BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(6)

Computer aided gene mining for gingerol biosynthesis

Priyanka James, Bincy Baby, Sona Charles, Lekshmysree Saraschandran Nair & Puthiyaveetil Abdulla Nazeem*

Bioinformatics Centre (DIC), College of Horticulture, Kerala Agricultural University, Vellanikkara, 680 656, Thrissur, Kerala, India; Puthiyaveetil Abdulla Nazeem – Email: bic@kau.in; Phone: +914872371994; Fax: +914872371994; *Corresponding author

Received April 23, 2015; Revised June 15, 2015; Accepted June 16, 2015; Published June 30, 2015

Abstract:
Inspite of the large body of genomic data obtained from the transcriptome of *Zingiber officinale,* very few studies have focused on the identification and characterization of miRNAs in gingerol biosynthesis. *Zingiber officinale* transcriptome was analyzed using EST dataset (38169 total) deposited in public domains. In this paper computational functional annotation of the available ESTs and identification of genes which play a significant role in gingerol biosynthesis are described. *Zingiber officinale* transcriptome was analyzed using EST dataset (38169 total) from ncbi. ESTs were clustered and assembled, resulting in 8624 contigs and 8821 singletons. Assembled dataset was then submitted to the EST functional annotation workflow including blast, gene ontology (go) analysis, and pathway enrichment by kyoto encyclopedia of genes and genomes (kegg) and interproscan. The unigene datasets were further exploited to identify simple sequence repeats that enable linkage mapping. A total of 409 simple sequence repeats were identified from the contigs. Furthermore we examined the existence of novel miRNAs from the ESTs in rhizome, root and leaf tissues. EST analysis revealed the presence of single hypothetical miRNA in rhizome tissue. The hypothetical miRNA is warranted to play an important role in controlling genes involved in gingerol biosynthesis and hence demands experimental validation. The assembly and associated information of transcriptome data provides a comprehensive functional and evolutionary characterization of genomics of *Zingiber officinale*. As an effort to make the genomic and transcriptomic data widely available to the public domain, the results were integrated into a web-based Ginger EST database which is freely accessible at http://www.kaubic.in/gingerest/.

Keywords: ESTs, miRNAs, transcriptomics, gingerol biosynthesis

Background:
Ginger (*Zingiber officinale* Rosc.) of the family Zingiberaceae is a significant tropical crop plant esteemed all over the world as the second most important spice and for its unique medicinal properties **[1, 2].** Ginger rhizome has several implementations in traditional systems of medicine, including ayurveda, chinese medicine, and western herbal medicine. It is conventionally used in the treatment of a wide array of ailments such as common cold, flu like symptoms, headaches, digestive disorders as well as muscular and rheumatic disorders **[3, 4].** It is important to emphasize its richness in a series of secondary metabolites including volatile and non-volatile phenolic compounds, the major ones possessing pungent characteristics which are responsible for the pharmacological activities of ginger. Gingerol is one of the most significant series of pungent oleoresin compounds. They are reported to possess essential physiological and pharmacological roles, despite which relatively little is known regarding its genome and gene expression patterns. Expressed Sequence Tags (ESTs) are direct evidences of gene expression and can provide useful resource for novel gene discovery, especially in non-model plants **[5].** EST database can be exploited for identifying and developing molecular markers such as SSRs, gene discovery, genome annotation and comparative genome analysis **[6, 7].** Comparative approaches based on ESTs derived from various kinds of tissues could significantly expedite gene discovery and genetic variations that may account for specific characteristics **[8, 9].** EST analysis also offer unique opportunities to elucidate the genes involved in biosynthetic pathways and the enzymes involved in pathways of secondary metabolite synthesis and

manipulation **[10].** Genes encoding enzymes involved in the biosynthesis of glycyrrhizin, ginsenosides, with anolides were successfully identified using EST analysis **[11, 12, 13].**

Despite the medicinal values of *Zingiber officinale*, information regarding its genome and transcriptome is limited. Many of the enzymes involved in biosynthesis of gingerol have not yet been identified. Here we describe the analysis of over 35,000 ESTs from rhizomes, leaves and roots that are available in dbEST (database of Expressed Sequence Tags). Functional annotations of ESTs were performed using BLASTn, BLASTx **[14],** Gene Ontology (GO) **[15]** and Kyoto Encyclopedia of Genes and Genomes (KEGG) **[16]** resources. Transcriptome similarity analysis between ginger and other plant species, identification of genes involved in gingerol biosynthesis and Simple Sequence Repeats (SSR) analysis were also performed.

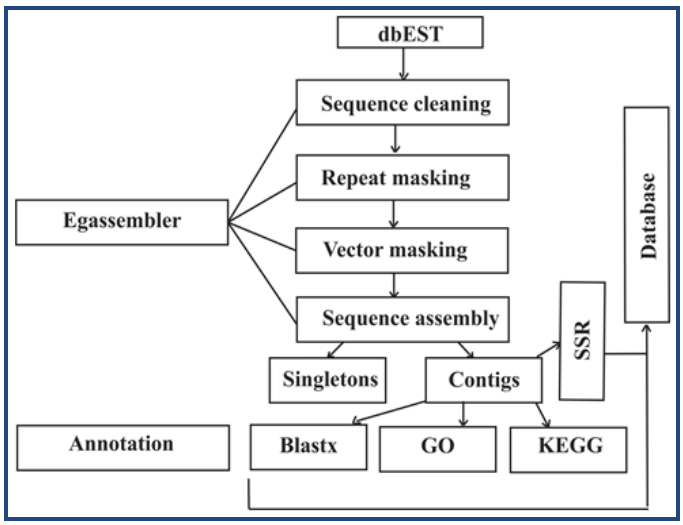

Figure 1: Schematic representation of the functional analysis pipeline for ginger EST pre-processing and database construction

Methodology:

Data generation and EST pre-processing

A total of 38,169 expressed sequence tags from different tissues including leaf, rhizome and root deposited in NCBI (http://www.ncbi.nlm.nih.gov/) EST database was obtained through keyword search ('*Zingiber officinale*', 'Rhizome', 'Root', 'Leaf'). Raw EST sequences contain contaminants such as poly A/T tail, low complexity sequences, vector sequences etc. The ESTs were thus pre-processed to generate unigene dataset which contains a set of non-redundant sequences composed of singletons and contigs. Different steps required for sequence processing integrated on a web server, EGassembler **[17]** were used for the generation of unigene dataset. ESTs obtained from different tissues were subjected to the server with default parameters. The sequences were cleaned, followed by masking of low quality sequence, removal of repetitive elements and organelle sequence masking. Pre-processed sequences were then assembled into 8624 contigs and 8821 singletons using CAP3 **[18]** software **(Figure 1)**.

Functional analysis

Blast2GO (B2GO), **[19]** comprehensive software for sequence analysis was used for the functional annotation of ginger ESTs. Putative functions were analyzed by performing blastx against non-redundant protein sequence database. For further perception of the functional classification, the EST contigs were submitted to Gene Ontology analysis (GO). Cellular component, biological process, and molecular function were classified for these contigs. Unigenes were further subjected to pathway analysis using KEGG. InterPro Scan **[20]** embedded with B2GO was utilized to obtain the protein domain information for the putative sequences.

SSR detection

Simple Sequence Repeats (SSRs) analysis on contigs of ginger was performed using SSR Finder from GRAMENE (ftp://ftp.gramene.org/ pub/gramene/software/scripts/ ssr.pl). The parameters were set for detection of di-, tri-, tetra-, penta-, and hexa-nucleotide units with a minimum of six, five, four, four and four repeats respectively **[21].** Primer pairs for each detected SSR were designed using Eprimer3 from EMBOSS software packages (http://emboss.sourceforge.net/). Parameters used to design flanking primers were primer length 18 to 24 nt with an optimum of 20 nt, annealing temperature with optimum 60°C, GC content ranging from 40 to 60% with an optimum 50% **[22].**

Identification of novel miRNAs

Raw EST sequences from *Zingiber officinale* and EST contigs obtained after data preprocessing were combined to form the subject dataset. All known Viridiplantae miRNAs from the publicly available EST database, miRBase (Release 20: June 2013, http://www.mirbase.org/) **[23]** was used as a reference set and performed homology searches using blastn. Blast parameter settings were as follows: expect value - 0.01; the maximum number of alignments -100. Python script was developed for the identification of miRNAs. EST sequences with n/n, n-1/n and n-2/n nucleotide mismatches were identified and retained for further analysis. A 60-400 nt sliding window was chosen to predict the secondary structure of pre-miRNAs using mfold algorithm (~14000 bp for each EST). If the length of EST sequence is <400nt then the whole sequence was taken as the candidate to check whether it is a miRNA precursor. Blastx was performed on these sequences to remove the protein coding sequences. The pre-miRNA secondary structure was predicted by using RNAfold (http:// rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) **[24]** by applying default parameters. The predicted hairpin-like structures were further assessed for their MFE, GC content, homology with mature miRNA sequence and ensured that the miRNA: miRNA* duplexes had less than 4 mismatches and it should arise from opposite arms and is present completely in the opposite arm. To ensure the stem-loop precursor could be precisely processed into mature miRNA, the predicted structures were examined according to the following criteria:
(1) An appropriate stem-loop structure can be assigned to the secondary structure; (2) The presence of mature miRNA sequence in one arm of the hairpin; (3) <=3 nucleotide substitutions when compared with existing miRNA; (4) < 6 mismatches with miRNA∗ of the opposite arm; (5) No loops should be present in miRNA∗; (6) Predicted secondary structures should have minimal folding free energy (MFE).

Database construction

Raw ESTs, assembled data and annotation results were integrated into a web based database GINGEREST. Front end

and backend were created using PhP and MySQL5.0 respectively. GINGEREST database was deployed on Apache HTTP server and runs on a server managed by the Windows operating system.

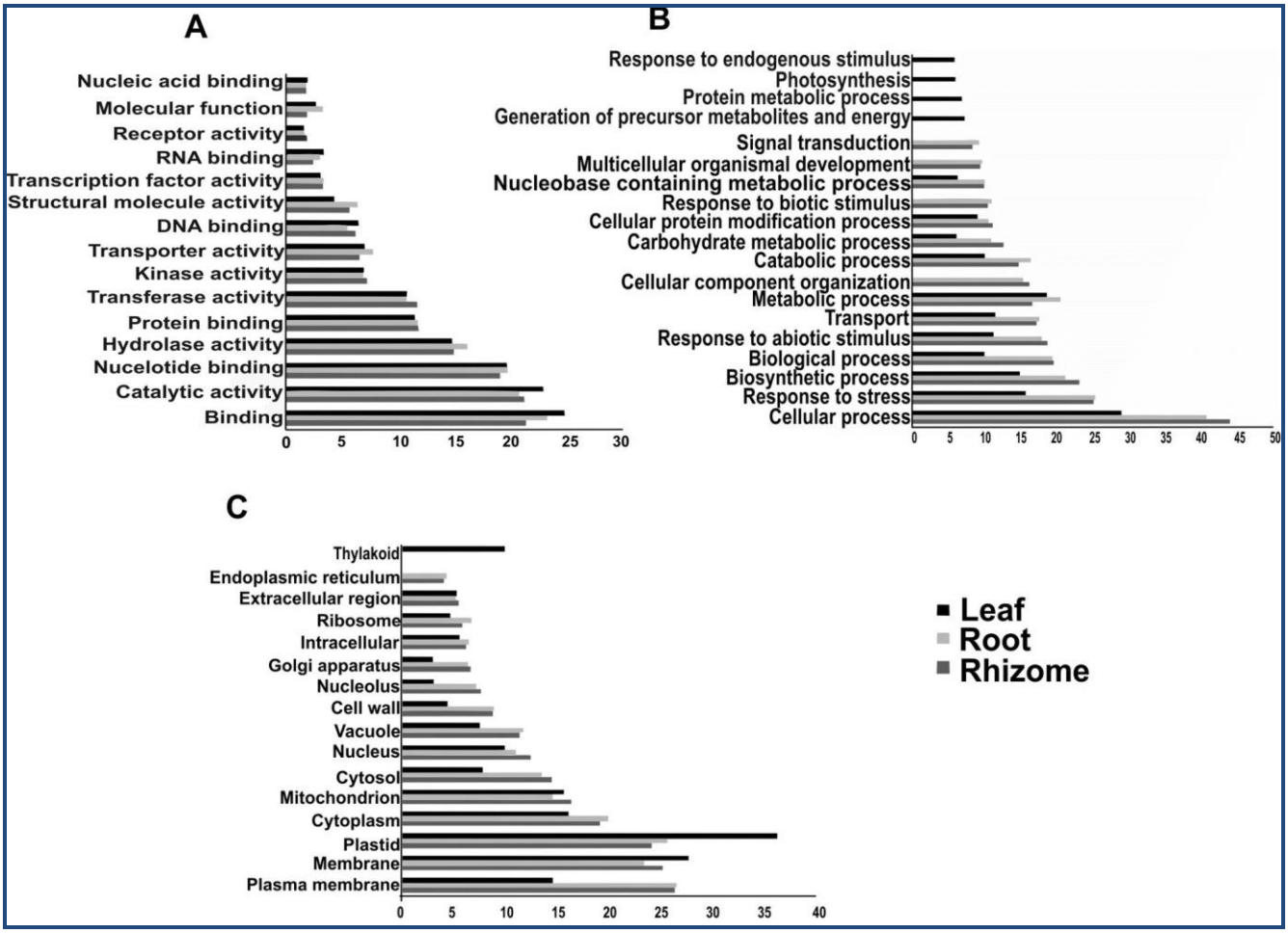

Figure 2: Percentage of unigenes from ginger ESTs involved in gene Ontologies. **(A)** Molecular Function **(B)** Biological Process **(C)** Cellular Component.

Results & Discussion:

Assembly and annotation of EST's

A total of 38,169 ginger (*Zingiber officinale*) ESTs from different tissues were downloaded from NCBI. The largest library of EST fragments was obtained from ginger leaf. Out of these EST sequences, 8624 contigs (unique sequences) were obtained after removing the non-informative sequence stretches. The CAP3 results included 8624 contigs and 8821 singlets. In order to capture the most informative and putative functions of ginger ESTs, all unique sequences were annotated using blastx against public non-redundant protein (nr) in NCBI using an E-value cutoff of 1e-15 and a maximum of 20 BLAST hits per sequence. Majority of BLASTX hits for sequences were from the following plants: *Vitis vinifera, Populus trichocarpa, Oryza sativa, Zea mays, Brachypodium distachyon, Ricinus communis, Sorghum bicolour, Glycine max.*

Functional analysis based on Gene Ontology

Gene Ontology annotation provides description of genes in terms of their associated molecular functions, cellular components, and biological processes. Functional classifications of ESTs in terms of gene ontology were retrieved for 8624 contigs using Gene Ontology (GO) analysis. GO term provides a broad overview of the groups of genes cataloged in the transcriptome. The total EST sequences were found coding for 2356 cellular compounds (CC), 928 molecular functions (MF), and 7258 biological processes (BP). Of these EST sequences, 640 were mapped under CC and MF, 2539 were mapped under CC and BP, 1511 were mapped under BP and MF and 4474 EST sequences were annotated according to all the three GO vocabularies. All annotations were simplified using plant-specific GOslim. EST contigs were grouped with respect to their GO terms, belonging to only one, a combination of two, or all three, and were organized in a venn diagram. A large number of unique sequences were found mapped under molecular functions such as binding, catalytic activity, nucleotide binding, protein binding, and transferase activity. The second most represented category was biological processes, including unique sequences associated with cellular process, response to stress, biosynthetic process, biological process, response to abiotic stress, transport and metabolic process. Finally, the cellular compound category was mapped for plasma membrane, plastid, cytoplasm, mitochondria, cytosol, and nucleus. According to annotation classification, the largest cellular component found for ginger ESTs was from plasmid (15.33%) and the smallest were from golgi apparatus and endoplasmic reticulum (2%). The majority of molecular functions identified were associated with binding (16.7%) and

catalytic activities (15.67%). The major biological processes were involved in cellular process (16.7%), response to stress (9.67%) and biosynthetic process (8.67%) **(Figure 2).**

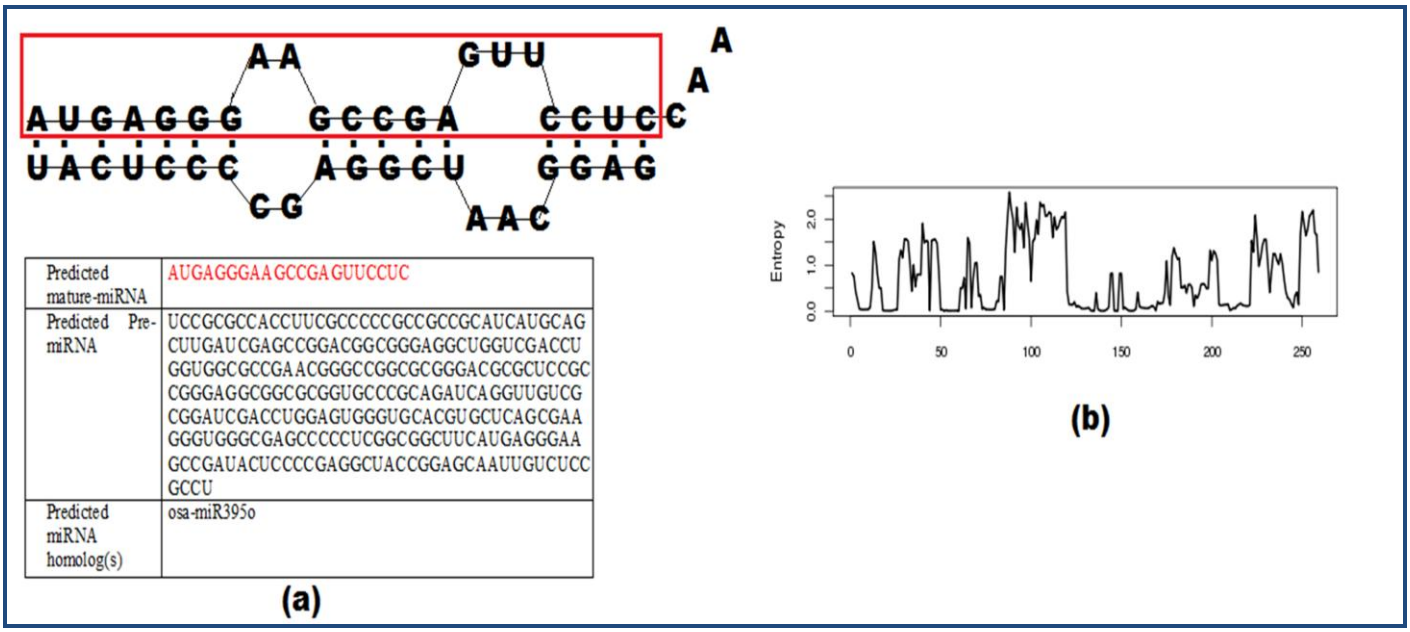

Figure 3: Predicted mature miRNA sequence from pre-miRNA in Zingiber officianale. **(a)** Mature miRNA sequence based paired with miRNA* **(b)** Positional entropy of each base in the pre-miRNA**.**

Protein family and functional domains

Annotation of EST sequences involves the identification of protein coding regions. InterPro Scan combines different protein family and domain signature identification method and can be used for the prediction of protein function. All sequences were subjected to InterPro Scan to identify functional domains associated with ginger EST sequences. InterPro Scan associated with Blast2GO was able to assign InterPro ID to 6071 sequences out of 8624 contigs whereas 2553 sequences were left without domain mapping. Gene ontology and protein domain information was obtained for 3021 ESTs, while 3050 ESTs were mapped by InterPro but not to a specific GO. Our analysis included secondary protein database like Pfam, SMART, Gene3D, PROFILE, ProDom, SUPERFAMILY, PANTHER, PIR, and TIGR. Along with functional and structural annotation of protein sequences, InterProScan incorporates prediction of signal peptide sequences (SignalP) and transmembrane helices (TMHMM).

Biochemical analysis based on KEGG pathways

KEGG is one of the most accessible biochemical pathway databases that provide functional annotation of genes according to their associated biochemical pathways. Putative *Zingiber officinale* transcripts from rhizome, root and leaf were subjected to search against KEGG database for annotations based on similarity search yielding KEGG enzyme commission (EC) number and pathway map. KEGG annotations were assigned to ESTs derived from Rhizome (395), Root (288), and Leaf (393) of which only 115 (Rhizome), 117 (Root) and 112 (Leaf) respective sequences were assigned to biochemical pathways. The metabolic pathways were well represented among unique sequences of *Zingiber officinale*, most of which were associated with amino acid metabolism, carbohydrate metabolism, lipid metabolism, energy metabolism, and the biosynthesis of secondary metabolites.

Genes involved in Gingerol Biosynthesis

Gingerols were identified as the major bioactive components in ginger rhizome and 6-gingerol is the most abundant constituent in the gingerol series. Biosynthesis of gingerol takes place through phenylpropanoid pathway. ESTs involved in phenylpropanoid pathway and gingerol biosynthesis pathway were identified from the ginger transcriptome (Rhizome, Root and Leaf) using *in silico* approaches. ESTs encoding enzymes associated with phenylpropanoid pathway, including phenylalanine ammonia lyase, polyketide synthases, p-coumaroyl shikimate transferase, p-coumaroyl quinate transferase, caffeic acid O-methyltransferase, and caffeoyl-CoA O-methyltransferase were observed in the sequences. Phenylalanine ammonia lyase, the most extensively studied enzyme in the phenylpropanoid pathway and it is the entry point into the potential pathways leading to the formation of gingerols **[25].** Phenylalanine ammonia lyase ESTs were found in ginger root alone and was not found in rhizome and leaf. Caffeoyl-CoA O-methyltransferase (CCOMT) ESTs were found in all the three tissues. Cinnamate 4-hydroxylase ESTs were most abundantly found in all tissues among the phenylpropanoid pathway enzymes. The pathway enrichment analysis of ginger ESTs based on KEGG annotation revealed 1,075 contigs involved in 345 different pathways. Out of the 1,075 contigs, 11 contigs were involved particularly in gingerol biosynthesis. Contigs encoding enzymes involved in the biosynthesis of gingerol is depicted in **Table 1 (see supplementary material)**.

SSR detection

Simple Sequence Repeats or SSRs are tandem repeats of a stretch of 2 to 6 nucleotide units and are important molecular markers being utilized in crop improvement programme. A total of 409 SSRs were identified from the three tissues,

rhizome, root and leaf. Trinucleotide SSRs were found larger in number in rhizome whereas dinucleotide SSRs were found more in leaf. The dominant motifs are TA/GGC/TC in leaf contigs, CT/AT/GCC/GAA in rhizome and TA/CT/TC in root contigs **Table 2 (see supplementary material)**.

Identification of novel miRNA

Scanning of ESTs for miRNAs revealed 4456 pre-miRNA sequences. Structures with minimal free energy and geometry were filtered and further optimized. A single possible miRNA was mapped to be present in rhizome tissue with MFE= -122.50 kcal/mol from a pre-miRNA sequence of length 259nt. The modelled miRNA structure along with the positional entropy of the pre-miRNA is described in **Figure 3**. Further target prediction was performed to see whether they regulate any significant genes involved in gingerol biosynthesis. However no genes were found to be targeted. The predicted miRNA obtained after computational analysis is to be validated experimentally.

Web based GINGEREST database

Ginger ESTs and its annotations were integrated into a web-based knowledge base in order to make the data widely available. Annotation results include contigs, blast results, Gene Ontology analysis, pathway maps and Simple Sequence Repeats for the three tissues, leaf, rhizome and root. Raw ESTs, assembled data and annotation results can be accessed freely through the web interface (http://www.kaubic.in/gingerest).

Conclusion:

Despite recent improvements in modern medicine, there is a growing interest in herbal drugs due to the undesired side effects. Gingerol obtained from *Zingiber officinale* is a widely used phytocompound that has profound physiological effects. The study aims at mining the major genes involved in gingerol biosynthesis pathway in a leaf, root and rhizome. The annotation of ESTs from *Zingiber officinale* was performed using computational methods. A total of 8624 assembled contigs have been successfully annotated and the genes involved in the gingerol synthesis were predicted. Out of the 8624 assembled contigs, 11 contigs were involved particularly in gingerol biosynthesis. The genes were further checked for the presence of regulatory miRNAs revealing one hypothetical miRNA wholes role might be significant in controlling gene expression in rhizome tissue. This analysis provides significant resources for gene discovery as well as identification of novel RNAs in gene silencing in *Zingiber officinale* and will pave the way to characterize the biosynthetic pathways of gingerol. The analytical results on ginger ESTs such as similarity search, gene ontology, pathway analysis and SSRs are integrated into a user friendly freely available web based database to share the transcriptomic data, gingerest.

Acknowledgements:

This project was supported by the Department of Biotechnology (DBT), Government of India.

Reference:

[1] Nayak S *et al. Z Naturforsch C*. 2005 **60**: 485 [PMID: 16047412]
[2] Harisaranraj R *et al. Glob J Mol Sci*. 2009 **4**: 103
[3] Ramirez-Ahumada MDC *et al. Phytochemistry* 2006 **67**: 2017 [PMID: 16890967]
[4] Tchombe L N *et al. ISESCO J Sci Technol*. 2012 **8**: 64
[5] Mondego J M *et al. BMC Plant Biol.* 2011 **11**: 30 [PMID: 21303543]
[6] Sreenivasulu N *et al. Curr Sci*. 2002 83.
[7] Qiu L *et al. BMC Plant Biol.* 2010 **10**: 278 [PMID: 21162723]
[8] Verza *et al. Plant Mol Biol.* 2005 **59**: 363 [PMID: 16247562]
[9] Mochida K & Shinozaki K, *Plant Cell Physiol* 2010 **51**: 497 [PMID: 20208064]
[10] Park *et al. Plant Sci*. 2004 **166**: 953
[11] Sudo H *et al. Plant Biotechnol*. 2009 **26**: 105
[12] Sathiyamoorthy S *et al. Mol Biol Rep*. 2010 **37**: 3465 [PMID: 19943115]
[13] Senthil K *et al. Mol Biol Rep*. 2010 **37**: 893 [PMID: 19669665]
[14] Altschul SF *et al. Nucleic Acids Res*. 1997 **25**: 3389 [PMID: 9254694]
[15] Ashburner M *et al. Nat Genet.* 2011 **25** :25 [PMID: 10802651]
[16] Kanehisa M & Goto S, *Nucleic Acids Res*. 2000 **28**: 27 [PMID: 10592173]
[17] Masoudi-Nejad A *et al. Nucleic Acids Res*. 2006 **34**: W459 [PMID: 16845049]
[18] Huang X, *Genome Res.* 1999 **9**: 868 [PMID: 10508846]
[19] Conesa A *et al. Bioinformatics* 2005 **21**: 3674 [PMID: 16081474]
[20] Quevillon E *et al. Nucleic Acids Res.* 2005 **33**: W116 [PMID: 16081474]
[21] Zeng S *et al. BMC Genomics.* 2010. **11**: 94 [PMID: 20141623]
[22] Rychlik W, *Mol Biotechnol.* 1995 **3**: 129 [PMID: 7620973]
[23] Griffiths-Jones S *et al. Nucleic Acids Res*. 2008 **36**: D154 [PMID: 17991681]
[24] Hofacker I L *et al. Monatshefte für Chemie / Chem Mon*. 1994 **125**: 167
[25] Koo HJ *et al. BMC Plant Biol*. 2013 **13**: 27 [PMID: 23410187]

Edited by P Kangueane

Citation: **James *et al.*** Bioinformation 11(6): 316-321 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Distribution of contigs from encoding enzymes

Tissue	Contig	Enzyme	Pathway	E.C. Number
Rhizome	Contig1294	Monooxygenase - Cinnamic Acid & Cinnamyl-CoA	Phenylpropanoid biosynthesis & GingerolBiosynthesis	ec:1.14.13.11
	Contig1614	O-methyltransferase – Caffeoyl-CoA & 5-Hydroxyferuloyl-CoA	Phenylpropanoid biosynthesis & Gingerol Biosynthesis	ec:2.1.1.104
	Contig587, Contig1506	Monooxygenase - Cinnamic Acid & Cinnamyl-CoA	Phenylpropanoid biosynthesis & Gingerol Biosynthesis	ec:1.14.13.11
	Contig571, Contig2872	O-methyltransferase – Caffeoyl-CoA , 5-Hydroxyferuloyl-CoA & caffeic acid 3-O-methyltransferase	Phenylpropanoid biosynthesis & Gingerol Biosynthesis	ec:2.1.1.104 ec:2.1.1.68
Root	Contig2772, Contig3034	ammonia-lyase - phenylalanine ammonialyase & phenylalanine/tyrosine ammonia-lyase	Phenylpropanoid biosynthesis	ec:4.3.1.24 ec:4.3.1.25
	Contig1304	Monooxygenase - Cinnamic Acid & Cinnamyl-CoA	Phenylpropanoid biosynthesis & Gingerol Biosynthesis	ec:1.14.13.11
Leaf	Contig2600, Contig814	O-methyltransferase – Caffeoyl-CoA, caffeic acid 3-O-methyltransferase	Phenylpropanoid biosynthesis & Gingerol Biosynthesis	ec:2.1.1.104 ec:2.1.1.68

Table 2: Distribution of SSRs in Ginger EST.

Tissue	Total	Tri	Di
Rhizome	109	88 (61 %)	37 (26 %)
Root	144	56 (51 %)	30 (28 %)
Leaf	147	86 (58 %)	39 (27 %)
Tissue	**Total**	**Tri**	**Di**
Rhizome	109	88 (61 %)	37 (26 %)
Root	144	56 (51 %)	30 (28 %)
Leaf	147	86 (58 %)	39 (27 %)