Plant Mol Biol (2014) 85:193–208
DOI 10.1007/s11103-014-0179-8

Characterization of mango (*Mangifera indica* L.) transcriptome and chloroplast genome

M. Kamran Azim · Ishtaiq A. Khan · Yong Zhang

Received: 26 October 2013 / Accepted: 4 February 2014 / Published online: 11 February 2014

Abstract We characterized mango leaf transcriptome and chloroplast genome using next generation DNA sequencing. The RNA-seq output of mango transcriptome generated >12 million reads (total nucleotides sequenced >1 Gb). De novo transcriptome assembly generated 30,509 unigenes with lengths in the range of 300 to ≥3,000 nt and 67× depth of coverage. Blast searching against non-redundant nucleotide databases and several Viridiplantae genomic datasets annotated 24,593 mango unigenes (80 % of total) and identified *Citrus sinensis* as closest neighbor of mango with 9,141 (37 %) matched sequences. The annotation with gene ontology and Clusters of Orthologous Group terms categorized unigene sequences into 57 and 25 classes, respectively. More than 13,500 unigenes were assigned to 293 KEGG pathways. Besides major plant biology related pathways, KEGG based gene annotation pointed out active presence of an array of biochemical pathways involved in (a) biosynthesis of bioactive flavonoids, flavones and flavonols, (b) biosynthesis of terpenoids and lignins and (c) plant hormone signal transduction. The mango transcriptome sequences revealed 235 proteases belonging to five catalytic classes of proteolytic enzymes. The draft genome of mango chloroplast (cp) was obtained by a combination of Sanger and next generation sequencing. The draft mango cp genome size is 151,173 bp with a pair of inverted repeats of 27,093 bp separated by small and large single copy regions, respectively. Out of 139 genes in mango cp genome, 91 found to be protein coding. Sequence analysis revealed cp genome of *C. sinensis* as closest neighbor of mango. We found 51 short repeats in mango cp genome supposed to be associated with extensive rearrangements. This is the first report of transcriptome and chloroplast genome analysis of any Anacardiaceae family member.

Keywords Transcriptome analysis · RNA-seq · Anacardiaceae · Plant genome

Introduction

Mango (*Mangifera indica* L.), a member of family Anacardiaceae is an important fruit crop which is commercially grown in over hundred tropical and subtropical countries (Mukherjee and Litz 2009). According to Food and Agriculture Organization (FAO) of United Nations, after banana, mango is the dominant tropical fruit variety produced worldwide, followed by pineapples, papaya and avocado (http://www.fao.org/docrep/006/y5143e/y5143e1a.htm). Major mango growing countries are India, China, Thailand, Pakistan, Australia, Indonesia, Bangladesh, Philippines, Nigeria, Myanmar and Egypt.

The mango tree is considered to have evolved in the rainforests of South and South-east Asia (Krishna and Singh 2007). Full-grown mango trees reach a height of 40 m and can stay alive for several 100 years. Mango leaves are exstipulate, simple and usually alternate; depending on the cultivar, leaf morphology is highly variable (Mukherjee

Electronic supplementary material The online version of this article (doi:10.1007/s11103-014-0179-8) contains supplementary material, which is available to authorized users.

M. K. Azim (✉) · I. A. Khan
Jamil-ur-Rehman Center for Genome Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan
e-mail: kamran.azim@iccs.edu; mkamranazim@yahoo.co.uk

Y. Zhang
BGI-Shenzhen, Beishan Road, Yantian District, Shenzhen 518083, China

and Litz 2009). The inflorescence of mango flowers is rigid and erect, usually 30 cm long, widely branched and is always cymose. Mango fruit varieties have been known for attractive colours, savouring smell, delightful taste and high nutritional value (Mukherjee and Litz 2009). Mango is an ever green dicot angiosperm. Although several tetraploid individuals were reported, mango is usually a diploid tree (Mukherjee 1950; Duval et al. 2005; Viruel et al. 2005; Schnell et al. 2005, 2006). It has 2n = 40 chromosomes with estimated genome size of 441 mega basepairs (http://data.kew.org/cvalues/).

During last few years, a number of reports addressed the genetic diversity of mango for application in cultivar identification using PCR and sequencing based techniques viz. RAPD, ISSR, DAMD etc. (Srivastava et al. 2012; Chinag et al. 2012; Souza et al. 2011; Rocha et al. 2012; Ravishankar et al. 2011; Hirano et al. 2010; Khan and Azim 2011). Despite its global importance, genomic sequence resources available for the mango tree are scarce. As of October 2013, there are only 684 highly redundant sequence entries in the GenBank for mango. Large scale discovery and characterization of functional genes via genome sequencing or global exploration of the transcriptome are required for better understanding of fundamental molecular biology of mango.

Recently development of RNA sequencing (RNA-seq) methodology has facilitated the analysis of transcriptomes of a number of crop and medicinal plants (Xu et al. 2013; Duangjit et al. 2013; Sara et al. 2010). RNA-seq is characterized by sequencing of the transcriptome using massively paralleled next generation DNA sequencing technology. It is among the most popular techniques of NGS (Strickler et al. 2012). RNA-seq generate millions of short cDNA reads which either aligned to a reference genome or reference transcripts, or assembled de novo to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene (Mortazavi et al. 2008). Sequencing of RNA has long been recognized as an efficient method for gene discovery and remains the gold standard for annotation of both coding and non-coding genes (Adams et al. 1991; Haas and Zody 2010). Furthermore, the RNA-seq method offers a holistic view of the transcriptome, revealing many novel transcribed regions, splice isoforms, single nucleotide polymorphisms (SNPs) and the precise location of transcription boundaries (Li et al. 2010; Wilhelm et al. 2010). RNA-seq is expected to revolutionize the manner in which eukaryotic transcriptomes are analyzed (Wang et al. 2009).

Here we report the characterization of mango leaf transcriptome and chloroplast genome using next generation sequencing. We generated over 1.0 billion bases of high quality DNA sequence of mango using RNA-seq technology and demonstrated the suitability of short-read sequencing for de novo assembly and annotation of genes without prior genome information. Moreover, we also sequenced the mango chloroplast genome with the help of a blend of Sanger and next generation sequencing. The results provide a cost effective and efficient way to global discovery of new functional genes in mango.

Materials and methods

Plant materials

Mangifera indica cultivar Langra used in this study was grown in Botanical Gardens of University of Karachi, Karachi, Pakistan. The leaf specimen of the tree used is preserved in the Herbarium of Department of Botany, University of Karachi, Karachi, Pakistan. Healthy and mature leaves from same tree were taken for RNA-seq and chloroplast DNA sequencing.

RNA isolation, cDNA synthesis and sequencing

Total RNA was isolated from mango leaves using RNAesy kit (Qiagen GmbH, Hilden, Germany). RNA integrity was confirmed using a 2100 Bioanalyzer (Agilent Inc., USA). Beads with oligo(dT) were used to isolate poly(A) mRNA from total RNA (Qiagen GmbH, Hilden, Germany). The purified mRNA was fragmented into short fragments using divalent cations under elevated temperature. The cDNA was synthesized with random hexamer primers and mRNA fragments as templates using Superscript®III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA). Short fragments were purified with the PCR extraction kit (Qiagen GmbH, Hilden, Germany) followed by end repair and poly(A) addition. The short fragments were then connected with sequencing adapters. The paired-end library (with 200 bp insert size) was prepared following the manufacturer's protocol (Illumina Inc., San Diego, CA, USA). Finally, the library was sequenced using Illumina HiSeq2000 (Illumina Inc., San Diego, CA, USA). The library was linked the flow-cell containing complementary adapters, and then bound fragments were amplified to create 'clusters'. The adapters were designed to allow selective cleavage of the forward DNA strand after resynthesis of the reverse strand during sequencing. The copied reverse strand was then used to sequence from the opposite end of the fragment. The raw reads were cleaned by removing adaptor sequences, empty read and low quality sequences.

Transcriptome de novo assembly

Transcriptome denovo assembly was carried out with short reads assembling program—Trinity (Grabherr et al. 2011).

Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly. Trinity applies these programs one after the other to process large volumes of RNA-seq reads. Firstly, Inchworm assembles the RNA-seq data into the unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts. Secondly, Chrysalis clusters the Inchworm 'contigs' into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptional complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs. Finally, Butterfly processes the individual graphs in parallel, tracing the paths that reads and pairs of reads take within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes. The resultant sequences are termed as 'unigenes'.

Annotation and classification of unigenes

Mango unigenes were analyzed by BLASTN (Zhang et al. 2000) against the NR database (NCBI non-redundant sequence database) with an E-value cut-off of 10^{-5}. The coding sequences in mango unigenes were also analyzed using BLASTN against genomic sequence datasets of *Citrus sinensis* (sweet orange), *Populus trichocarpa* (black cottonwood), *Vitis vinifera* (Grapevine), *Ricinus communis* (castor bean), *Glycine max* (soya bean), *Medicago truncatula* (Barrel Medic) and *Arabidopsis thaliana*. Unigene sequences were further aligned by BLASTX to protein databases; SwissProt, KEGG (Kanehisa et al. 2008) and COG. This step retrieved proteins with the highest sequence similarity with the given unigenes along with their functional annotations. In case of disagreement between databases, a priority order of NR, Swiss-Prot, KEGG and COG was followed. For unigenes that did not align to any of the above databases, ESTScan software (Iseli et al. 1999) was used to predict their coding regions and decide sequence direction.

The Blast2GO and WEGO programs were used for GO functional annotations, KEGG and COG analysis of mango unigenes (Conesa et al. 2005; Ye et al. 2006). The analysis mapped annotated unigenes to GO terms and calculated the number of unigenes associated with every term.

Comparison with Genbank *M. indica* sequence entries

Six hundred and eighty four mango sequences (ESTs and nucleotide sequences) were downloaded from the GenBank and used for nucleotide BLAST search against 30,509 unigenes using an E-value cut-off of 10^{-5}.

Mango chloroplast genome sequencing

A combination of Sanger-based and next-generation sequencing strategies were used for mango chloroplast DNA (cpDNA) sequencing. The mango leaves (5.0 grams) were used for isolation of total DNA using AxyGen multisource DNA miniprepration kit (Axygen Scientific, USA). Initially, a primer walking strategy termed as "ASAP: amplification, sequencing and annotation of plastomes" (Dhingra and Folta 2005) was used for amplification and Sanger-based sequencing of inverted repeat (IR) and large single copy (LSC) regions of cpDNA. Briefly, purified mango DNA was used for generation of 6.0 kb amplicons with consensus set of primers (Supplementary data) (Dhingra and Folta 2005). The 6.0 kb amplicons were then used for generation of 1.0 kb fragments using internal sets of primers (Supplementary data) corresponding to 6.0 kb amplicons. Later on, gap filling primers were designed to fill the gaps within the inverted repeat region (Supplementary data). The Sanger-based sequencing of the above mentioned fragments was carried out by CEQ8000 Genetic Analyzer (Beckman Coulter Inc., USA). For cycle sequencing reactions, the DTCS kit (Beckman Coulter Inc., USA) was used, with conditions as recommended by the suppliers. Further mango cpDNA sequencing was carried out by next-generation sequencing technology of GS FLX System (Roche Inc., USA) using Titanium Mini kit and 7.0 μg of purified DNA.

The sequences obtained from Sanger-based sequencing were assembled using the Lasergene package version 7.1 (DNASTAR Inc., Madison, WI, USA). The sequencing data from the GS FLX system was assembled using CLC Genomics Workbench version 3.5.1 (CLCbio, Denmark). The assembled sequences obtained from Sanger-based and next generation sequencing were combined using CLC Genomics Workbench (CLC bio, Denmark). Genome annotation was performed through the DOGMA server (Dual Organellar Genome Annotator; (Wyman et al. 2004), ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/), and BLAST (Altschul et al. 1990). Repeat analysis was performed using the REPuter program (Kurtz et al. 2001). A circular genome map of mango cpDNA was constructed using the GenomeVx tool (Conant and Wolfe 2008). Construction of multiple alignments and phylogenetic trees of complete cpDNA sequences was carried out by the mVISTA comparative genomics tool (Frazer et al. 2004).

Results and discussion

RNA-seq and de novo assembly of mango transcriptomic sequences

To characterize the mango transcriptome, total RNAs were isolated from leaves. After DNase treatment and

Table 1 Output statistics of mango RNA-seq experiment

Total raw reads	Total clean reads	Total clean nucleotides (nt)	Q20 (%)	N (%)	GC (%)
15,851,736	12,153,196	1,093,787,640	91.96	0.01	44.73

* Total clean nucleotides = total clean reads1 × read1 size + total clean reads2 × read2 size

confirmation of RNA integrity using bioanalyzer, total RNA was used for mRNA preparation, fragmentation and cDNA synthesis. After cleaning and quality checks, Illumina NGS sequencing generated 12,153,196 sequence reads, encompassing 1,093,787,640 nucleotides, with each sequence read averaging ca. 90 bp in length (Table 1). This dataset has been submitted to the NCBI Short Read Archive with accession number SRR947746.

Transcriptome de novo assembly was carried out with short reads assembling program—Trinity (Grabherr et al. 2011). Using this method, initially 85,651 contigs with a mean length of 238 were generated (Table 2). As shown in the contig length distribution in Fig. 1, the contig count is inversely proportional to length. Later on, the contigs sequences were assembled into 30,509 unigenes. Mean size of unigenes was 536 bp with lengths in the range of 300 to >3,000 bp (Fig. 2). Sum of all unigenes length was 16,354,267 nucleotides with depth of coverage of 66.8× [depth of coverage was calculated by dividing number of clean nucleotides i.e. 1,093,787,640 by sum of nucleotides in 30,509 unigenes (16,354,267 nt)]. The unigenes were divided into two clusters. In cluster-1, unigenes with sequence homology >70 % were grouped (prefixed CL); whereas cluster-2 contained singleton unigenes (prefixed unigene). The assembled mango transcriptome sequences have been deposited in the Genbank with Transcription Shotgun Assembly number SUB363843.

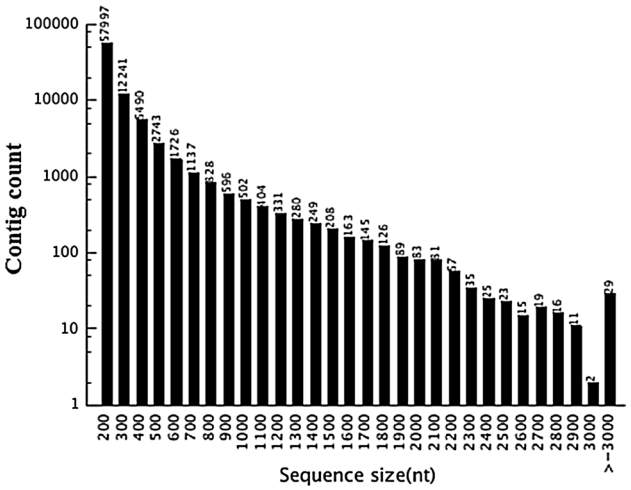

Fig. 1 Length distribution of contigs obtained after assembly of mango RNA-seq reads

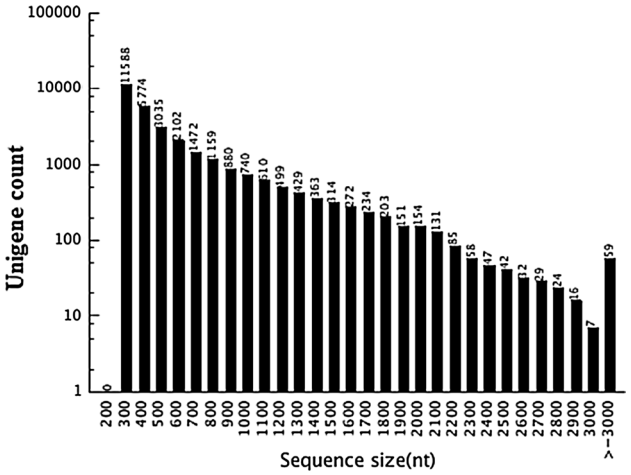

Fig. 2 Length distribution of assembled mango Unigene sequences

Comparison of assembled unigenes with mango sequences in Genbank

As of October 2013, the Genbank contained 684 sequence entries of mango (both gene sequences and ESTs); many of them are redundant, analyzed in the frame of phylogenetic studies and yet unpublished. After removing the redundancy, Genbank mango sequences were divided into complete and partial gene sequences. Subsequently, 30 and 59 nonredundant complete and partial mango gene sequences were found respectively (indicating 89 nonredundant mango gene sequences in Genbank). These sequences were submitted to the BLAST searches against 30,509 unigenes in present mango dataset. Out of 89 nonredundant mango gene sequences in Genbank, 76 (85.4 %) matched with assembled unigenes with a cutoff E-value of 10^{-5}. This analysis provided an evaluation of the quality of unigene sequences in present dataset. Further analysis showed that

Table 2 Statistics of assembly of NGS reads using Trinity

	Total number	Total length (nt)	Mean length (nt)	N50	Total consensus sequences	Distinct clusters	Distinct singletons
Contig	85,651	20,364,178	238	291	–	–	–
Unigene	30,509	16,354,267	536	687	30,509	11,403	19,106

Table 3 Summary of mango unigene annotation with the nucleotide and protein sequence databases NR, NT, Swiss-Prot, KEGG, COG and GO

NR	NT	Swiss-Prot	KEGG	COG	GO	ALL
24,593	21,974	14,447	13,561	7,594	21,054	25,453

The number of unigenes annotated with each database along with total number

majority of matched sequences had percent identities more than 95 % and E-value lower than 10^{-40} indicating reliability of alignment.

Recently mass spectrometry based proteomic analysis identified 538 proteins in mango leaves (Renuse et al. 2012). This mango leave proteome data contained 151 non-redundant protein sequences (length of the peptides used for protein matching were in the range of 09–34 amino acids; mean = 17). Comparative analysis showed that the present mango leaf transcriptome dataset was in full agreement with proteomic data. The integration of proteomic and transcriptomic data further evaluate the quality of assembled unigenes.

Annotation of unigenes

BLAST searching of mango unigene sequences against nucleotide databases (NR and NT), protein database Swiss-Prot, as well as KEGG and COG was performed with Evalue cutoff 10^{-5}. BLAST searches against NR database annotated 24,593 unigenes out of total 30,509 (80 %) unigenes (Table 3). Sequence analyses with NR database and several Viriplantae species genomic datasets showed that 37 % of *C. sinensis* coding region sequences matched with mango sequences, followed by *P. trichocarpa* (22.5 %), *V. vinifera* (18.3 %), *R. communis* (17.9 %), *G. max* (17.3 %), *M. truncatula* (10.5 %) and *Arabidopsis thaliana* (6.7 %) (Fig. 3; Table 4) (10.4 % of mango transcriptome sequences matched with sequences of other species).

Based on sequence homology, 21,054 mango unigenes were categorized into 57 functional groups, belonging to three main GO ontologies: molecular function, cellular component and biological process (Table 5). The results showed a high percentage of genes from categories of "cellular process", "metabolic process", "cell/cell part", "organelle", "catalytic", and "binding" with only a few genes related to "locomotion" and "nucleoid". On the other side, genes were not grouped in the categories of "cell killing", "extracellular matrix", "metallochaperone activity", "nutrient reservoir activity", "protein tag" and "translation regulator" (Table 5).

To further evaluate the function of the assembled unigenes, we searched the annotated unigenes involved in Clusters of Orthologous Groups (COG). Out of 24,593 NR Blast hits, 7,594 unigenes had a COG classification (Fig. 4). Among the 25 COG categories, the cluster for "general function prediction" represented the largest group, followed by three categories related to transcription, translation and posttranslational modification (categories J, K and O; see Fig. 4). Other most predicted gene functions were replication, recombination, repair and signaling (categories L and T; see Fig. 4). The categories "cell motility", "extracellular structures" and "nuclear structure" were least represented groups.

To identify active biochemical pathways in mango, the unigenes were mapped to the reference canonical pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2008). The KEGG database contains systematic analysis of inner-cell metabolic pathways and functions of gene products, which aid in studying the complex biological behaviors of genes. In total, 13,561 unigenes were assigned to 293 KEGG pathways. The pathways with most representation were "metabolic pathways", "photosynthesis", "transcription/translation", "DNA repair/recombination", "signal transduction" and "cell cycle". Considerable numbers of unigenes were identified as "vitamin metabolism", "N- and O-linked glycan biosynthesis", "ubiquitin mediated proteolysis/proteasome", "endocytosis" and "circadian rhythm". KEGG analysis discovered an array of biochemical pathways involved in biosynthesis of secondary metabolites/natural products known to participate in plant hormonal, flavors, aromatic and other cellular processes.

Characterization of natural products biosynthetic pathways

Annotation of mango unigenes identified many genes involved in phenylpropanoid, flavonoid, flavone/flavonol, isoflavonoid, terpenoid and carotenoid biosynthetic pathways (Pandit et al. 2010; Andrade Jde et al. 2012).

Phenylpropanoids comprise a large group of plant based natural compounds, derived from phenylalanine (Michal 1999). First step in the phenylpropanoid biosynthesis pathways is formation of cinnamic acid from phenylalanine which leads the formation of cinnamoyl-CoA, *p*-coumaroyl-CoA, feruloyl-CoA and sinapoyl-CoA. These CoA activated compounds are starting metabolites for the synthesis of lignins (the second most frequent class of compounds in biosphere after cellulose), flavonoids, flavones and flavonols (Michal 1999). KEGG analyses of mango transcriptome sequences revealed presence of 13 genes involved in biosynthesis of above mentioned CoA activated

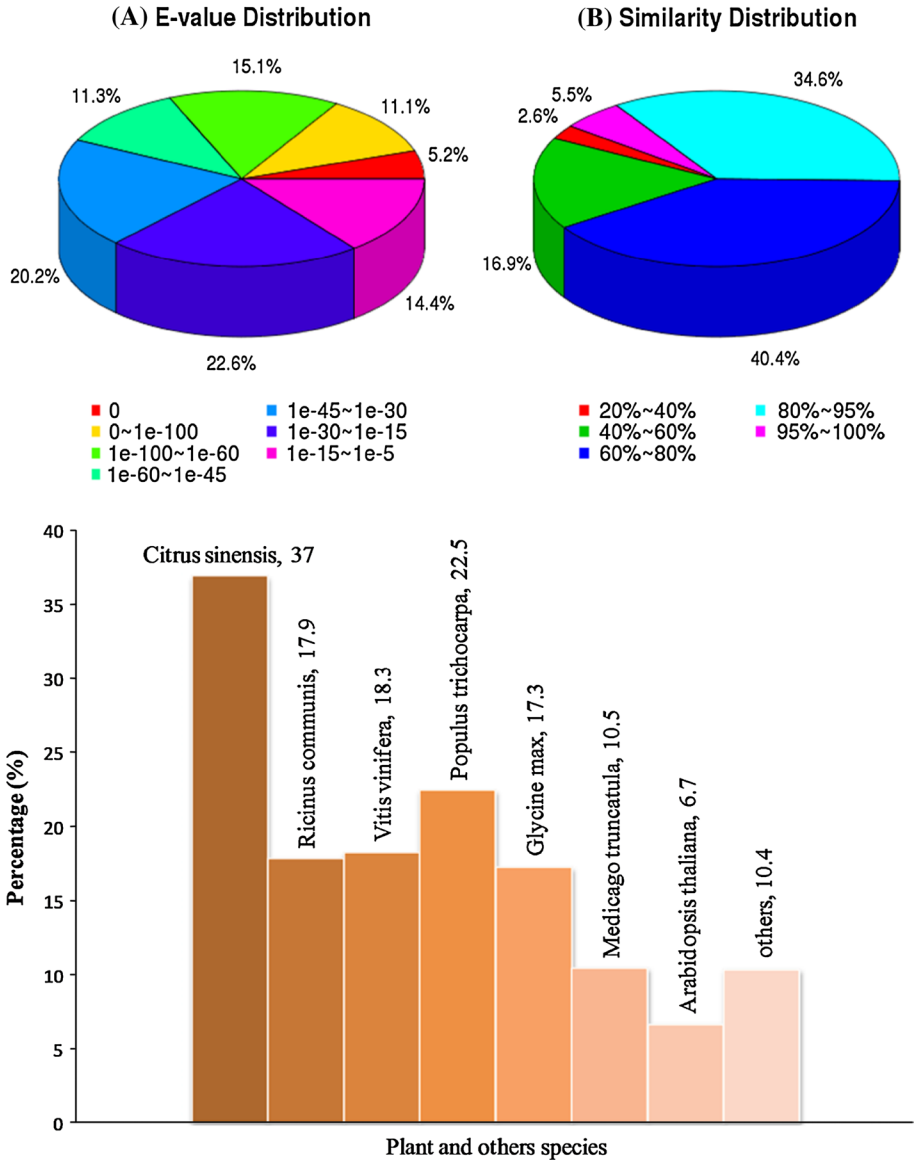

Fig. 3 Annotation of mango unigenes with nonredundant nucleotide database NR. Pie diagram showing **a** E-value distribution, **b** percentage identity distribution and **c** species distribution of mango unigenes

compounds. Analysis showed active pathways for the synthesis of several lignins in mango including guaiacyl lignin, 5-hydroxyguaiacyl lignin, synringyl lignin, *p*-hydroxyphenyl lignin.

Several phenylpropanoid and flavonoid biosynthetic pathways genes found in the mango transcriptome dataset were methyltransferases. Mining of mango sequence data revealed >200 unigenes corresponding to different classes of methyltransferases. The methyltransferases transfer a methyl group from a donor to an acceptor. In plants, methylation occurs on nucleotide bases in DNA and amino acids in proteins (Lam et al. 2007). This process is involved in many cellular functions including epigenetic modification and gene regulation. Methyltransferase are also involved in methylation of plant secondary metabolites (such as salicyclic acid) that are important contributors to taste and aroma of many fruits and flowers (Tieman et al. 2010). Identification of >200 methyltransferases in mango provided a wealth of data related to this important class of enzymes. The high number of unigenes encoding methyltransferases indicated variability in structures and functions which in turn reflected the diversity in epigenetic mechanisms and secondary metabolites turnover in mango.

Moreover, mango transcriptome dataset contained 12 enzymes required for the synthesis of following flavonoids (including flavones and flavonols). Pinostrobin, pinobanksin, buetin, dihydrofisetin (futin), naringenin, luteolin, afzelechin, epiafzelechin, catechin, epicatechin, homoeriodictyol, eriodictyol, quercetin and garbanzol (Table 6). Several of these flavonoids are known to exhibit antioxidant, anti-inflammatory, antimutagenic etc. properties (see Table 6 for references). For instance, butein is proposed in

Table 4 Summary of BLASTN searches of mango coding region sequences (n = 24,642) with the coding sequences of seven Viridiplantae species

Name of plant species	*Citrus sinensis* (sweet orange)	*Populus trichocarpa* (black cottonwood)	*Vitis vinifera* (Grape-vine)	*Ricinus communis* (Castor beans)	*Glycine max* (soya bean)	*Medicago truncatula* (Barrel Medic)	*Arabidopsis thaliana*
Source	Citrus Genome DB (www.citrusgenomedb.org)	Plant Genome DB (www.plantgdb.org)	www.genoscope.fr)	J. Craig Ventor Institute (castorbean.jcvi.org/index.php)	Plant Genome DB (www.plantgdb.org)	J. Craig Ventor Institute, Castor (castorbean.jcvi.org/index.php)	Plant Genome DB (www.plantgdb.org)
No. of transcript/coding region sequences	46,147	45,033	26,346	31,225	55,787	62,319	41,671
No. of mango coding sequences matched (%)	9,141 (37.0)	5,568 (22.5)	4,529 (18.3)	4,427 (17.9)	3,634 (17.3)	2,606 (10.5)	1,663 (6.7)

the treatment of breast cancer due to its ability to inhibit aromatase in the human body (Wang 2005). Interestingly, two compounds in this list (i.e. homoeriodictyol, eriodictyol) are bitter-masking flavonones (Ley et al. 2005).

The present dataset contained sequences encoding enzymes for biosynthesis of precursors in Isoflavonoid biosynthetic pathway (i.e. liquiritigenin and naringenin), Flavone and flavonol biosynthetic pathway (i.e. apigenin and kaemferol) and Anthocyanin biosynthetic pathway (i.e. cyanidin).

A number of terpene and benzenoid metabolism related genes including geranylgeranyl pyrophosphate synthase, Farnesyl pyrophosphate synthase and isochorismate hydrolase were also found in mango unigenes. Terpenoids and benzenoids are important natural products in mango. These genes are reported to involve in fruit ripening process (Pandit et al. 2010; Kulkarni et al. 2013). However, they are not limited to tissues of fruit ripening and transcribe in leaf tissues as well. Since most these genes are comprised of large families, most likely different genes would be expressed in leaf and fruit tissues.

Twenty-two enzyme sequences involved in terpenoid backbone biosynthetic pathways were present in mango unigenes. Mevalonate pathway for the synthesis of Isopentenyl-pyrophosphate was found active in mango. Moreover, enzymes required for syntheses of other key intermediates i.e. geranyl-pyrophosphate, farnesyl-pyrophosphate and geranyl-geranyl-pyrophosphate were also present. The farnesal biosynthetic pathway from farnesyl-pyrophosphate was also found active as all required enzymes were present. Moreover, enzymes for the synthesis of dehydrodelichol-pyrophosphate from farnesyl-pyrophosphate; phytyl-pyrophosphate and nona-prenyl-pyrophosphate from greanyl-geranyl-pyrophosphate were also present. Among several monoterpenoid biosynthetic pathways, enzyme for production of Linalool from geranyl-pyrophosphate was found. As a terpene alcohol, Linalool has a pleasant scent and found in many flowering and spice plants (Lewinshon et al. 2001). Triterpenoid biosynthetic pathway from farneyl-pyrophosphate was found active in mango. However, enzymes needed for sesqui terpenoid biosynthesis was not detected in present dataset.

Plant hormone signal transduction pathways in mango

The subsystem based gene annotation identified active presence of a number of plant hormone signal transduction pathways in the present dataset. These plant hormones included auxin, cytokinin, gibberillin, abscisic acid, ethylene, brassinosteroid, jasmonic acid, salicylic acid. Genes encoding receptors, enzymes and others proteins of these signaling pathways and found in the present mango dataset are given in Table 7.

Table 5 Gene ontology (GO) classification of mango transcriptome unigenes

No.	Functional class	Unigenes	No.	Functional class	Unigenes	No.	Functional class	Unigenes
GO ontology: biological_process			*GO ontology: cellular_component*			*GO ontology: molecular_function*		
01	Biological adhesion	195	26	Cell	17,541	43	Antioxidant activity	109
02	Biological regulation	5,891	27	Cell junction	1,183	44	Binding	10,722
03	Carbon utilization	08	28	Cell part	17,541	45	Catalytic activity	10,009
04	Cell killing	01	29	Extracellular matrix	19	46	Electron carrier activity	353
05	Cell proliferation	206	30	Extracellular matrix part	03	47	Enzyme regulator activity	227
06	Cellular component organization or biogenesis	4,265	31	Extracellular region	1,056	48	Metallochaperone activity	03
07	Cellular process	14,285	33	Extracellular region part	20	49	Molecular transducer activity	329
08	Death	592	34	Macromolecular complex	2,844	50	Nucleic acid binding transcription factor activity	670
09	Developmental process	4,286	34	Membrane	7,249	51	Nutrient reservoir activity	13
10	Establishment of localization	4,334	35	Membrane part	2,698	52	Protein binding transcription factor activity	102
11	Growth	1,053	36	Membrane-enclosed lumen	1,076	53	Protein tag	04
12	Immune system process	1,090	37	Nucleoid	38	54	Receptor activity	142
13	Localization	4,561	38	Organelle	14,293	55	Structural molecule activity	581
14	Locomotion	31	39	Organelle part	4,393	56	Translation regulator activity	07
15	Metabolic process	13,425	40	Symplast	1,175	57	Transporter activity	1,631
16	Multi-organism process	2,054	41	Virion	03			
17	Multicellular organismal proc.	4,285	42	Virion part	03			
18	−ve regulation of biological proc.	1,445						
19	+ve regulation of biological proc.	1,276						
20	Regulation of biological process	5,403						
21	Reproduction	2,552						
22	Reproductive process	2,548						
23	Response to stimulus	7,526						
24	Signaling	2,265						
25	Single-organism process	5,756						

From work in model plants, it is known that the ethylene receptors (ER) are negative modulators of the plant hormone ethylene, and therefore likely to play an important part in plant cell physiology. In Arabidopsis, ER is perceived by a family of 05 receptors, divided into two subfamilies (Bleecker et al. 1998). The type-I subfamily include ETR1 and ERS1 and the type-II subfamily receptors include ETR2, ERS2 and EIN4. Mining of mango transcriptome dataset identified a total of five receptor genes. Multiple alignment and phylogenetic comparisons of mango ER sequences with representative ER sequences from other fruit species identified two ETR1 genes (CL3814 and CL5014), one ETR2 gene (CL3814), ETR2-like gene (Unigene15424) and EIN4 gene (Unigene2479) each (Fig. 5). Hence the analysis showed that both ER subfamily members were present in mango. This study demonstrates the potential for the whole genome sequence to be used as a resource for characterization of large multi-gene families in mango.

Proteolytic enzymes in mango

Proteases or peptidases are involved in myriad important cellular functions. Few studies have been reported on proteolytic enzymes in mango (Mehrnoush et al. 2012). The current mango transcriptome analysis revealed 235 unigenes (0.8 % of all unigenes) corresponding to proteases. These unigenes were grouped in five catalytic classes of peptidases/proteases. Number of unigenes related to different classes of peptidases were as follows; serine peptidases (n = 89), metallo peptidases (n = 72), cysteine peptidases (n = 25), aspartic peptidases (n = 40) and threonine peptidases (n = 09) (Table 8). This list contains several plant-specific proteases including Xylem serine protease,

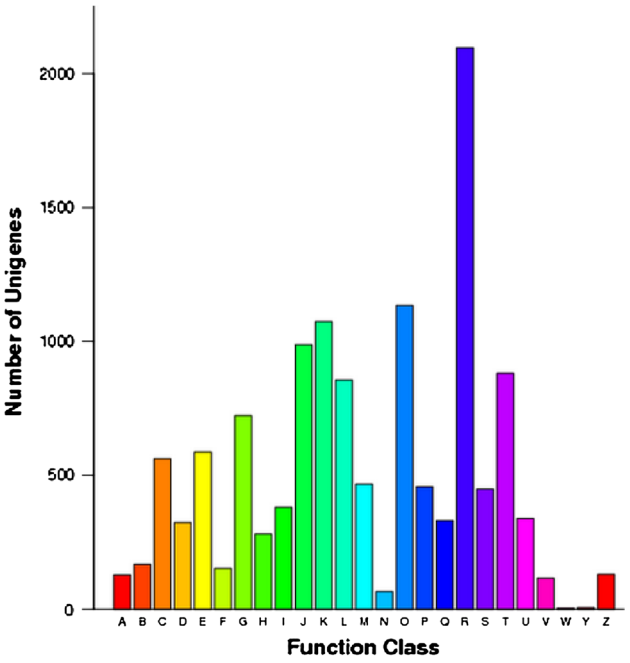

Fig. 4 COG functional classification of mango Unigene sequences. *A* RNA processing and modification; *B* Chromatin structure and dynamics; *C* Energy production and conversion; *D* Cell cycle control, cell division, chromosome partitioning; *E* Amino acid transport and metabolism; *F* Nucleotide transport and metabolism; *G* Carbohydrate transport and metabolism; *H* Coenzyme transport and metalbolism; *I* Lipid transport and metabolism; *J* Translation, ribosomal structure and biogenesis; *K* Transcription; *L* Replication, recombination and repair; *M* Cell wall/membrane/envelope biogenesis; *N* Cell motility; *O* Post-translational modification, protein turnover, chaperones; *P* Inorganic ion transport and metabolism; *Q* Secondary metabolites biosynthesis, transport and catabolism; *R* General function prediction only; *S* Function unknown; *T* Signal transduction mechanisms; *U* Intracellular trafficking, secretion, and vesicular transport; *V* Defense mechanisms; W Extracellular structures; *Y* Nuclear structure; *Z* Cytoskeleton

thalakoidal processing peptidase, Chloroplast processing peptidase, and Germination-specific cysteine protease. Highest number of unigenes for a specific protease was estimated for ATP dependent protease ClpAP (n = 36). The ClpAP is a two component system energy dependent protease system composed of ClpP, the peptidase and ClpA, the ATPase. ClpAP is involved in intracellular proteolysis. We found 25 unigene sequences of ClpP (both cytosolic and chloroplast) which indicated presence of several isoforms of ClpP probably due to alternative splicing. The ClpP isoforms would be responsible for degradation of different proteins due to variation in substrate specificity as a result of sequence variation at the substrate binding cleft of ClpP. Mango contained several unigenes related to papain-like cysteine proteases. These include actinidin homologues with >65 % sequence similarity (CL3299 and Unigene4699) and cathepsins L/H homologues with 40 % sequence similarity (CL3398, CL3842 and CL4516).

Stress response genes

Several stress response genes identified in mango unigene sequences included metallothionein (04 unigenes), Ubiquitin-protein ligase (03 unigenes), cysteine proteinase inhibitor (03 unigenes), FtsJ-like methyltransferase (several unigenes), 14-3-3 protein (12 unigenes), small heat shock protein (04 unigenes) and chitinase (13 unigenes). Two out of four metalothionin (MTH) sequences (i.e. Unigene10572 and Unigene12274) were classified as plant MTH type 1. Sequence motif analysis showed that Unigene12272 codes for MTH type-2 and Unigene10535 codes for MTH type-3 protein. Therefore present data provided evidence of presence of type-1, -2 and -3 MTH in mango.

Table 6 Analysis of mango transcriptome sequences revealed active biosynthetic pathways for bioactive flavonoids mentioned in the table

No.	Name of flavonoid compounds	Known bioactivities
1	Pinostrobin, Pinobanksin	Antioxidant flavonoids that inhibit peroxidation of low density lipoprotein (Fahey and Stephenson 2002)
2	Homoeriodictyol, eriodictyol	Bitter-masking flavanones (Ley et al. 2005)
3	Afzelechin	A flavan-3ol, a type of flavonoids, found in *Bergenia ligulata* (aka Paashaanbhed in Ayurveda traditional Indian medicine) (en.wikidepia.org/wiki/afzelechin)
4	Garbanzol	Antimutagenic flavonoid (Park et al. 2004)
5	Butein	A chalconoid with antioxidative, aldose reductase and advanced glycation endproducts inhibitory effects. (Lee et al. 2008; Wang 2005)
6	Catechin, epicatechain, gallocatechin	Flavan-3-ol compounds with antioxidant and number of other health benefits (en.wikidepia.org/wiki/catechin)
7	Dihydrofisetin (also known as fustin)	A flavanonol, a type of flavonoid; showed protective effects on 6-hydroxydopamine-induced neuronal cell death (Park et al. 2007)
8	Quercetin	Antioxidant (Edwards et al. 2007)
9	Naringenin	A flavanone, a type of flavonoid. It has antioxidant, free radical scavenging, anti-inflammatory, carbohydrate metabolism promoting, and immune system modulating activities (Mulvihill et al. 2009)
10	Luteolin	A flavones with antioxidant and anti-inflammatory activities (López-Lázaro 2009)

Table 7 List of genes involved in plant hormone signaling found in mango mango transcriptome dataset

No.	Plant hormone	Plant hormone signal transduction pathways genes found in mango transcriptome dataset
1	Auxin	auxin influx carrier (AUX1 LAX family), auxin response factor, auxin responsive GH3 gene family, SAUR family protein
2	Cytokinin	histidine-containing phosphotransfer protein, two-component response regulator ARR-A family protein, two-component response regulator ARR-B family protein
3	Gibberillin	gibberellin receptor GID1, DELLA protein
4	Abscisic acid	abscisic acid receptor PYR/PYL family, protein phosphatase 2C [EC:3.1.3.16], serine/threonine-protein kinase SRK2 [EC:2.7.11.1], ABA responsive element binding factor
5	Ethylene	ethylene receptor, serine/threonine-protein kinase CTR1 [2.7.11.1], mitogen-activated protein kinase 6 [2.7.11.24], ethylene-insensitive protein 2, EIN3-binding F-box protein, ethylene-insensitive protein 3
6	Brassinosteroid	protein brassinosteroid insensitive 1, BR-signaling kinase [2.7.11.1], brassinosteroid resistant 1/2
7	Jasmonic acid	jasmonic acid-amino synthetase, coronatine-insensitive protein 1, jasmonate ZIM domain-containing protein, transcription factor MYC2
8	Salicylic acid	regulatory protein NPR1, transcription factor TGA, pathogenesis-related protein 1

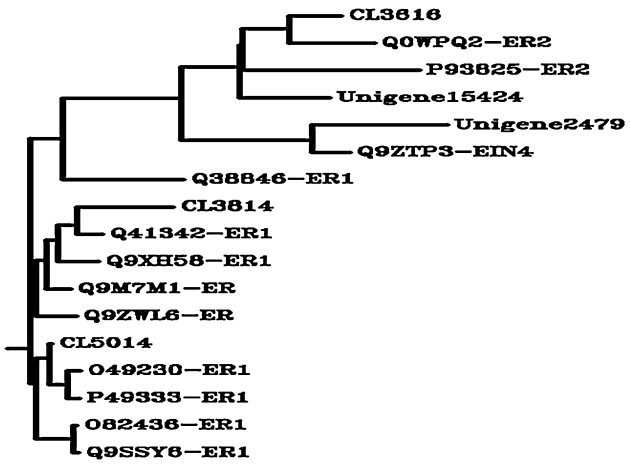

Fig. 5 Phylogenetic tree generated based on multiple alignment of 12 representative ethylene receptor (ER) sequences and five mango ER sequences found in present RNA-seq dataset

Mango chloroplast genome

We carried out chloroplast genome sequencing of mango using Sanger-based and next-generation sequencing methods. Chloroplast genome contains a pair of inverted repeat (IR) regions separated by small and large single copy regions (SSC and LSC). Initially, 22,918 bp of the inverted repeat (IR) region were sequenced using the ASAP protocol (Dhingra and Folta 2005). For this, primers reported by Dhingra and Folta (2005) were used. Additionally, gap filling primers was designed. Consequently, complete IR region of mango cpDNA was sequenced with size of 27,093 bp. Furthermore, 5,783 bp of LSC region was also sequenced using same strategy. Therefore, collectively 32,876 bp of complete IR region and partial LSC region of mango chloroplast genome was obtained using Sanger-based sequencing.

To get more sequence coverage, the mango cpDNA was subjected to pyrosequencing based 454 technology adopted in GS FLX genome sequencer. The raw data of GS FLX was analyzed to get finished sequences. GS FLX sequencer generated sequence data of 10.573 megabases containing 26,988 reads with average length of 400 nucleotides. This data along with the 32,876 bp sequence obtained from Sanger sequencing was used to generate a circular map of mango chloroplast genome containing 151,173 base pairs. However, a complete cpDNA

Table 8 Proteolytic enzymes found in mango transcriptome dataset

No.	Protease	Unigenes	No.	Protease	Unigenes
Metallo-peptidases			*Serine peptidases*		
1	Zn metallopeptidase (endoproteinase)	12	19	ATP dependent protease Clp; proteolytic subunit (18), proteolytic subunit (chloroplast) (7), ATPase subunit (5)	36
2	FTSH4 (2), FTSH9 (3), FTSH2 (1), FTSH6 (2), FTSH10 (2), FTSH (Chloroplast) (1)	11	20	ATP-dependent protease La (Lon)	9
3	Methionine aminopeptidase (6) Methionine aminopeptidase 1A-like (2)	8	21	Serine carboxypeptidase (6), carboxypeptidase type III (2),	8
4	Oligopeptidase A	7	22	Subtilisin like	5
5	Xaa-Pro aminopeptidase	6	23	Protease Do-like 7-like (2), Protease Do-like 2-like (1), Protease Do-like 9-like (1)	4
6	Mitochondrial processing peptidase	6	24	Thalakoidal processing peptidase 2	4
7	Aminopeptidase (2), Aminopeptidase N-like (2)	4	25	Xylem serine proteinase1	3
8	Carboxypeptidase A2-like	3	26	Proline iminopeptidase	3
9	Puromycin-sensitive aminopeptidase	3	27	Mitochondrial inner membrane protease	3
10	Glutamate pro-X carboxypeptidase 2-like	3	28	Site-1 protease	2
11	Caax prenyl protease ste24	2	29	Serine endopeptidase depp2	2
12	Zinc metaloprotesae SLR1821-like isoform 2	1	30	Prolyl oligopeptidase like	2
13	Protease ecfe (RseP peptidase)	1	31	Dipeptidyl peptidase 8-like	2
14	Zinc protease PQQL-like protease	1	32	Ubiquitin-specific protease 21(ESD4-like) serine protease HtrA 2	1
15	Membrane-bound transcription factor protease Site-2	1	33	Serine endopeptidase II-2	1
16	Endoplasmic reticulum metallopeptidase1	1	34	Leucine endopeptidase	1
17	Aspartyl aminopeptidase-like	1	35	Lysosomal pro-X carboxypeptidase	1
18	Chloroplast processing peptidase	1	36	Glutamyl endopeptidase (choloplastic)	1
			37	Protease 4-like	1
Cysteine proteases			*Aspartic proteases*		
38	Cysteine peptidase-like	7	49	Aspartic proteinase-nepenthesin-1 (7), aspartic proteinase-nepenthesin-2 (4)	11
39	Cysteine protease (5), Cysteine protease like (3), Cysteine protease ATG4B (2),	5	50	Aspartic proteinase-like protein2-like	7
40	Sentrin/sumo-specific protease (2), sentrin-specific protease (1)	2	51	Signal peptide peptidase-like 2B	6
41	UFM1-specific protease-like	2	52	D-Alanyl-D-alanine endopeptidase	6
42	Cysteine protease	2			
43	pyrrolidone-carboxylate peptidase	2	53	Aspartic proteinase (2), Aspartic proteinase 1 (1), Aspartic proteinase 2 (1)	4
44	Germination-specific cysteine protease 1-like	1	54	Aspartic proteinase-Asp1	3
45	Cysteine proteinase 15A-like	1	55	Aspartic Proteinase-like protein1-like	3
46	Legumain-like proteinase	1	*Threonine proteases*		
47	PPPDE peptidase domain-containing protein2 like	1	56	26S proteasome regulatory subunits 6A, S10B, 6B, 7	7
48	Asparaginyl endopeptidase	1	57	Isoarpartyl peptidase/L-aspargenase2	1
			58	Gamma-glutamyltranspeptidase1	1

sequence could not be obtained as the 151,173 bp mango cpDNA sequence contained 30 gaps when compared to Citrus cpDNA (17 gaps located in intergenic homopolymer repeat regions whereas 13 intragenic gaps). Preliminary sequence comparison revealed close relationship of mango and *C. sinensis* cpDNA sequences. The chloroplast genome size of *C. sinensis* is 160,129 base pairs. This comparison indicated that 95 % of mango chloroplast genome sequences have been obtained resulting from the current study. The draft sequence of mango chloroplast genome was submitted in GenBank with accession number FJ212316.

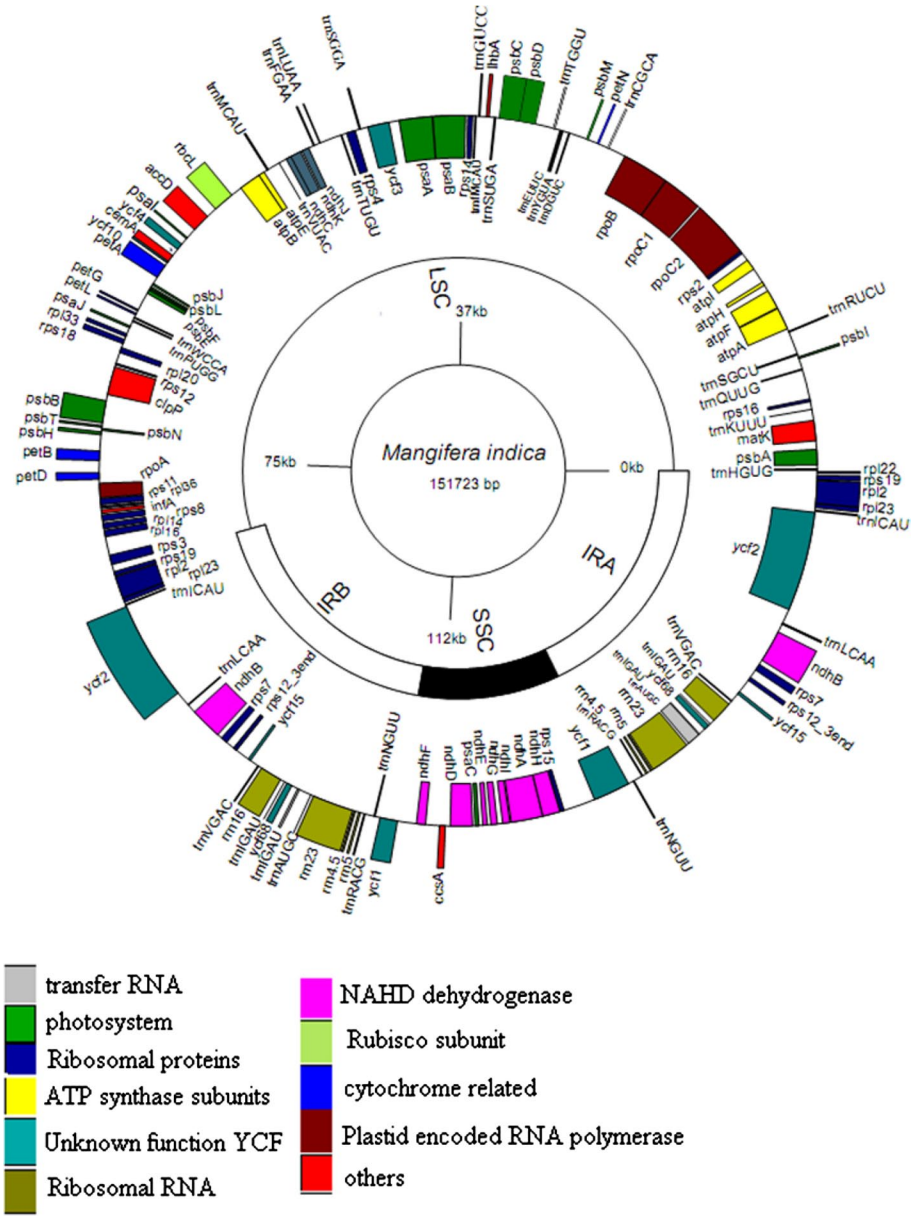

Fig. 6 Circular map of mango chloroplast DNA indicating LSC, SSC and IR regions and the color scheme for gene indication. The map has been constructed by GenomeVx at http://wolfe.gen.tcd.ie/GenomeVx/ (Conant and Wolfe 2008)

The available sequence of mango cpDNA (151,173 bp; GC 38.18 %) contains a pair of inverted repeats (IRA and IRB) of 27,093 bp separated by small and large single copy (SSC, LSC) regions, respectively (Fig. 6). The inverted repeat of mango chloroplast is larger as compared to several previously reported angiosperms for example length of cpDNA inverted repeat of *A. thaliana* is 26,264 bp (Sato et al. 1999), *Solanum tuberosum* is 25,595 bp (Chung et al. 2006), *Nicotiana tabacum* is 25,339 bp (Shinozaki et al. 1986) and *Gossypium barbadense* is 25,591 bp (Ibrahim et al. 2006). The length of mango cpDNA IR region is only 97 base pairs larger compared to its closest neighbour *C. sinensis* which has 26,996 bp (Bausher et al. 2006). This finding support the previous observations that contraction and expansion of IR region of chloroplast DNA is a major source of size variation of cpDNA among angiosperms (Chung, et al. 2006).

A total of 139 genes were detected in mango cpDNA sequence (119 single copy genes while 20 duplicated genes in inverted repeat regions). 91 genes code for proteins, including nine duplicated genes in the inverted repeats. There were four rRNA genes and 29 distinct tRNAs, 7 of which are duplicated in the inverted repeat. Notably, *M. indica* cpDNA contains the *infA* gene which code for a translation initiation factor and not present in its closest neighbour *Citrus* genome (Bausher et al. 2006). The detailed list of gene contents is present in Table 9.

Chloroplast genome-wide comparative analysis was carried by multiple alignment of cp DNA sequence from mango and 16 representatives of land plants and 2

Table 9 Genes encoded by mango chloroplast DNA

transfer RNA	trnH-GUG, trnK-UUU, trnQ-UUG, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnM-CAU, trnS-UGA, trnG-UCC, trnfM-CAU, trnS-GGA, trnT-UGU, trnL-UAA, trnF-GAA, trnV-UAC, trnM-CAU, trnT-GGU, trnW-CCA, trnP-UGG, trnI-CAU, trnL-CAA, trnV-GAC, trnI-GAU, trnI-GAU, trnA-UGC, trnR-ACG, trnA-UGC, trnI-GAU, trnI-GAU, trnV-GAC, trnL-CAA, trnL-UAG[a], trnI-CAU
Photosystem I	*psa* A[a], *psa* B, *psa* C, *psa*J, *psa*I
Photosystem II	*psb* A, *psb* B, *psb* C, *psb* D, *psb* E, *psb* F,*psb* H,*psb* I,*psb* J, *psb* L, *psb* M, *psb* N, *psb* T
Assembly stability of photosystem I	Ycf3[a], ycf4
Ribosomal proteins	*rps* 2, *rps* 3, *rps* 4, *rps* 7, *rps* 8, *rps* 11, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*
ATP synthase subunits	atpA,atpB,atpE, atpF, atpH, atpI
Unknown function YCF	Ycf1[a], ycf2, ycf15
Ribosomal RNA	rrn23, rrn16, rrn5, rrn4.5
NAHD dehydrogenase	ndhA[a], ndhB, ndhC, ndhD, ndhE, ndhF, ndhG[a], ndhH,ndhI, ndhJ, ndhK
Rubisco subunit	rbcL
cytochrome related	PetA, PetB[a], PetD, PetG, PetL, PetN,ccsA
Plastid encoded RNA polymerase	*rpo* A[a],*rpo* B,*rpo* C1[a], *rpo* C2
Maturase	*matK*[a]
Acetyl-CoA carboxylase subunit	*accD*
ATP dependnt protease subunit	*clpP*
Gene for inorganic carbon uptake	*cemA*[a]

[a] Partial gene sequences due to gaps in respective regions of mango chloroplast genome

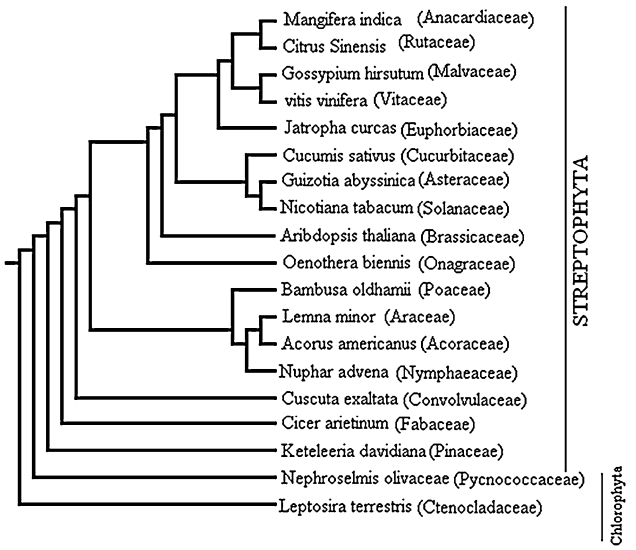

Fig. 7 Phylogenetic tree of cpDNA from 19 plant species (including mango) using VISTA comparative genomics server (http://genome.lbl.gov/vista/mvista/submit.shtml)

representatives of Chlorophyta using VISTA web server (Asif et al. 2013; Khan et al. 2012; Frazer et al. 2004). The phylogenetic tree also indicated grouping of mango cpDNA sequences with *C. sinensis* (citrus sp.), *Gossypium hirsutum* (cotton) and *V. vinifera* (red grape). However, the most closely related sequence is *C. sinensis* cpDNA (Fig. 7).

*Repeats in **M. indica** chloroplast genome*

Along with two large inverted repeats in chloroplast genome sequences i.e. IRA and IRB, a large number of relatively small repeats have been recently observed (Haberle et al. 2008). In mango chloroplast genome 51 direct repeats (Supplementary data) could be found using RePuter server http://bibiserv.techfak.uni-ielefeld.de/reputer/submission.html (Kurtz et al. 2001). The RePuter program provides software solutions to compute and visualize repeats in whole genomes or chromosomes. It provides interactive as well as static images of the results. Figure 8 indicates the position of repeats along with their sizes and orientation. RePuter analysis of mango cpDNA revealed 15 repeats of size >50 bp while rest were less than 50 bp. These repeats occur as the part of genes as well as intergenic spacers. It is interesting to note that presence of hundreds of repeats in *Trachelium caeruleum* chloroplast genome is supposed to be associated with extensive rearrangements (Haberle et al. 2008).

Conclusion

Mango genomic research lags that of other crops of economic importance. To facilitate biochemical and molecular biological research in mango, we characterized mango leaf transcriptome and cpDNA. Most of the resultant unigenes

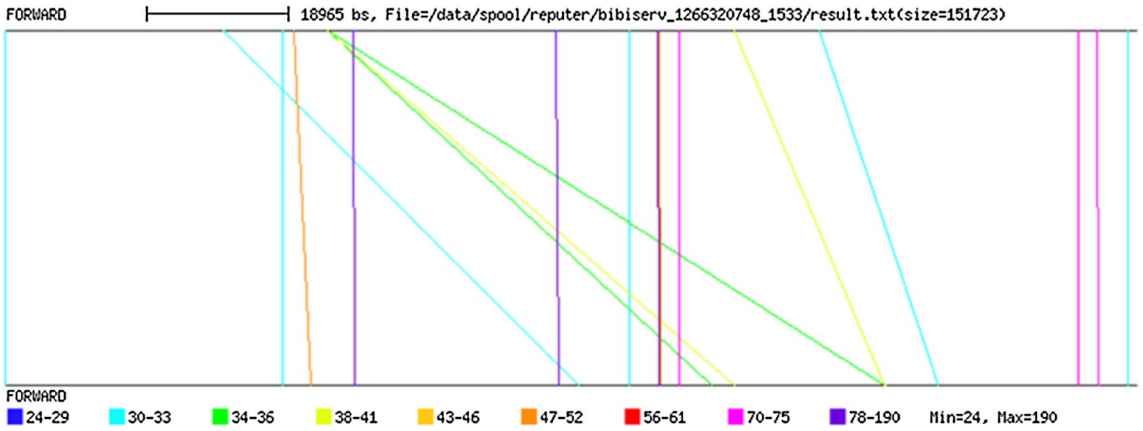

Fig. 8 Diagrammatic representation of short repeats in mango chloroplast genome sequence. The graph outlines the length and location of repeats. The *lines* indicating repeats are *colored* according to length. *Each part* of a repeat is displayed on a separate strand to keep the starting position visible

were aligned with mango sequences deposited in the Genbank, whereas all of coding regions in unigenes were matched with proteins sequences identified in mango proteomic dataset. Subsystem based gene annotation provided information for the production of a number of bioactive flavonoids, carotenoids and terpenoids in mango. These bioactive natural products are known to have a range of health beneficial properties. The large number of unigenes identified in this study provides an important resource for future studies on mango biology. The advancements in transcriptomic, genomic, epigenomic, proteomic resources of non-model plants would greatly facilitate research in plant biology.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013):1651–1656

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andrade Jde M, Toledo TT, Nogueira SB, Cordenunsi BR, Lajolo FM, do Nascimento JR (2012) 2D-DIGE analysis of mango (*Mangifera indica* L.) fruit reveals major proteomic changes associated with ripening. J Proteomics 75:3331–3341

Asif H, Khan A, Iqbal A, Khan IA, Heinze B, Azim MK (2013) The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. Tree Genet Genomes 9:867–877

Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H (2006) The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biol 6:21. doi:10.1186/1471-2229-6-21

Bleecker AB, Esch JJ, Hall AE, Rodríguez FI, Binder BM (1998) The ethylene-receptor family from Arabidopsis: structure and function. Philos Trans R Soc Lond B Biol Sci 353(1374):1405–1412

Chinag YC, Tasi CM, Chen YK, Lee SR, Chen CH, Lin YS, Tasi CC (2012) Development and characterization of 20 new polymorphic microsatellite markers from mangifera indica (Anacardiaceae). Am J Bot 99(3):e117–e119

Chung HJ, Jung JD, Park HW, Kim JH, Cha HW, Min SR, Jeong WJ, Liu JR (2006) The complete chloroplast genome sequences of Solanum tuberosum and comparative analysis of with solanaceae species identified the presence of 241 bp deletion in cultivated potato chloroplast DNA sequence. Plant Cell Rep. doi:10.1007/s0029-006-0196-4

Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24:861–862

Conesa A, Gotz S et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676

Dhingra A, Folta KM (2005) ASAP: amplification, sequencing and annotation of plastomes. BMC Genom 6:176

Duangjit J, Bohanec B, Chan AP, Town CD, Havey MJ (2013) Transcriptome sequencing to produce SNP-based genetic maps of onion. Theor Appl Genet. doi:10.1007/s00122-013-2121-x

Duval M, Bunel FJ, Sitbon C, Risterucci AM (2005) Development of microsatellite markers for mango (*Mangifera indica* L.). Mol Ecol Notes 5:823

Edwards RL, Lyon T, Litwin SE, Rabovsky A, Symons JD, Jalili T (2007) Quercetin reduces blood pressure in hypertensive subjects. J Nutr 137(11):2405–2411

Fahey JW, Stephenson KK (2002) Pinostrobin from honey and Thai ginger (Boesenbergia pandurata): a potent flavonoid inducer of mammalian phase 2 chemoprotective and antioxidant enzymes. J Agric Food Chem 50(25):7472–7476

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–W275

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29(7):644–652

Haas BJ, Zody MC (2010) Advancing RNA-seq analysis. Nat Biotechnol 28(5):421–423

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and trna genes. J Mol Evol 66:350–361

Hirano R, Htun Oo T, Watanabe KN (2010) Myanmar mango landraces reveal genetic uniqueness over common cultivars from Florida, India, and Southeast Asia. Genome 53(4):321

Ibrahim RIH, Azuma JI, Sakamoto M (2006) Complete nucleotide sequence of the cotton (*Gossypium barbadense* L.) chloroplast genome with a comparative analysis of sequences among 9 dicot plants. Genes Genet Syst 81:311–321

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 138–148

Kanehisa M, Araki M et al (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36(Database issue):D480–D484

Khan IA, Azim MK (2011) Variations in intergenic spacer rpl20-rps12 of Mango (*Mangifera indica*) chloroplast DNA: implications in cultivar identification. Plant Evol Syst 292(3–4):249–255

Khan A, Khan IA, Heinze B, Azim MK (2012) The chloroplast genome sequence of date palm (*Phoenix dactylifera* L. cv. 'Aseel'). Plant Mol Biol Rep 30:666–678

Krishna H, Singh SK (2007) Biotechnological advances in mango (*Mangifera indica* L.) and their future implication in crop improvement: a review. Biotechnol Adv 25:223–243

Kulkarni R, Pandit S, Chidley H, Nagel R, Schmidt A, Gershenzon J, Pujari K, Giri A, Gupta V (2013) Characterization of three novel isoprenyl diphosphate synthases from the terpenoid rich mango fruit. Plant Physiol Biochem 71:121–131

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29(22):4633–4642

Lam KC, Ibrahim RK, Behdad B, Dayanandan S (2007) Structure, function, and evolution of plant O-methyltransferases. Genome 50(11):1001–1013

Lee EH, Song DG, Lee JY, Pan CH, Um BH, Jung SH (2008) Inhibitory effect of the compounds isolated from *Rhus verniciflua* on aldose reductase and advanced glycation endproducts. Biol Pharm Bull 31(8):1626–1630

Lewinshon E, Schalechet F, Wilkinson J, Matsui K, Tadmor Y, Nam K, Amar O, Lastochkin E, Larkov O, Ravid U, Hiatt W, Gepstein S, Pichersky E (2001) Enhanced levels of the aroma and flavor compound S-linalool by metabolic engineering of the terpenoid pathway in tomato fruits. Plant Physiol 127:1256–1265

Ley JP, Krammer G, Reinders G, Gatfield IL, Bertram HJ (2005) Evaluation of bitter masking flavanones from Herba Santa (*Eriodictyon californicum* (H. and A.) Torr., Hydrophyllaceae. J Agric Food Chem 53(15):6061–6066

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) RNASeq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493–500

López-Lázaro M (2009) Distribution and biological activities of the flavonoid luteolin. Mini Rev Med Chem 9(1):31–59

Mehrnoush A, Mustafa S, Sarker MZ, Yazid AM (2012) Optimization of serine protease purification from mango (Mangifera indica cv. Chokanan) peel in polyethylene glycol/dextran aqueous two phase system. Int J Mol Sci 13:3636–3649

Michal G (1999) Biochemical pathways, an atlas of biochemistry and molecular biology. Spektrum Akademischer, Heidelberg

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

Mukherjee SK (1950) Mango: its allopolyploid nature. Nature 4213:196–197

Mukherjee SK, Litz RE (2009) Introduction: Botany and Importance. In: Litz RE (ed) The mango botany, production and uses, 2nd edn. CBI International, Wallingford, pp 1–18

Mulvihill EE, Allister EM, Sutherland BG, Telford DE, Sawyez CG, Edwards JY, Markle JM, Hegele RA, Huff MW (2009) Naringenin prevents dyslipidemia, apolipoprotein B overproduction, and hyperinsulinemia in LDL receptor-null mice with diet-induced insulin resistance. Diabetes 58(10):2198–2210

Pandit SS, Kulkarni RS, Giri AP, Kollner TG, Degenhardt J, Gershenzon J, Gupta VS (2010) Expression profiling of various genes during the fruit development and ripening of mango. Plant Physiol Biochem 48:426–433

Park KY, Jung GO, Lee KT, Choi J, Choi MY, Kim GT, Jung HJ, Park HJ (2004) Antimutagenic activity of flavonoids from the heartwood of *Rhus verniciflua*. J Ethnopharmacol 90(1):73–79

Park BC, Lee YS, Park HJ, Kwak MK, Yoo BK, Kim JY, Kim JA (2007) Protective effects of fustin, a flavonoid from Rhus verniciflua Stokes, on 6-hydroxydopamine-induced neuronal cell death. Exp Mol Med 39(3):316–326

Ravishankar KV, Mani BH, Anand L, Dinesh MR (2011) Development of new microsatellite markers from mango (*Mangifera indica*) and cross-species amplification. Am J Bot 98(4): e96–e99

Renuse S, Harsha HC, Kumar P, Acharya PK, Sharma J, Goel R, Kumar GSS, Raju R, Prasad TSK, Slotta T, Pandey A (2012) Proteomic analysis of an unsequenced plant-*Mangifera indica*. J Proteomics 75:5793–5796

Rocha A, Salomao LC, Salomao TM, Cruz CD, de Siqueira DL (2012) Genetic diversity of 'uba' mango tree using ISSR markers. Mol Biotechnol 50(2):108–113

Sara Z, Alberto F, Enrico G, Luciano X, Marianna F, Giovanni M, Diana B, Mario P, Massimo D (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-seq. Plant Physiol 152:1787–1795

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast Genome of *Arabidopsis thaliana*. DNA Res 6:283–290

Schnell RJ, Olano CT, Quintanilla WE, Meerow AW (2005) Isolation and characterization of 15 microsatellite loci from mango (*Mangifera indica* L.) and cross-species amplification in closely related taxa. Mol Ecol Notes 5:625

Schnell RJ, Brown JS, Olano CT, Meerow AW, Campbell RJ, Kuhn DN (2006) Mango genetic diversity analysis and pedigree inferences for Florida cultivars using microsatellite markers. J Am Soc Hort Sci 131:214

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki J, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Souza IG, Valente SE, Britto FB, de Souza VA, Lima PS (2011) RAPD analysis of the genetic diversity of mango (Mangifera indica) germplasm in Brazil. Genet Mol Res 10(4):3080–3089

Srivastava N, Bajpai A, Chandra R, Rajan S, Muthukumar M, Srivastava MK (2012) Comparison of PCR based marker systems for genetic analysis in different cultivars of mango. J Environ Biol 33(2):159–166

Strickler SR, Aureliano Bombarely A, Mueller LA (2012) Designing a transcriptome next-generation sequencing project for a non-model plant species. Am J Bot 99(2):257–266

Tieman D, Zeigler M, Schmelz E, Taylor MG, Rushing S, Jones JB, Klee HJ (2010) Functional analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor volatile methyl salicylate. Plant J 62:113–123

Viruel MA, Escribano P, Barbieri M, Ferri M, Hormaza JI (2005) Fingerprinting, embryo type and geographic differentiation in mango (*Mangifera indica* L., Anacardiaceae) with microsatellites. Mol Breeding 15:383

Wang Y (2005) The plant polyphenol butein inhibits testosterone-induced proliferation in breast cancer cells expressing aromatase. Life Sci 77(1):39–51

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev 10(1):57–63

Wilhelm BT, Marguerat S, Goodhead I, Bahler J (2010) Defining transcribed regions using RNA-seq. Nat Protoc 5(2):255–266

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17):3252–3255

Xu J, Li Y, Ma X, Ding J, Wang K, Wang S, Tian Y, Zhang H, Zhu X-G (2013) Whole transcriptome analysis using next-generation sequencing of model species *Setaria viridis* to support C4 photosynthesis research. Plant Mol Biol. doi:10.1007/s11103-013-0025-4

Ye J, Fang L et al (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34(Web Server issue):W293–W297

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1–2):203–214