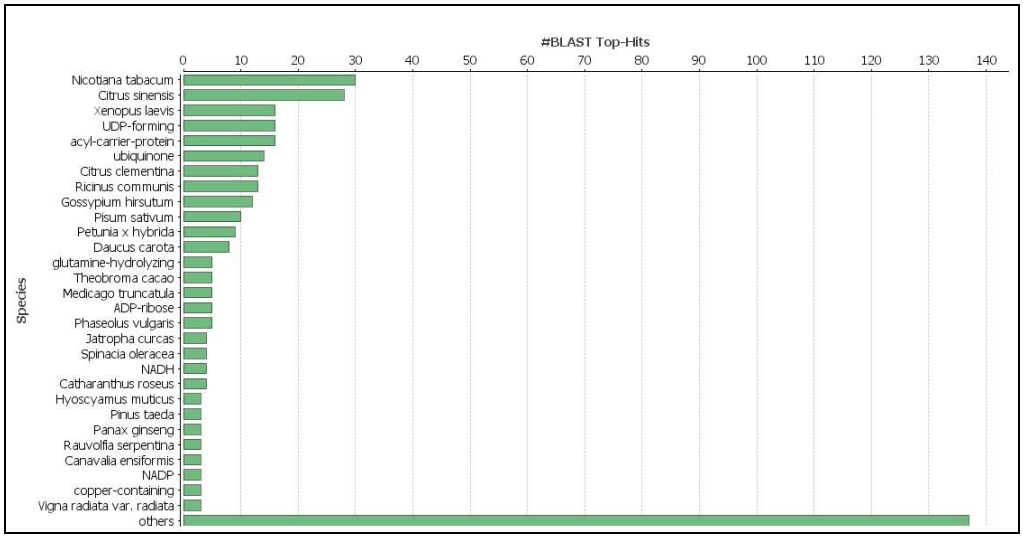

Fig 5. Blast analysis showing similarity between longan contigs and Swissprot database.

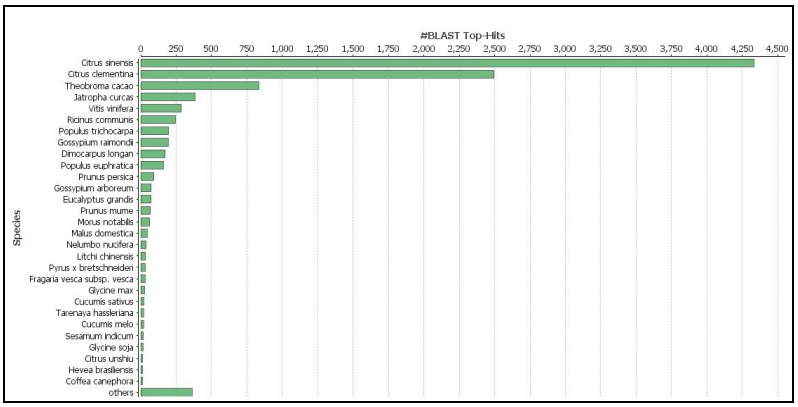

Fig 6. Similarity found between longan pooled transcriptomic contigs and Non-Redundant database.

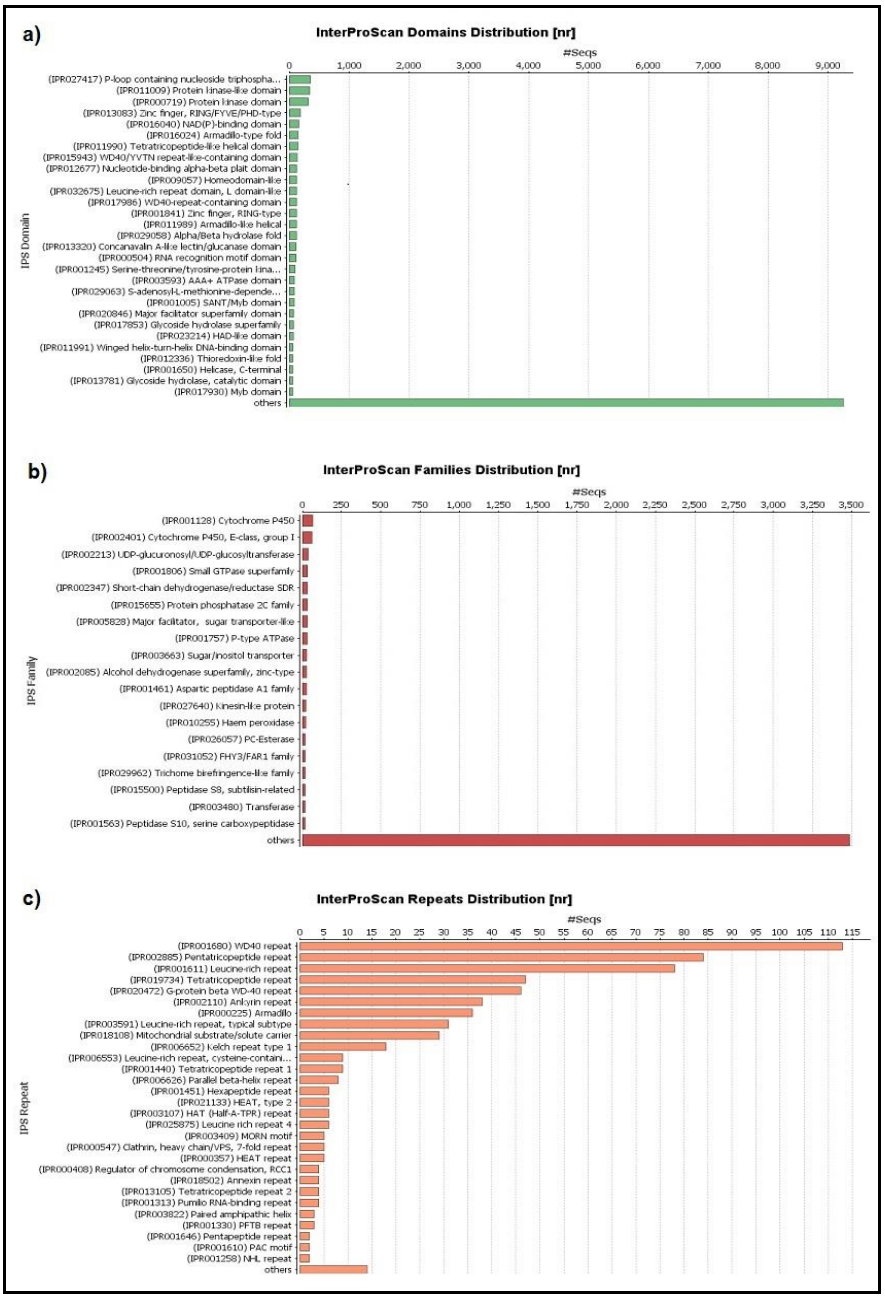

Fig 7. Graphs depicts the search results for longan putative transcripts via interpro scan using Blast2GO Pro suite:

a. Major protein domains and their respective distribution among contigs.
b. Distribution of putative transcripts among the protein families.
c. Repeated regions found in longan transcript sequences.

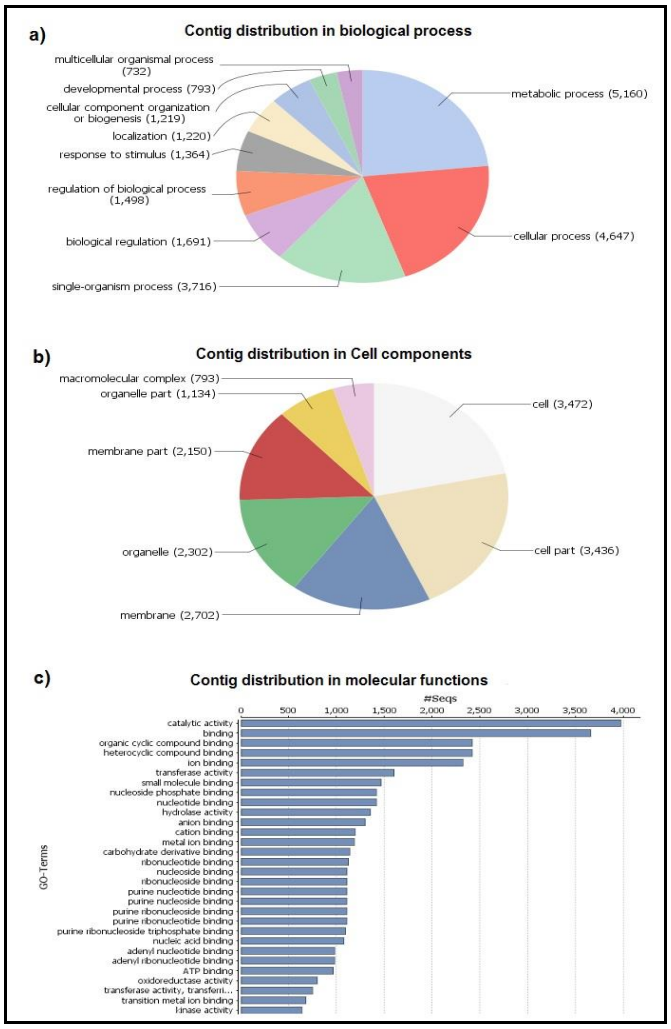

Fig 8. Functional classification of longan contigs based on Gene Ontology (GO) terms are summarized in three main categories:

a. Biological processes
b. Cellular components
c. Molecular functions

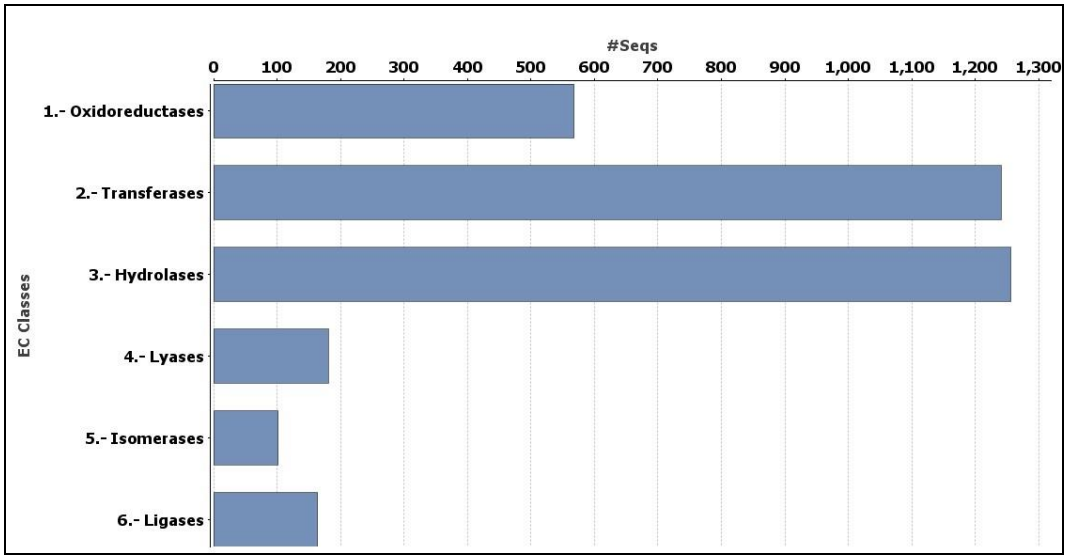

Fig 9. Distribution of Enzyme classes and the numbers of contigs laying in these classes.

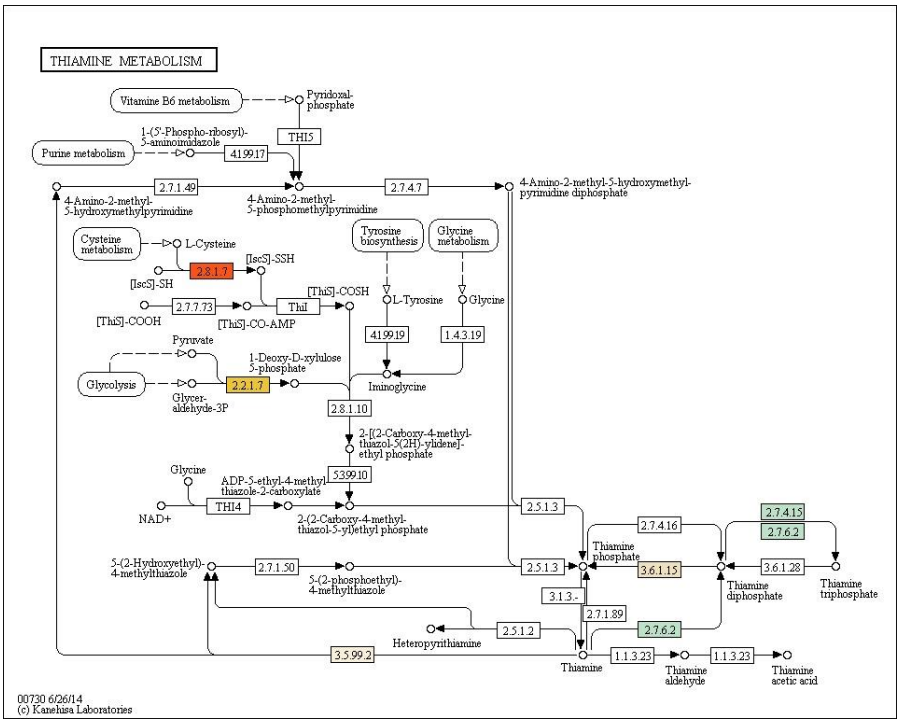

Fig 10. Thiamine Metabolism pathway. 407 contigs were assigned to thiamine metabolism pathway generated by KEGG.

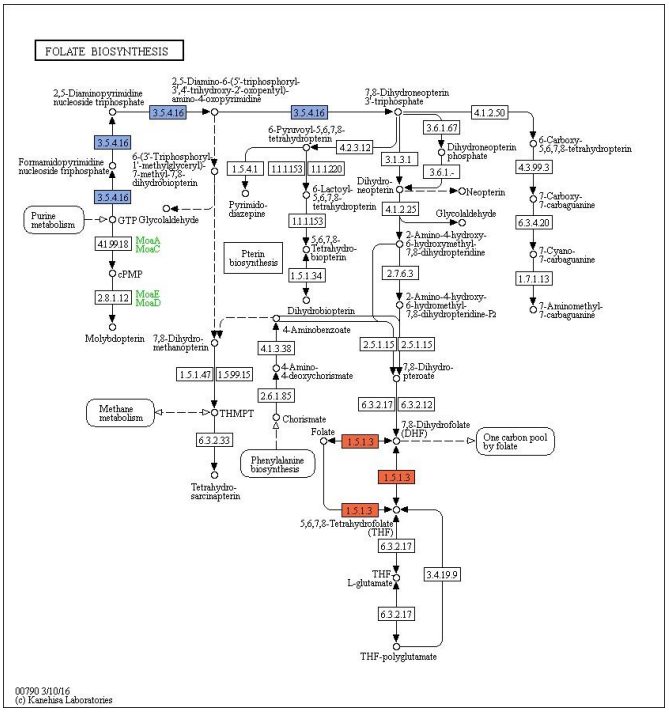

Fig 11. Folate biosynthesis pathway. Three contigs were associated with Folate biosynthesis pathway found in longan crop.

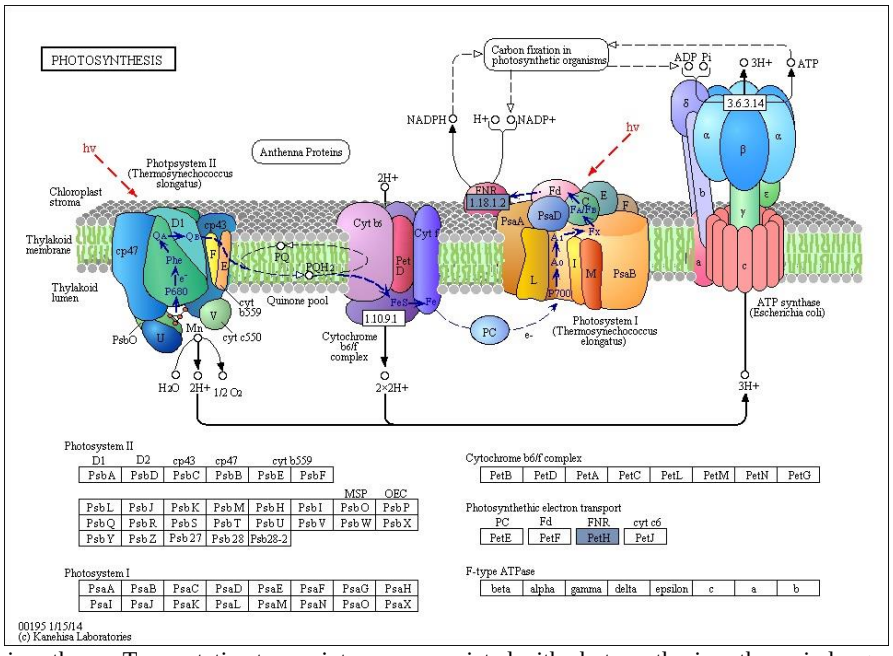

Fig 12. Photosynthesis pathway. Two putative transcripts were associated with photosynthesis pathway in longan tissues.

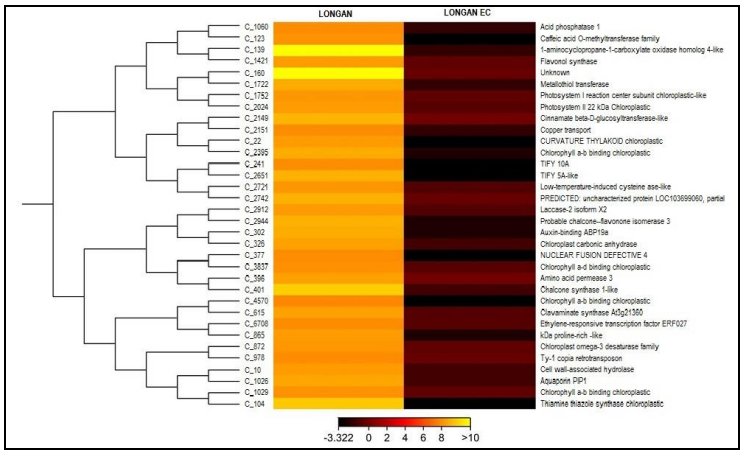

Fig 13. Heat map representation of relative expression based on RPKM values, log2 fold change and FDR p-value correction of differentially expressed genes. All the upregulated (Yellow colour) longan genes showed relatively downregultion in longan-EC (Red colour). Color scale at the bottom represents the expression level. Contig number (starting with C_) and their annotation are given on the left and right side of the heat map respectively.

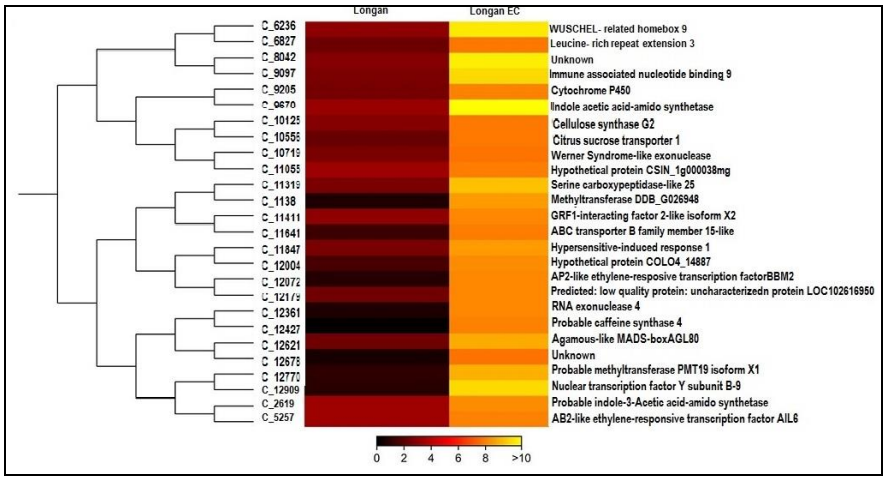

Fig 14. Heat map representation of relative expression based on RPKM values, log2 fold change and FDR p-value correction of differentially expressed genes. All the downregulted (red colour) longan genes showed relatively upregulation in longan-EC (yellow colour). Color scale at the bottom represents the expression level. Contig number (starting with C_) and their annotation are given on the left and right side of the heat map respectively.

The Biological processes were classified into 10 sub-categories. Among them, maximum number of contigs were found to be associated with metabolic process (5160, 84.6%) followed by cellular process (4647; 76.2%), single-organism process (3716; 61%) and biological regulation (1691; 27.7%) and regulation biological process (1498; 24.5%). Moreover a very few putative transcripts were assigned to sub-categories such as developmental process (793) and multicellular organismal process (732).

The transcripts engaged with molecular functions found to have associated with catalytic activity (3976; 66.9%) and binding function (3664; 61.6%).

On the other hand cellular component included 7 sub-categories with 4909 contigs. The most represented sub category was cell (3472 contigs; 70.7%), followed by cell part (3436; 69.9%), membrane (2702; 55%), organelle (2302; 46.9%), membrane part (2150; 43.8%), organelle part (1134; 23.1%) and macromolecular complex (793; 16.15%).

Furthermore, considering enzymes (coded by contigs) code distribution among enzyme classes, highest number of contig sequences was found to have the genes which code for hydrolytic enzymes. This indicates that longan have the property to digest proteins, carbohydrates and esters which can be one of its remarkable health benefits.

Metabolic Pathways assigned to longan transcripts

KEGG database was used to analyse gene products of metabolic processes and related cellular processes. Putative transcripts were screened against the KEGG database, which resulting in, 139 KEGG pathways assigned to 9,304 contigs. Primarily among 139 pathways, the pathway most represented by contigs were related to metabolism (2913; 31.3%) which includes Purine metabolism (494), Thiamine metabolism (407) (Fig. 10), Starch and Sucrose metabolism (175), Amino benzoate degradation (143), Drug metabolism (81), Glycolysis/ Gluconeogenesis (80), Pyrimidine metabolism (66), Glycolipid metabolism (38), Inositol Phosphate metabolism (33), metabolism of Xenobiotics by Cytochrome P450 (33), Retinol metabolism (22), Nitrogen metabolism (18), Biotin metabolism (15), Seleno compound metabolism (13) etc. Collectively metabolism of all the 20 amino acids was assigned to 379 contigs (4.07%). In contrast, only a few putative transcripts were associated with Caffeine metabolism (8), Steroid degradation (7), Taurine and Hypotaurine metabolism (6), C5-Branched dibasic acid metabolism (6), Vitamin B6 metabolism (6), Taulene degradation (5), Ethyl Benzene degradation (2) etc.

Many longan contigs (852; 9.15%) were classified into pathways related to biosynthesis of secondary metabolites such as biosynthesis of Antibiotics (288), Phenylpropanoid biosynthesis (74), Flavonoid Biosynthesis (30), Fatty acid biosynthesis (30), Terpenoid backbone biosynthesis (24), biosynthesis of steroid hormone (21), Streptomycin biosynthesis (16), Ubiquinone and other terpenoid–quinone biosynthesis (14), Flavone and other flavonal biosynthesis (11). Some other biosynthetic pathways were observed to have association of very small number of contigs viz. Monobactam biosynthesis (9), Carotenoid biosynthesis (8), Steriod biosynthesis (8), Stilbenoid diarylheptanoid and Gingerol biosynthesis (3), Anthocynine biosynthesis (3), Folate biosynthesis (3) (Fig. 11).

Furthermore 151 contigs (5.38%) were involved in some other pathways related to signalling systems which includes T-cell receptor signalling pathway (108), Phosphotydyl Inositol signalling pathway (29) and mTOR signalling pathway (14). Other important pathways includes Pentose and Glucurorante inter conversion pathway (74), Pentose phosphate pathway (48), oxidative phosphorylation pathway (48), Carbon fixation in photosynthetic organism (44), carbon fixation pathway in prokaryotes (35), Citerate cycle (33), and Photosynthesis (2) (Fig. 12). On the other hand only few contigs were assigned to photosynthesis (2). Aminoacyl –t RNA biosynthesis pathway is related to genetic information and found to have 37 putative transcripts.

Discovered defence responsive genes in longan transcriptome

A large number of defensive genes were observed in contigs from longan transcriptome. Proteins encoded by these genes were summarized in the suppl. table 1. These proteins were reported to have resistant as well as responsive properties against salt stress, high heat, viruses, bacteria, fungus, and light.

Defence response proteins were found to have different conserved domains such as NBS-LRR, TIR-NBS-LRR, LRR, NBS-ARC, CC-NBS-LRR etc. These conserved domains represent major R-gene classes of Plant resistant genes (Gururani et al., 2012). Additionally some other disease resistant proteins (DRPs) were also exists. Although their specific defence response activity was not found but for future perspective being defence responsive protein, they might play an important role in plant genetics. Among these proteins six proteins (contigs: 442, 3087, 4089, 5352, 7438, 8490) contain NB-ARC domain followed by four other proteins (contigs: 2413, 3133, 5762, 8066) encapsulating NBS-LRR domain and one TIR-NBS-LRR domain based protein (2380). Moreover two disease resistant proteins were also detected with domain CC-NBS-LRR. Among DRPs one RPP13-like protein 1, two Dirigent like proteins, a At3g14460 like isoform X1 and a TMV N like were also observed. Transgenic technology is a primary tool to deploy resistant genes and their ability in different plant species to promote or acquire the resistant against various environmental stresses. Previously similar study has been carried out for Bs2 resistant gene against *B.compestris* which was originally identified in pepper (Tai et al., 1999). Moreover new resistant genes have been created by putting single point mutation which can be introduced in deficient plant (Hammond and Kanyuka 2007). Most promising application of functional genomics could not only provide understanding of plant defence mechanism but also reveals the site of plant-pathogen interactions.

Longan and litchi belongs to the same family i.e. in the Sapindaceae. Evidentially we have observed that global transcriptome profile of both the plants is quite similar in the context of transcript abundance in Gene Ontology categorisation (Li et al., 2013).

Identification of differentially expressed genes in longan and longan-EC transcriptomes

Differential expressison of genes in longan was performed in contrast with genes expressed in longan-EC. With RPKM > 100 (longan) and < 100 (longan-EC) , log2 fold change > 1 and FDR P-value < 0.005, a set of sixty five genes were obtained to be highly upregulated in longan transcriptome as compared to longan-EC transcriptome. Among these, only 34 genes possessing RPKM value > 200 in longan and < 10 in longan-EC, log2 fold change between 2 to 60 were selected for hierarchical feature clustering (Fig. 13). Similarly 519 genes were discovered with low expression profile in longan in contrast with longan-EC using RPKM < 100 (longan) and

>100 (longan-EC), log2 fold change >1 and FDR P-value < 0.005. For feature clustering of downregulated genes only 26 genes out of 519 genes were choosen on the basis of RPKM value < 10 in longan and > 200 in longan-EC with similar fold change and p-value used for upregulated genes categorization (Fig. 14).

All the selected genes were further mapped and annotated using Blast2GO Pro suite against *Arabidopsis thaliana* database. The gene with highest expression profile was unknown having RPKM: 6,355 followed by 1-aminocyclopropane-1-carboxylate oxidase homolog 4-like (RPKM: 6,355). Other genes were flavonol synthase, acid phasphatase1, metallothiol transferase, copper transport, auxin-binding ABP19a, laccase-2 isoform X2, nuclear fusion defective4, clavaminate synthase At3g21360, aquaporin PIP1, ethylene-responsive transcription factor ERF027 and cell wall-associated hydrolase etc. On the other hand poorly expressed genes in longan included probable caffeine synthase 4 (RPKM: 1.03), methyltransferase DDB_G026948 (RPKM: 1.54) followed by RNA exonuclease 4 (RPKM: 1.56). Some other downregulated genes were WUSCHEL-related homebox 9, Immune associated nucleotide binding 9, cytochome p450, Hypersensitive-induced response 1, Werner Syndrome-like exonuclease etc.

Discussion

Transcriptome sequencing is a major eye catching area of research now days. Although many high throughput data have been developed for sequencing and characterization of transcriptomes but still there is huge lack of transcriptomic data for many organisms. Transcriptome sequencing, assembly and annotation allows us to strengthen the information content of transcribed regions at low cost and time. *In silico* based computational tools such as CLC Genomics workbench, Velvet, Trinity etc. have been developed in conjunction with extensive trancriptome assembly and their further application in biotechnological research particularly in genetic engineering and genome editing. In the current study only CLC Genomics Workbench was used for sequence assembly since many previous studies have been demonstrated that CLC genomics workbench is most efficient tools based on different assessment parameters specifically N50 length and average contig length (Misner et al., 2013). Transcritome assembly for various organisms including *Cleome spinosa* and *Cleome gynandra* (Brautigam et al., 2011), *Amycalatopsis orientalis* (Jeong et al., 2013), *Dodonaea viscosa* (Christmas et al., 2015) and 12 Citrus Species (Terol et al., 2016) have already been done using same approach.

In this study, a total of 90,762,826 paired end reads were assembled into 9,304 contigs (≥ 1000 bp) whose N50 length (1,671 bp) was far longer than the length documented in other studies using similar technology but for different plant species such as *Capsicum annum* (1,647; Ashrafi et al., 2012), *Cicer arietinum* (1501; Agarwal et al., 2012) and *Cleome gynandra* and *C. spinosa* (732; Brautigam et al., 2011). Our Results showed that ~92% of the contigs were matched with functional annotation in NR database which is comparable with results of other studies using same species but different approach i.e. 99% matches with NR db in *D. longan* transcriptome (Zhang et al., 2016). Moreover *Theobroma cacao* and *Vitis vinifera* were among the top 5 NR db hits identical with the findings of Zhang et al., (2016). Cytochrome P450 (IPR001128) and UDP-glucosyltransferase (IPR002213) protein domain were found more prevalently in longan transcriptome. Similar protein domains were identified in *Coffea arabica* and *Coffea canephora* (Mondego et al., 2011). We observed that more transcripts are involved in the metabolic process followed by cellular process in longan. These results are consistent with the outcomes of longan embryogenic callus transcriptome (Lai and Lin 2013), *Litchi Chinensis* (Li et al., 2013), *D. longan* (Zhang et al., 2016). Additionally in our exlploration, multicellular organismal process in among the least mapped category similar to the study in *Litchi Chinensis* by Li et.al. (2013). Among the metabolic pathways, our observations illustrate that the metabolism of starch and sucrose is one of the top 3 most prominent metabolic pathways that are active in longan. This is highly invariable with *Siraitia grovenorii* (Tang et al., 2011), *Litchi chinensis* (Li et al., 2013), *Phoenix dactylifera* (Yin et. al. 2012) and *Myrica rubra* (Feng et al., 2012).

Disease Resistance genes (R genes) in plants are mostly encoded with leucine rich repeats (NBS-LRR) proteins. These repeats are known to have immune response against pathogen attack (Belkhadir et al., 2004; McHale et al., 2006). In the present study we have analysed 33 disease resistance genes in longan trancriptome. CC-NBS-LRR and TIR-NBS-LRR protein domains were observed frequently which is quite analogous with R genes found in *C. arabica and C. canephora* (Mondego et al., 2011). All other defensive genes are showing similarity with *Arabidopsis thaliana, Citrus sinensis, Zea mays* and *D. Longan* callus etc. (Suppl. Table 1). Similar results were found in other studies such as in longan-EC (Lai et al., 2013) and rice (Yuli Li et al., 2014).
Callus transcriptomes have been obtained for number of plant species including *A. Thaliana* (Che et al., 2006), *D. longan* (Lia and Lin et al., 2013) and *Picea balfouriana* (Li and Wang et al., 2014). In rice Yuli Li et al., (2014) identified expressed genes based on the comparison of 2 samples representing callus and leaf which is similar to our study conducted for pooled transcriptome of longan seven different tissues (young and adult leaves, young and mature fruit, floral buds and vegetative buds, and roots) and trancriptome of longan-EC. During expression analysis FDR p- value < 0.005 (Wei et al., 2015), RPKM value 10- 200 and >200 (Sweetman et al., 2012) and log2 fold change >2 (Li and Wang et al., 2014; Wei et al., 2015) were used. No such studies have been conducted for DGE analysis between longan and longan-EC till date. In the present study many genes which were found to be highly detectable in longan were exhibit least expression in EC and vice versa suggested that some genes played role in development of all the seven tissues but during callus embryogenesis they are inactive.

Materials and Methods

Plant Material and SRA data downloading

The dataset of transcriptomic reads of different types of tissues (young and adult leaves, young and mature fruit, floral buds and vegetative buds, and roots) of longan (ID: SRR2864836) and longan-EC (ID: SRR412534) were downloaded using SRA tool kit from SRA database (Sequence Reads Archive), NCBI (www.ncbi.nlm.nih.gov/). These reads were developed using total RNA extracted from longan-EC (Lai et al., 2013) and seven different types of tissues of longan (Zhang et al., 2016) plant through next-generation sequencing (NGS) technology ILLUMINA (Illumina HiSeq 2000: run: 45.4M spots, 9.1G bases, 5.7 GB downloads). Sequence data was converted into FASTQ format by executing command line version of SRA tool kit

for storing the output of high-throughput sequencing instruments Illumina Genome Analyzer (Cock et al., 2009).

Data filtering and de-novo assembly of longan reads

Downloaded paired SRA reads in FASTQ format were analysed to retrieve information about read length distribution, GC content, nucleotide base ambiguity, sequence quality, sequence duplication levels etc using CLC genomics workbench (www.qiagenbioinformatics.com /products/clc-genomics-workbench/). The raw reads containing adapters, unknown or low quality bases, and contaminants were trimmed using adapter trimming tool of CLC Genomics Workbench. Keeping statistical parameters as default, trimmed reads were used as input file for further *de-novo* assembly using CLC Genomics workbench and reads were assembled to form contiguous consensus sequences (contigs) from collections of overlapping reads (Yandell and Ence, 2012).

Functional Annotation of longan transcriptome

Functional annotation of assembled longan contiguous consensus sequences (contigs) was carried out using Blast2GO Pro software (Conesa et al., 2005) keeping statistical parameters as default.

Sequence alignment via cloud blast

Contigs were used as queries in Blastx algoritham (Altschul et al., 1997) using Blast2GO Pro software against the non-redundant (NR) and Swissprot database at NCBI. Resulting blast hit with an e-value of ≤1.0E-3 were considered as a significant match for further functional annotation of contigs.

Domain search via InterPro scan

Protein signature sequences recognition was carried through InterPro domain search (Mulder et al., 2003) directly on the FASTA input file of the contigs and the Gene Ontology terms were assigned to the identified domains. Interpro scan was queried against nine databases viz. a viz. BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart HMMTigr, Profile Scan, ScanRegExp and SuperFamily (Gotz et al., 2008).

Gene ontology mapping and annotation

Functional information for each contig was retrieved from Gene Ontology (GO) database encapsulating millions of functionally annotated gene products for several different species. Moreover, GO database contain an evidence code qualifier which provides information related to the quality of this functional assignment. Blast2GO Pro annotation was performed with default parameters after Gene Ontology mapping which enumerated GO annotation score for each candidate GO term.

Biological pathways classification

Large-scale sequencing datasets generated by high-throughput experimental technologies were integrated and interpretated for molecular interactions and biological functions by Kyoto Encyclopedia of Genes and Genomes (KEGG) server (Kanehisa et al., 2016; Kanehisa et al., 2000). Reference standard pathways stored in KEGG database were used to annotate longan transcripts and role of all the contigs in various metabolic pathways were elucidated.

Study of defensive genes in longan transcriptome

To explore the information about defensive genes/ disease resistance genes for each transcript against various environmental stresses, fungal, bacterial and other infections on longan, annotation was carried out. To validate the results, all the contigs were queried against the databases of *Arabidopsis thaliana*, *Zea mays* and *Oryza sativa* via Blast2GO Pro mapping and annotation suite.

Longan digital gene expression profiling in contrast with longan-EC

Digital Profiling for genes that were expressed differently in longan (ID: SRR2864836) as compared to longan-EC (ID: SRR412534) was done using CLC genomics workbench. Contigs for both longan and longan-EC were mapped to reference transcripts which were assembled by high quality reads of longan and longan-EC collectively through RNA-seq tool of CLC workbench. Distribution of paired end distances among the reads was calculated for longan (114 to 277 bp) and longan-EC (145 to 243 bp). This RNA–seq file was used as input to set up an experiment to examine differential gene expression between both the samples using Microarray and Small RNA Analysis module of CLC genomics workbench. While generating expression data RPKM (reads per kilo base of exon per million mapped reads) was used to determine transcript abundance (Mortazavi et al., 2008). Consequently read count based statistical analysis was performed using kal's Z test. All the differentially expressed genes were filtered by restricting RPKM value > 100, FDR p-value correction < 0.005 and log2 fold change > 2. Filtered differentially expressed genes from longan (ID: SRR2864836) and longan-EC (ID: SRR412534) exhibiting high expression as well as low expression were compared and clustered into a tree.

Conclusion

This study explored the large scale transcriptome dataset of longan for functional annotation, classification and metabolic pathways associated with the dataset of putative transcripts of longan provide its detailed molecular and functional insights. Moreover, 33 contigs related to disease resistance property were also identified in longan. Furthermore 34 upregulated and 26 downregulated genes were discovered in longan transcripts in contrast with longan-EC transcriptome. Thus a further research can be facilitated by plant defence responsive genes against various environmental stresses and various other functionally active genes found in longan transcriptome. Keeping in view that longan and litchi belongs to the same family this study can be a new platform for future research in litchi and other similar plants through different biotechnological approaches for the betterment of plant sustainability and yield. Therefore new insights into molecular processes of longan transcriptome would be of high value in context of enhancement for its economic factor as well as to make use of its medicinal importance for other plants and further to intensify their stress resistance calibre.

Acknowledgement

We sincerely acknowledge the Vice Chancellor, Sardar Vallabhbhai Patel University of Agriculture and Technology,

Meerut- 250110, Uttar Pradesh; and Bioinformatics infrastructure facility (DBT funded), India, for providing financial support and the facilities to carry out this research work.

References

Jiang YM, Zhang ZQ, Joyce DC, Ketsa S (2002) Postharvest biology and handling of longan fruit (*Dimocarpus longan* lour.). Postharvest Biol Technol. 26: 241–252.

Huang HB (1995) Advances in fruit physiology of the arillate fruits of litchi and longan. Annu Rev for Hortic Sci. 1: 107–120.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nuc Aci Res. 38 (6): 1767–1771

Tindall HD (1994) Sapindaceous fruits: botany and horticulture. Hortic Rev. 16: 143–196.

Yaounde C (2011). Tropical fruits compendium. Committee on commodity problems, intergovernmental group on bananas and tropical fruits. Food and agriculture organization. Fifth Session.

Menzel CM and Waite GK (2005) Litchi and longan: botany, production and uses. Trowbridge: Cromwell Press.

Yang C, He N, Ling X, Ye M, Zhang C, Shao W, Yao C, Wang Z, Li Q (2008) The isolation and characterization of polysaccharides from longan pulp. Sep Purif Technol. 63: 226–230.

Park SJ, Park DH, Kim DH, Lee S, Yoon BH, Jung WY, Lee KT, Cheong JH, Ryu JH (2010) The memory-enhancing effects of Euphoria longan fruit extract in mice. J Ethnopharmacol. 128(1): 160-165.

Yang B, Zhao MM, Shi J, Yang N, Jiang YM (2008) Effect of ultrasonic treatment on the recovery and DPPH radical scavenging activity of polysaccharides from longan fruit pericarp. Food Chem. 106: 685–690.

Yang B, Jiang Y, Shi J, Chen F, Ashraf M (2011) Extraction and pharmacological properties of bioactive compounds from longan (*Dimocarpus longan* lour.) fruit — a review. Food Res Int. 44 (7): 1837–1842.

Lai Z and LinY (2013) Analysis of the global transcriptome of longan (*Dimocarpus longan* lour.) embryogenic callus using Illumina paired-end sequencing. BMC Genomics. 14:561.

www.ncbi.nlm.nih.gov/

www.qiagenbioinformatics.com/products/clc-genomics-workbench/

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z *et al.,* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nuc Aci Res. 25: 3389–3402.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO Pro: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674-3676.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nuc Aci Res. 44: 457-462.

Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nuc Aci Res. 28: 27-30.

Li C, Wang Y, Huang X, Li J, Wang H, Li J (2013) *De novo* assembly and characterization of fruit transcriptome in *Litchi chinensis* Sonn and analysis of differentially regulated genes in fruit in response to shading. BMC genomics. 14: 552.

Gotz S, Gomez JMG, Terol J, Williams TD, Nagraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughtput functional annotation and data minning with the Blast2GO Pro suite. Nuc Aci Res. 36(10): 3420-3435.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D *et al.,* (2003) The Interpro database, 2003 brings increased coverage and new features. Nuc Aci Res. 31: 315–318.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I accuracy assessment. Genome Res. 8(3): 175–185.

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. error probabilities. Genome Res. 8(3): 186–194.

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Gen. 13: 329-342.

Bittner-Eddy PD, Crute IR, Holub EB, Beynon JL (2000) RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in peronospora parasitica. Plant J. 21: 177-188.

Murray SL, Ingle RA, Petersen LN, Denby KJ (2007) Basal resistance against pseudomonas syringae in Arabidopsis involves WRKY53 and a protein with homology to a nematode resistance protein. Mol Plant Microbe Interact. 20: 1431-1438.

Gou X, He K, Yang H, Yuan T, Lin H, Clouse SD, Li J (2010) Genome-wide cloning and sequence analysis of leucine-rich repeat receptor-like protein kinase genes in *Arabidopsis thaliana*. BMC Genomics. 11:19-19.

Sung DY, Vierling E, Guy CL (2001) Comprehensive expression profile analysis of the Arabidopsis Hsp70 gene family. Plant Physiol. 126: 789-800.

Xu Q, Chen LL, Ruan X *et al.,* (2013) The draft genome of sweet orange (Citrus sinensis). Nat Genet. 45(1): 59-66.

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell. 15: 809-834.

Shirasu K, Lahaye T, Tan MW, Zhou F, Azevedo C, Schulze-Lefert P (1999) A novel class of eukaryotic zinc-binding proteins is required for disease resistance signaling in barley and development in C. elegans. Cell. 99: 355-366.

Ye W, Lin Y, Lai Z (2014) Cloning and analysis NBS resistance genes from embryogenic callus in *Dimocarpus longan* Lour. EMBL/GenBank/DDBJ databases.

Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Mol Biol. 60(1): 69-85.

Van der Biezen EA and Jones JD (1998) The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol. 8(7): 226-227.

Marzec M, Eletto D, Argon Y (2012) GRP94: An HSP90-like protein specialized for protein folding and quality control in the endoplasmic reticulum. Biochim Biophys Acta. 1823(3): 774-787.

Lin X, Kaul S, Rounsley SD, Shea TP, Benito MI, Town CD, Fujii CY, Mason TM, Bowman CL, Barnstead ME, Feldblyum TV, Buell CR, Ketchum KA, Lee JJ, Ronning CM, Koo HL, Moffat KS, Cronin LA, Venter JC (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature. 402: 761-768.

Glavinas H, Krajcsi P, Cserepes J, Sarkadi B (2004) The role of ABC transporters in drug resistance, metabolism and toxicity. Curr Drug Deliv. 1(1): 27-42.

Liao Y, Zou HF, Wei W, Hao YJ, Tian AG, Huang J, Liu YF, Zhang JS, Chen SY (2008) Soybean GmbZIP44, GmbZIP62 and GmbZIP78 genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic Arabidopsis. Planta. 228(2): 225-240.

Zhou L, Cheung MY, Li MW, Fu Y, Sun Z, Sun SM, Lam HM (2010) Rice hypersensitive induced reaction protein 1 (OsHIR1) associates with plasma membrane and triggers hypersensitive cell death. BMC Plant Biol. 10: 290.

Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, Kawashima K, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakayama S, Nakazaki N, Naruo K, Fransz PF (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature. 408: 823-826.

Goritschnig S, Krasileva KV, Dahlbeck D, Staskawicz BJ (2012) Computational prediction and molecular characterization of an oomycete effector and the cognate Arabidopsis resistance gene. PLoS Genet. 8(2).

Salanoubat M, Lemcke K, Rieger M *et al.,* (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. Nature. 408(6814): 820-822.

Yang Y, Wu X, Xuan H, Gao Z (2016) Functional analysis of plant NB-LRR gene L3 by using E. coli. Biochem Biophys Res Commun. 478(4): 1569-1574.

Chiu CC, Chen LJ, Su PH, Li HM (2013) Evolution of chloroplast J proteins. PLoS One. 8(7).

Sato S, Kaneko T, Kotani H, Nakamura Y, Asamizu E, Miyajima N, Tabata S (1998) Structural analysis of *Arabidopsis thaliana* chromosome 5. IV. sequence features of the regions of 1,456,315 bp covered by nineteen physically assigned P1 and TAC clones. DNA Res. 5: 41-54.

Lopez ZJ, Forczek E, Hoen DR, Juretic N, Bureau TE (2012) A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in *Arabidopsis thaliana.* PLoS-Genet.

Mayer K, Schüller C, Wambutt R *et al.,* (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature. 402(6763): 769- 777.

Kobe B and Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol. 11(6): 725-732.

Badri DV, Chaparro JM, Manter DK, Martinoia E, Vivanco JM (2012) Influence of ATP binding cassette transporters in root exudation of phytoalexins, signals, and in disease resistance. Front Plant Sci. 3:149.

Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey Sk, Park SW (2012) Plant disease resistance genes: current status and future directions. Physiol Mol Plant Path. 78: 51-65.

Tai TH, Dahlbeck D, Clark ET *et al.,* (1999) Expression of the Bs2 pepper-gene confers resistance to bacterial spot disease in tomato. Proc Natl Acad Sci. 96 (141): 53-58.

Hammond KKE and Kanyuka K (2007) Resistance genes (R genes) in plants.

Mondego JMC, Vidal RO , Carazzolle MF, Tokuda EK, Parizzi LP, Costa GGL , Pereira LFP , Andrade A C, Colombo CA, Vieira LGE, Pereira GAG (2011) An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. BMC Plant Biol. 11:30.

Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. Curr Opin Plant Biol. 7(4): 391-399.

McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol. 7(4): 212.

Brautigam A, Mullick T, Schliesky S, Weber APM (2011) Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. J Exp Bot. 62(9): 3093–3102.

Yin YX, Zhang XW, Fang YJ, Pan LL, Sun GY, Xin CQ, Abdullah MMB, Yu XG, Hu SN, Al-Mssallem IS, *et al.,* (2012) High-throughput sequencing-based gene profiling on multi-staged fruit development of date palm (*Phoenix dactylifera*, L.). Plant Mol Biol. 78: 617–626.

Feng C, Chen M, Xu CJ, Bai L, Yin XR, Li X, Allan AC, Ferguson IB, Chen KS (2012) Transcriptomic analysis of chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq. BMC Genomics. 13: 19.

Tang Q, Ma XJ, Mo CM, Wilson IW, Song C, Zhao H, Yang YF, Fu W, Qiu DY (2011) An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics. 12: 343.

Ashrafi H, Hill T, Stoffel K, Kozik A, Yao J, Chin-Wo RS, Deynze AV (2012) *De novo* assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. BMC Genomics. 13:571.

Agarwal G, Jhanwar S, Priya P, Singh VK, Saxena MS, Parida SK, Garg R, Tyagi AK, Jain M (2012) Comparative Analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. PLOS One. 7(12).

Terol J, Tadeo F, Ventimilla D, Talon M (2016) An RNA-Seq-based reference transcriptome for Citrus. Plant Biotechnol J. 14: 938–950.

Christmas MJ, Biffin Ed, Lowe AJ (2015) Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscose*. BMC Genomics. 16: 803.

Zhang HN, Shi SY, Li WC, Shu B, Liu LQ, Xie JH,Wei YZ (2016) Transcriptome analysis of 'Sijihua' longan *(Dimocarpus longan* L.) based on next-generation sequencing technology. J Hortic Sci Biotechnol.

Li Y, Wang X, Li C, Hu S, Yu J, Song S (2014) Tramscriptome-wide N^6-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. RNA Biol. 11(9): 1180-1188.

Li Q, Zhang S, Wang J (2014) Transcriptome analysis of callus from *Picea balfouriana*. BMC Genomics. 15: 553.

Sweetman C, Wong DCJ, Ford CM, Drew DP (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. BMC Genomics. 13: 691.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5(7): 621–628.

Wei H, Chen X, Zong X, Shu H, Gao D, Liu Q (2015) Comparative Transcriptome Analysis of Genes Involved in Anthocyanin Biosynthesis in the Red and Yellow Fruits of Sweet Cherry (*Prunus avium* L.). PLOS One.