Plant Mol Biol (2012) 79:413–427
DOI 10.1007/s11103-012-9919-9

De novo transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis

Tae Kyung Hyun · Yeonggil Rim · Hui-Jeong Jang · Cheol Hong Kim · Jongsun Park · Ritesh Kumar · Sunghoon Lee · Byung Chul Kim · Jong Bhak · Binh Nguyen-Quoc · Seon-Won Kim · Sang Yeol Lee · Jae-Yean Kim

Received: 24 February 2012 / Accepted: 29 April 2012 / Published online: 12 May 2012

Abstract The ripe fruit of *Momordica cochinchinensis* Spreng, known as gac, is featured by very high carotenoid content. Although this plant might be a good resource for carotenoid metabolic engineering, so far, the genes involved in the carotenoid metabolic pathways in gac were unidentified due to lack of genomic information in the public database. In order to expedite the process of gene discovery, we have undertaken Illumina deep sequencing of mRNA prepared from aril of gac fruit. From 51,446,670 high-quality reads, we obtained 81,404 assembled unigenes with average length of 388 base pairs. At the protein level, gac aril transcripts showed about 81.5 % similarity with cucumber proteomes. In addition 17,104 unigenes have been assigned to specific metabolic pathways in Kyoto Encyclopedia of Genes and Genomes, and all of known enzymes involved in terpenoid backbones biosynthetic and carotenoid biosynthetic pathways were also identified in our library. To analyze the relationship between putative carotenoid biosynthesis genes and alteration of carotenoid content during fruit ripening, digital gene expression analysis was performed on three different ripening stages of aril. This study has revealed putative phytoene synthase, 15-cis-phytone desaturase, zeta-carotene desaturase, carotenoid isomerase and lycopene epsilon cyclase might be key factors for controlling carotenoid contents during aril ripening. Taken together, this study has also made availability of a large gene database. This unique information for gac gene discovery would be helpful to facilitate functional studies for improving carotenoid quantities.

Keywords Carotenoid · Digital gene expression · *Momordica cochinchinensis* Spreng · Next generation sequencing · Transcriptome

Introduction

Plants produce a wide variety of secondary metabolites which are non-essential to plant growth and development, but they play a very important role in various aspects such as the resistance against pests and diseases, the attraction of pollinators and the interaction with symbiotic microorganisms (Dixon 2001; Harborne 2001; Zhang et al. 2005). Because of the valuable nutritional and pharmacological properties of secondary metabolites in human beings, these plant secondary metabolites have been given great importance recently. Since secondary metabolites are usually produced only at low concentration in plants, the alternative of more efficient metabolic pathways by genetic modification has been suggested as a tool for improving their biosynthetic capacity (Zhang et al. 2005). Advances in our understanding about the secondary metabolite biosynthetic genes combined with high resolution DNA sequences data have created great opportunities to achieve detailed metabolite profiles.

Electronic supplementary material The online version of this article (doi:10.1007/s11103-012-9919-9) contains supplementary material, which is available to authorized users.

T. K. Hyun · Y. Rim · H.-J. Jang · R. Kumar · S.-W. Kim · S. Y. Lee · J.-Y. Kim (✉)
Division of Applied Life Science (BK21-WCU Program), Plant Molecular Biology and Biotechnology Research Center, Gyeongsang National University, Jinju 660-701, Korea
e-mail: kimjy@gnu.ac.kr

C. H. Kim · J. Park · S. Lee · B. C. Kim · J. Bhak
Theragen Bio Institute, Suwon 443-766, Korea

B. Nguyen-Quoc
Biotechnology Center of Hochiminh City, Hochiminh City, Vietnam

Carotenoids, isoprenoid pigments, are secondary metabolites that not only play a significant role in protecting the photosynthetic apparatus from photo-oxidation (Frank and Cogdell 1996), but they also have fundamental roles in human nutrition as metabolic precursors and antioxidants (Fraser and Bramley 2004). Their consumption is associated with many health benefits like the prevention of cancer (Giovannucci 2002), the maintenance of the immune system (Chew and Park 2004) and the protection from a wide range of diseases (Fraser and Bramley 2004; Giuliano et al. 2008). Although fruits and vegetables are good sources of carotenoids, cereal grains are deprived of carotenoid leading to deficiency diseases in the developing countries where cereals are the staple diet (Giuliano et al. 2008). For these reasons, during last decades, many metabolic engineering efforts have been undertaken in modifying cereal crops to improve the carotenoid contents (Giuliano et al. 2008; Zhu et al. 2008). Carotenoid biosynthetic pathway and gene regulation have been studied in various plant species, such as *Arabidopsis thaliana* (Park et al. 2002), *Solanum lycopersicum* (Isaacson et al. 2002; Ronen et al. 1999), *Capsicum annuum* (Romer et al. 1993), *Nicotiana tabaccum* (Busch et al. 2002), *Prunus armeniaca* (Marty et al. 2005), *Crocus sativus* (Castillo et al. 2005), *Daucus carota* (Clotault et al. 2008) and *Tagetes erecta* (Del Villar-Martinez et al. 2005; Moehs et al. 2001).

Momordica cochinchinensis Spreng, commonly known as gac, belongs to the family *Cucurbitaceae*. This fruit is indigenous to Southeast Asia and grown in Vietnam, Thailand, Laos, Cambodia, China, India and other tropical areas (Nantachit and Tuchinda 2009). Gac is a bright-red fruit that grows as large as a cantaloupe and covered with short spines, and has been used as a food and traditional herbal medicine in East and South Asia from many decades (Iwamoto et al. 1985). The seeds and seed membranes of gac have been employed for a variety of internal and external purposes by human beings (Ishida et al. 2004). Gac is esteemed as "the fruit from heaven" (Burke et al. 2005) because of high concentration of carotenoids, especially beta-carotene and lycopene (Ishida et al. 2004; Vuong et al. 2002). Phytochemical analysis of gac aril has shown that its lycopene concentration is much higher than that found in field grown tomatoes and tomato products, which are the major source of lycopene in Western countries (Aoki et al. 2002; Ishida et al. 2004; Kubola and Siriamornpun 2011; Markovic et al. 2006; Vuong et al. 2002). This finding suggests that gac might contain the gene resources for improving lycopene and beta-carotene content using biotechnological applications and metabolic engineering, although genomic information is very limited for gac in the public database.

Since the application of next-generation sequencing (NGS) has proven to be an effective method to analyze functional gene variation in non-model plants (Angeloni et al. 2011; Garg et al. 2011; Sun et al. 2010; Wenping et al. 2011), a combination of de novo assembly of transcriptome using short-read sequencing data and digital gene expression (DGE) analyses has been introduced as a powerful method for indentifying candidate genes encoding biosynthetic enzymes of secondary metabolites in a non-model plants (Tang et al. 2011; Yang et al. 2011). Thus, we employed the RNA-Seq method using the Illumina HiSeq^{TM} 2000 to generate large amount of sequenced transcriptome of gac aril. We have generated 51,466,670 sequence reads that were assembled into 81,404 unigenes. Comprehensive analysis indicated that most of the known genes encoding enzymes involved in the biosynthesis of terpenoid C5-backbone (C5-unit isopentenyl diphosphate, IPP) and carotenoid, existed in the transcriptome library of gac aril. To further validate the findings, expression analysis of each gene in different ripening stages of aril was carried out using DGE analysis. The transcriptome sequences along with DGE profile provide the first report of discovering functional genes in the gac aril transcriptome. In future, this study would be of great importance by providing the gene resources for improving carotenoid content in other plants.

Materials and methods

Sample collection and preparation

The fully expanded unripe fruit collected from the plantation in Hochiminh City, Vietnam was used for deep RNA sequencing. For transcriptome sequencing, the arils from 45 days after flowering (DAF) *M. cochinchinensis* Spreng were used, whereas the arils from 35 DAF, 45 DAF and 55 DAF from fruit (see Table 1) were collected for DGE analysis. The arils from each sample were immediately frozen in liquid nitrogen until further processing. For measurement of carotenoids, the arils from each sample were then cut into small pieces (50 mg) and were stored in 1.5 ml of ethanol containing 0.1 % ascorbic acid (w/v).

Table 1 Description of fruit sample growth stages used in this study

Sample	Description
Stage 1	Fruit age between 30 and 40 DAF, green peel, white pulp and orange aril
Stage 2	Fruit age between 41 and 50 DAF, green peel, white pulp and red aril
Stage 3	Fruit age between 51 and 60 DAF, green peel, light orange pulp and red aril

RNA isolation and Illumina sequencing

Total RNA was isolated using RNEasy Plant Mini kit (Qiagen, USA). To avoid genomic DNA contamination RNA was treated with RNase-free DNase I (Promega, USA). Total RNA content, purity and degradation were assessed by Nanodrop 2000C spectrophotometer (Thermo scientific, USA) and the quality of RNA was confirmed by agarose gel electrophoresis before proceeding. For transcriptome analysis, mRNA from total RNA was enriched by using the oligo(dT) magnetic beads, fragmented to short pieces (about 200 bp). Then the first strand cDNA was synthesized by random hexamer-primer using the mRNA fragments. After synthesizing second strand cDNA using DNA polymerase I and RNase H, cDNA fragments went through an end repair process and ligation of sequencing adapters. These products were purified by agarose gel electrophoresis and enriched by PCR to create the final cDNA library. The cDNA library was sequenced on the Illumina HiSeq^{TM}2000 sequencing platform.

De novo assembly

The raw reads were cleaned by removing 3′adaptor sequences, empty reads, low quality reads (with ambiguous sequences '*N*') and the reads with more than 10 % Q < 20 bases. Transcriptome de novo assembly was carried out with short reads assembling program SOAPdenovo (Li et al. 2010). SOAPdenovo firstly combined reads with certain length of overlap to form longer fragments without N, called contigs. Then the reads were mapped back to contigs; with paired-end reads it was able to detect contigs from the same transcript as well as the distances between these contigs. SOAPdenovo connected the contigs using N to represent unknown sequences between each two contigs, and then Scaffolds were made. Paired-end reads were used again for gap filling of scaffolds to get sequences with least Ns and cannot be extended on either end. Such sequences are defined as unigenes. Finally, blastx alignment (E value $< 10^{-5}$) between unigenes and protein databases like NCBI non-redundant protein (NR) database (http://www.ncbi.nlm.nih.gov), Swiss-Prot protein database (http://www.expasy.ch/sprot), the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (http://www.genome.jp/kegg) and Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG) was performed, and the best aligning results were used to decide sequence direction of unigenes. In any case, if results of different databases conflicted with each other, a priority order of NR, Swiss-Prot, KEGG and COG should be followed while deciding sequence direction of unigenes. When a unigene was found to be unaligned to none of the above databases, ESTScan (Iseli et al. 1999) was used to predict its coding regions as well as to decide its sequence direction.

Construction of phylogenomic tree and similarity analysis

To draw phylogenomic tree with gac tracriptome sequences, we used 20 plant genomes, *Medicago truncatula*, *Lotus japonicus*, *Glycine max*, *Cucumis sativus*, *Vitis vinifera*, *Prunus persica*, *Carica papaya*, *Populus trichocarpa*, *Ricinus communis*, *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Lycopersicon esculentum*, *Mimulus guttatus*, *Aquilegia coerulea*, *Orayza sativa japonica*, *Zea mays*, *Setaria italic*, *Sorghum bicolor*, *Physcomitrella patens* and *Selaginella moellendorffii*. Based on 793,704 proteins including Gac transcriptome sequences, we clustered them using Tribe-MCL resulting in 119,626 clusters. Among them 2,926 clusters which contains proteins from all 21 sources were selected. Using ClustalW 2.1, each set of proteins were aligned and merged into one file. We drew Neighbor joining phylogenetic tree with 5,000 bootstrap repeats using MEGA5 (Tamura et al. 2011). To analyze similarity between gac tracriptome and other plants, the proteome data sets for Arabidopsis, rice, cucumber, grape, maize, medicago, poplar, tomato, soybean, papaya, sorghum and physcomitrella were downloaded from their respective genome project websites. The gac transcripts were searched against proteome sequences using BLASTX program.

Functional annotation of unigenes

For the functional annotation, the remaining sequences that might putatively encode proteins were searched against NR, Swiss-Prot, KEGG and COG using the BLASTX algorithm. A typical cutoff value of $E < 10^{-5}$ was used. With Nr annotation, Blast2GO program (Conesa et al. 2005) was used to get GO annotation of unigenes according to component function, biological process and cellular component ontologies. After getting GO annotation for every unigene, we used WEGO software (Ye et al. 2006) to do GO functional classification for all unigenes and to understand the distribution of gene functions of the species from the macro level.

mRNA tag (digital gene expression, DGE) library preparation and sequencing

Tag library preparation for the three different ripening stage samples (35 DAF, 45 DAF and 55 DAF) was performed using Illumina gene expression sample preparation kit. mRNA from total RNA was purified using oligo(dT) magnetic beads, and then oligo(dT) was used as primer to

synthesize the first and second-strand cDNA. The bead-bound cDNA was subsequently digested with restriction enzyme NlaIII. The cDNA fragments with 3′ ends were purified with magnetic beads and the Illumina adaptor 1 was ligated to the 5′ end of the digested bead-bound cDNA fragments. The junction of Illumina adaptor 1 and CATG site is the recognition site of MmeI, which is a type of endonuclease and cuts 17 bp downstream of CATG site, producing tags with adaptor 1. After removing 3′ fragment with magnetic beads precipitation, Illumina adaptor 2 was ligated to the 3′ ends of tag, acquiring tags with different adapters of both ends to form a tag library. After 15 cycles of linear PCR amplification, 95 bp fragments were purified by PAGE gel electrophoresis. After denaturation, single-chain molecules were fixed onto the Illumina sequencing chip for sequencing. The reproducibility of DGE was >0.99.

Analyzing DGE tags and evaluation of DGE library

For DGE analysis, we initially filtered all sequences to remove adaptor sequence, empty reads (reads with only 3′ adaptor sequences but no tags), low quality sequences (tags with unknown sequences '*N*') and tags with a copy number of 1 (probably sequencing error) as described previously by Hao et al. (2011). All clean tags were mapped to the reference sequences, and DGE counts were normalized using the RPKM (reads/Kb/Million) method (Mortazavi et al. 2008; Nagalakshmi et al. 2008). The formula is shown as follows:

$$\mathrm{RPKM} = \frac{10^5 C}{NL/10^3}$$

Total number of reads that uniquely aligned to all genes is *N*, number of bases on gene A is *L*, and number of reads that uniquely aligned to gene A is *C*.

A statistical analysis of frequency of each tag in the different DGE libraries was preformed to identify differentially expressed genes between each different ripening stage samples. Statistical comparison was performed using the method described by Audic and Claverie (1997). The number of unambiguous clean tags from gene A was denoted as *x*, given every gene's expression occupies only a small part of the library, $p(x)$ closely follows the Poisson distribution.

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \,(\lambda \text{ is the real transcript of the genes})$$

The total clean tag number of the sample 1 is N_1, and total clean tag number of sample 2 is N_2; gene A holds *x* tags in sample 1 and *y* tags in sample 2. The probability of gene A expressed equally between two samples was calculated with:

$$2\sum_{i=0}^{i=y} p(i|x) \text{ or } 2\left(1-\sum_{i=0}^{i=y} p(i|x)\right)\left(\text{if } \sum_{i=0}^{i=y} p(i|x) > 0.5\right)$$

$$p(y|x) = \left(\frac{N_2}{N_1}\right) y \frac{(\mathrm{x}+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}$$

p value corresponds to differential gene expression test. The threshold of *p* value in multiple tests was determined by false discovery rate (FDR) value (Benjamini and Yekutieli 2001). We used "FDR $\leq$0.001 and the absolute value of $\log_2$Ratio $\geq$1" as the threshold to judge the significance of gene expression difference. In addition, we performed cluster analysis of gene expression patterns with Cluster software (Eisen et al. 1998) and Java Treeview (Saldanha 2004). To analyze gene expression profiling, hypergeometric test was employed to find significantly enriched GO terms in DGEs comparing to the *M. cochinchinensis* transcriptome background. The formula used is:

$$P = 1 - \sum_{i-1}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where *N* is the number of all genes with GO annotation; *n* is the number of DGEs in *N*; *M* is the number of all genes that are annotated to the certain GO terms; m is the number of DGEs in *M*. For pathway enrichment analysis, we mapped all differentially expressed genes to the terms in the KEGG database and looked for significantly enriched KEGG terms compared with whole transcriptome background.

Extraction and high performance liquid chromatography analysis of carotenoids from *M. cochinchinensis*

The carotenoid content in the samples was measured using the method described in Tuan et al. (2011). A sample (50 mg) was placed in a vessel, protected from light, and mixed with 1.5 ml of extraction solvent (ethanol containing 0.1 % ascorbic acid (w/v)). This mixture was vortexed for 10 min and then incubated in a water bath at 60 °C for 5 h and changing the extraction solvent 3 times. The extract was filtered through 0.45 μm membrane filters and 10 μl was injected for high performance liquid chromatography (HPLC) analysis.

For HPLC analysis, the extract was separated on Shimadzu 20AT system with Symmetry C18 (250 mm × 4.6 mm, 5 mm, Waters, Milford, USA), Sentry guard C18 (15 mm × 4.6 mm, 5 mm, Waters, Milford, USA) and UV/Vis detector. The mobile phase was composed of

acetonitrile/methanol (30:70 v/v) at a flow rate of 1.5 ml min^{-1}. The column temperature was 40 °C and the reading was taken at 475 and 454 nm wavelength.

Quantitative real-time PCR analysis

Twelve novel transcripts related to terpenoid backbone and carotenoid biosynthesis were chosen to determine their expression patterns in the three different ripening stage samples. Specific primer pairs used in real-time PCR were designed as shown in Supplemental Table 1. Real-time PCR was performed on a ECO™ Real-time PCR system (Illumina, USA) with the SsoFast™ EvaGreen supermix (Bio-Rad, USA) under the following condition: 95 °C for 10 min followed by 40 cycles of 95 °C for 10 s, 55 °C for 30 s and 72 °C for 15 s. The expression levels of different genes were normalized to the constitutive expression level of ubiquitin.

Results

Illumina sequencing and de novo assembly of *M. cochinchinensis* transcriptome

To obtain an overview of the *M. cochinchinensis* aril transcriptome, cDNA library constructed from the total RNA of aril in 45 DAF fruit was subjected to pair-end read with the Illumina platform. After the adapter sequences were trimmed and reads shorter than 75 bp were removed, 51,466,670 clean pair-end reads with total of 4,632,000,300 bps were obtained for assembly. The Illumina reads generated in this study are available at the web site (http://gac.kimjy.gnu.ac.kr). 5.2 % of a total reads were filtered out with the Q20 conditions, and an overview of the sequencing and assembly is given in Table 2. Based on high quality reads, a total of 477,695 contigs were assembled with an average length of 137 bp (Supplemental Fig. 1a). The contigs were then joined into 119,366 scaffolds (average length of 303 bp) using paired-end information as well as gap-filling process. Scaffolds with the length ranging from 100 to 500 bp accounted for 86.93 % (Supplemental Fig. 1b), and 79.8 % scaffolds had zero gap (Supplemental Fig. 2a). We also determined that scaffold N50 was 380 bp long by counting the estimated intra-scaffold gaps (Table 2). To further shorten the remaining gaps, we used those pair-end reads to fill the scaffold gaps in order to obtain unigenes with the least Ns and which could not be extended on either end. Finally, de novo assembly yielded 81,404 unigenes with a final N50 length of 450 bp, and total length of 31.57 Mb (Table 2). There were 65,775 unigenes (80.80 % of all unigenes) with length varying from 200 to 500 bp (Supplemental Fig. 1c) and 71,096 unigenes (87.33 % of all unigenes) showed no gap (Supplemental Fig. 2b).

Table 2 Overview of the sequencing and assembly

Total number of reads	51,466,670
Total nucleotides (nt)	4,632,000,300
GC percentage	45.38 %
Q20 percentage	94.8 %
Step-wise assembly	
Contig	
Total number of contigs	477,695
Length of all contigs (nt)	65,480,881
Average sequence size of contigs (nt)	137
Contigs N50 (nt)	117
Scaffold	
Total number of scaffolds	119,366
Length of all scaffolds (nt)	36,153,015
Average sequence size of scaffolds (nt)	303
Scaffolds N50 (nt)	380
Unigene	
Total number of unigenes	81,404
Length of all unigenes (nt)	31,579,810
Average sequence size of unigenes (nt)	388
Unigenes N50 (nt)	450

Sequence similarity of gac aril transcripts with other plants

Phylogenomic analysis allows us to statistically evaluate the variability between species or comparative analyses between large groups of species (Bininda-Emond 2004). Recently, it has been shown that large-scale transcriptome data are a potential source of information for multigene phylogenomic analysis (Logacheva et al. 2011). For phylogenomic analysis, gac aril transcriptome has been clustered with 20 plant genomes using Tribe-MCL, and the 2,926 clusters were selected for generating of phylogenetic tree using Neighbor joining method (Saitou and Nei 1987). As expected, gac aril transcriptome closely related to *Cucumis sativus* genome, which belongs to the family *Cucurbitaceae* (Fig. 1a). In addition, gac aril transcriptome exhibited 81.47 % similarity with *Cucumis sativus* at protein level, whereas the monocot proteomes have about 63 % similarity with gac aril transcriptome (Fig. 1b). The significant homology with *C. sativus* proteins and gac aril transcriptome indicates their evolutionary relationship.

Functional annotation

To identify the putative function of assembled unigenes, sequencing similarity search was conducted against NCBI non-redundant (NR) database and Swiss-Prot protein

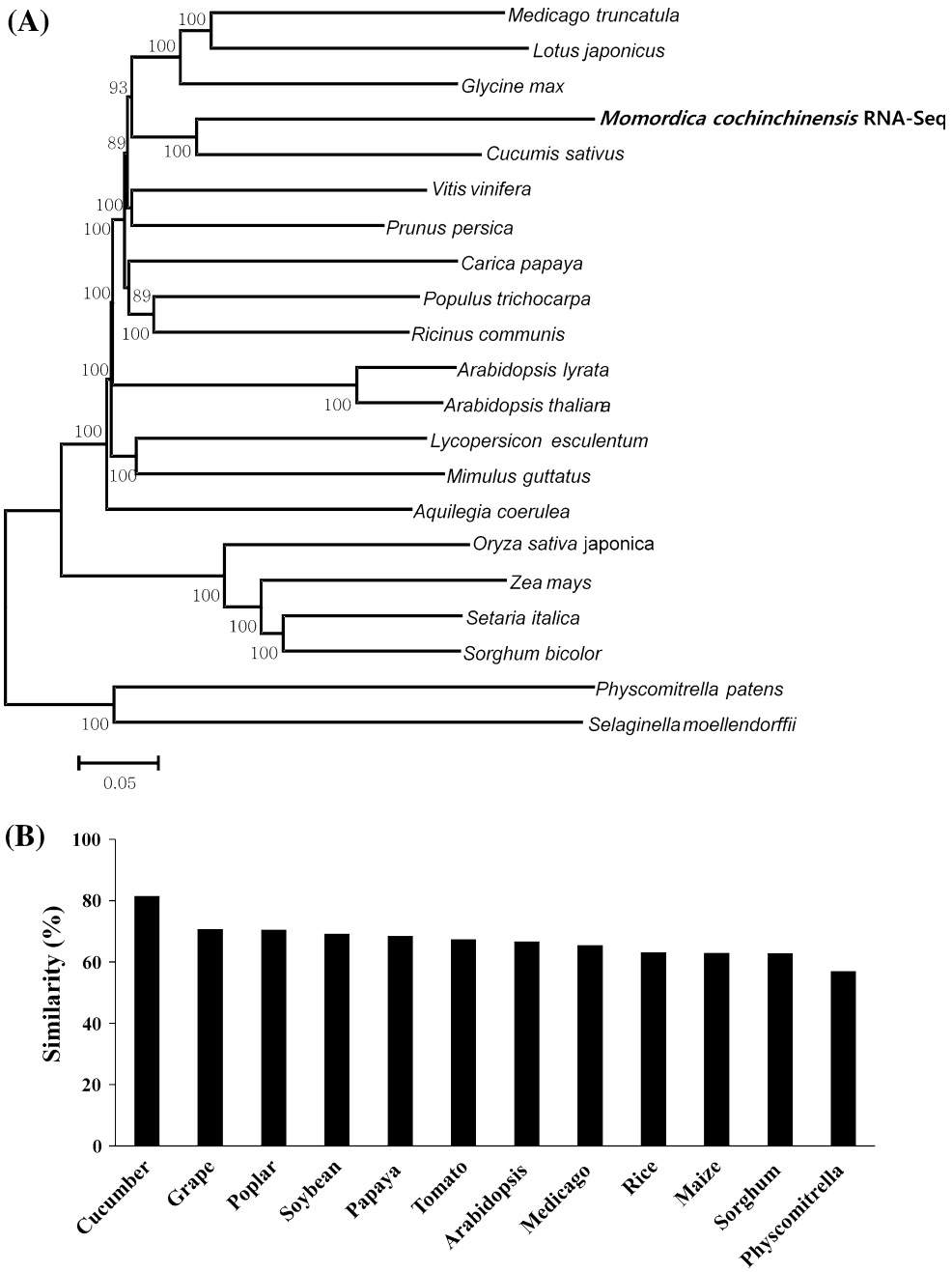

Fig. 1 Sequence conservation of gac transcripts with other plant species. **a** Phylogenomic tree of gac aril transcriptome sequences with different plants species. Phylogenomic analysis was carried out using the neighbor-joining method with 1,000 bootstrap repeats using MEGA5. **b** Sequence conservation of gac aril transcripts with annotated proteins of completely sequenced plant species. The percentage of transcripts showing significant similarity (E value $< 10^{-5}$) in BLASTx searche are shown

database using BLASTX search with a cut-off E value of 10^{-5} (Supplemental Table 2). A total of 39,385 (43.38 % of all unigenes) showed significant similarity to known proteins in NR database (Table 3). Due to limitation of genome and EST information in *Cucurbitaceae*, 56.62 % of unigenes could not be matched to known genes. As expected, the low percentage of unigenes (29.13 % of all unigenes) was found for the search against Swiss-Prot protein database.

Gene ontology (GO), which is an international standardized gene functional classification system, was used

Table 3 Summary of annotations of *M. cochinchinensis* unigenes in public protein databases

Public protein database	No. of unigene hit	Percentage
NR	39,385	43.38
KEGG	17,104	21.01
COG	9,678	11.89
Swiss-Prot	23,712	29.13
GO	20,420	25.08
Total	39,586	48.63

for classifying the function of the predicted *M. cochinchinensis* genes. Based on sequence homology, a total of 20,420 (25.08 % of all unigenes) transcripts were assigned at least one GO term including 44 functional groups (Table 3, Supplemental Fig. 3). Under the biological process category, metabolic process and cellular process were prominently represented. Cell compartments and organelles represented the majorities of terms in the cellular component category, whereas the vast majority is related to binding and catalytic activity in the molecular function category. In our database, 10,136 unigenes were annotated as related to metabolic processes, which suggests that this finding has provided abundant information on novel genes involved in the metabolic pathways including secondary metabolism.

To predict and classify possible function, all unigenes were aligned to the Cluster of Orthologous Groups (COG) database where the orthologous gene products were classified. Out of 39,385 NR hits, 9,678 sequences were assigned to the COG classification (Table 3). Because some of these unigenes were annotated with multiple COG functions, total 15,582 functional annotations were produced (Supplemental Fig. 4). Among the 25 COG categories, the cluster for "general functions prediction only" (2,395, 15.37 %) associated with basic physiological and metabolic functions represented the largest group, followed by "Replication, recombination and repair" (1,442, 9.25 %), "Posttranslational modification, protein turnover, chaperones" (1,320, 8.47 %), "transcription" (1,317, 8.45 %) and "Translation, ribosomal structure and biogenesis" (1,054, 6.76 %), whereas only a few unigenes were assigned to "Extracellular structures" and "Nuclear structure". In addition, 408 unigenes (2.62 %) were assigned to "Secondary metabolites biosynthesis, transport and catabolism" (Supplemental Fig. 4). The most abundant sequences in this category were cytochrome P450 (COG2124), with a total of 84 unigenes involved in various metabolic pathways.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database is a knowledge base for systematic analysis of gene functions in terms of networks of genes and molecules in the cells and their variants specific to particular organisms (Ogata et al. 1999). Genes within the same pathway usually cooperate with each other to exercise their biological function, indicating that pathway-based analysis is a useful tool to further understand the biological functions of genes (Wenping et al. 2011). Therefore, to identify the biological pathways activated in the aril of *M. cochinchinensis*, the assembled unigenes were annotated with corresponding Enzyme Commission (EC) numbers using BLASTx alignments against KEGG with a cut-off *E* value of 10^{-5}. In total, out of 39,385 Nr hits, 17,104 unigenes had significantly matched in the database, and were assigned to 115 KEGG pathways (Table 3; Fig. 2). As shown in Fig. 2a, the mapped unigenes represented metabolism pathways of major biomolecules such as carbohydrates, amino acids, lipids, etc. The pathways with most representation by unigenes were carbohydrate metabolism (2,286 unigenes) and biosynthesis of secondary metabolites (1,972 unigenes) in metabolism categories. In carbohydrate metabolism, starch and sucrose metabolism (508 unigenes) represented the most abundant classification. In the non-green heterotrophic fruit tissues, sucrose is cleaved and used to sustain cell metabolism and growth (Butowt et al. 2003). The utilization of sucrose as a source of carbon and energy depends on its cleavage by either invertase (β-fructofuranosidase, EC 3.2.1.26) or sucrose synthase (SuSy, EC 2.4.1.13) (Hyun et al. 2009). It has been shown that these enzymes link the transport of sucrose into cytosol, vacuole and apoplast (Kim et al. 2000; Nguyen-Guoc and Foyer 2001), indicating that these enzymes play a vital role in cellular processes such as carbohydrate transport and partitioning in plants. In *M. cochinchinensis* aril transcriptome, we found unigenes potentially encoding acid invertases (17 unigenes) and neutral/alkaline invertases (8 unigenes), whereas 18 unignens were aligned to SuSy (Supplemental Table 3). In addition, the mapped unigenes were also sorted to genetic information processing (transcription, translation, folding, sorting and degradation, replication and repair), environmental information processing (membrane transport, signal transduction, immune system and environmental adaptation) and cellular processes (transport and catabolism, cell motility, cell growth and death). Furthermore, the KEGG pathway analysis also showed that 1,503 out of 1,972 unigenes were classified into 19 sub-categories belonging to a secondary metabolite category (Fig. 2b). Among them, the cluster for 'Phenylpropanoid biosynthesis' represents the largest group followed by 'Limonene and pinene degradation' and 'Stilbenoid, diarylheptanoid and gingerol biosynthesis'. These findings indicated that transcriptome sequencing studies of *M. cochinchinensis* these annotations has facilitated the discovery of novel genes involved in the secondary metabolic pathways.

Analysis of differential gene expression during different ripen stages

To reveal the molecular events during different ripen stages, three digital gene expression (DGE) libraries were made from the arils of three ripening stages of fruits, and sequenced using Illumina technology. Stage 1, 2 and 3 are 35 DAF, 45 DAF and 55 DAF, respectively. After Illumina sequencing, we generated about 12,638,129, 11,944,453, and 11,999,671 raw tags for each of the three samples, respectively (Supplemental Table 4). Following transformation of raw sequences into clean tags, the total number

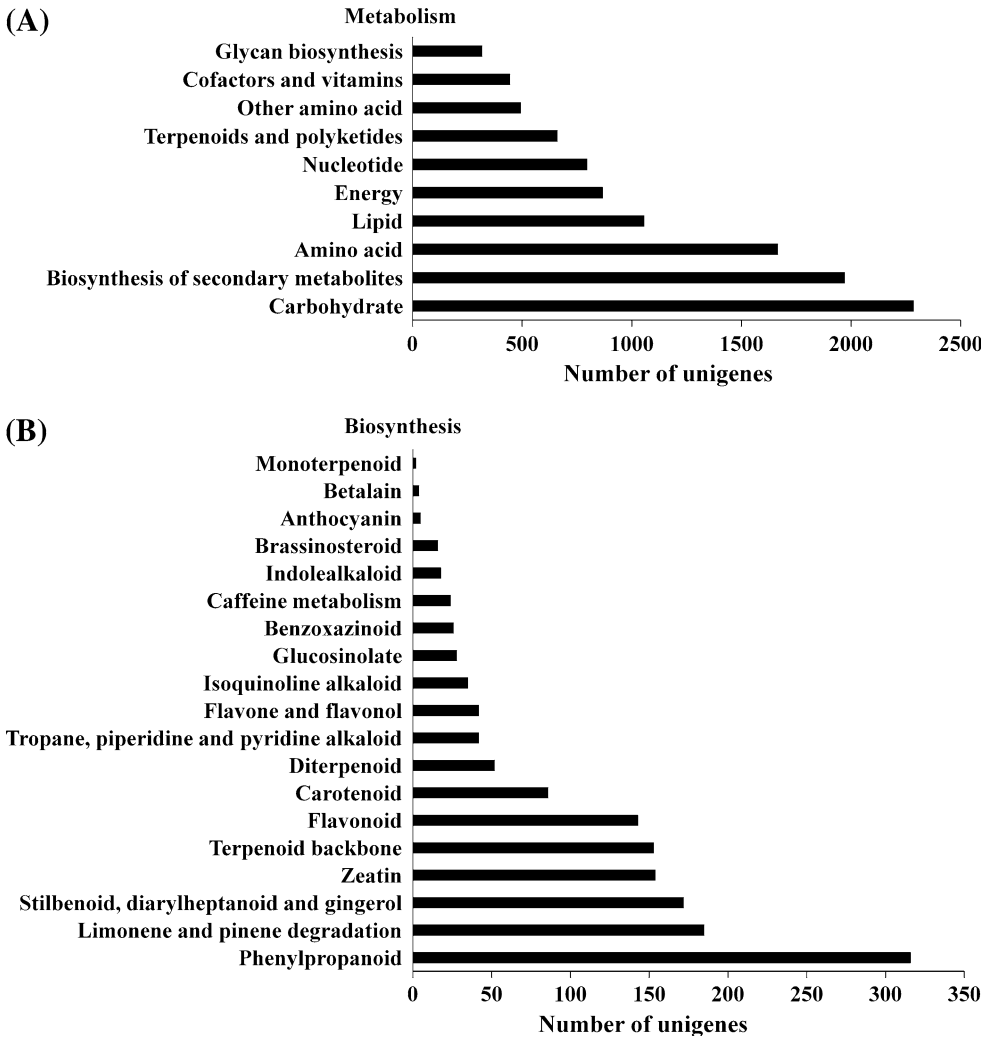

Fig. 2 Pathway assignment based on the Kyoto Encyclopedia of Genes and Genomes (KEGG). **a** Classification based on metabolism categories. **b** Classification based on categories of secondary metabolite biosynthesis

of tags per library ranged from 11,810,758 to 12,514,405 and more than 98 % of the total tags in the three libraries were clean tags (Supplemental Fig. 5). Then we mapped the tag sequences of the three DGE libraries to our transcriptome reference database that contains 81,404 unigene sequences. Differentially expressed tags between samples were identified by calculating the number of unambiguous tags for each gene and then normalizing this to the number of transcripts per millions tags (TPM). Following the algorithm developed by Audic and Claverie (1997), differentially expressed tags between two samples were identified, and filtered with FDR $\leq$0.001 and log2 Ratio $\geq$1. As shown in Fig. 3a, a total of 10,895 differentially expressed genes were detected between stage 1 and stage 2 libraries, and these included both up-regulated (6,962 genes) and down-regulated genes (3,933 genes). In addition, the comparative analysis between stage 2 and stage 3 libraries have been shown that 3,187 genes highly expressed in the stage 3, whereas the expression of 6,919 genes was down-regulated. 58.1 % (4,048 genes) of 6,962 genes which were up-regulated in stage 2 have been down-regulated in stage 3 (compared with stage 2). Interestingly, 94.7 % of 4,048 genes exhibit the similar expression level in stage 3 compared to stage 1. Thus, the relatively smaller changes (Fig. 3a) in gene expression between stage 1 and stage 3 is explained by that the same genes that were up-regulated in stage 2 were later down-regulated in stage 3. Based on KEGG pathway database, ontological analysis of up-regulated and down-regulated transcript groups revealed various affected metabolic processes, mainly carbohydrate and amino acid metabolism, and biosynthesis of secondary metabolites (Fig. 3b).

Candidate enzymes involved in carotenoid biosynthesis and other enzymes responsible for the biosynthesis of terpenoid C5-backbones

As shown in Fig. 5a, carotenoids are derived from isopentenyl diphosphate (IPP) which is synthesized by two independent pathways, methyl-eryhritol phosphate (MEP) pathway and mevalonate (MVA) pathway in higher plants (Flores-Pérez et al. 2010). Four molecules of IPP are

converted to geranylgeranyl diphosphate (GGPP, C20) by the action of IPP isomerase together with other enzymes like geranyl diphosphate synthase (GPS), farnesyl diphosphate synthase (FPS), geranylgeranyl pyrophosphate synthase (GGPS). In the first step of carotenoid pathway, phytoene synthase (PSY) catalyses the condensation of two molecules of GGPP to produce a 15-cis phytoene (C40) (Cunningham 2002). Based on the KEGG pathway assignment, we found 80 unigenes encoding enzymes involved in terpenoid backbones and carotenoid biosynthesis from our *M. cochinchinensis* transcriptome (Table 4). More than one unigene sequence was annotated as the same enzyme. This indicates that such unigene sequences represent different fragments of a single transcript, different members of a gene family, or both. Although most of them were partial sequences, we were able to identify 36 putative enzymes involved in terpenoid backbones biosynthetic pathways and carotenoid biosynthetic pathway.

Alteration of carotenoid biosynthetic pathway during fruit ripening in *M. cochinchinensis*

Based on our DGE analysis during different ripening stages, ripening represents coordinated modification of numerous metabolic pathways associated with carbohydrate and secondary metabolisms (Fig. 3b). It was therefore important to determine the expression of various genes involved in carotenoid biosynthetic pathway during fruit ripening in *M. cochinchinensis*. HPLC with UV/Vis detector was used to identify the content of lycopene and beta-carotene in the extract from different ripen stages of *M. cochinchinensis* arils (Supplemental Fig. 6). As shown in Fig. 4, trans-lycopene content increased from stage 1 (17.6 μg g^{-1} wet weight) to stage 2 (111.98 μg g^{-1} wet weight), and then maintained in stage 3 (110.47 μg g^{-1} wet weight), whereas fully ripe fruit of tomato contained 75.24 μg g^{-1} wet weight of lycopene. In addition, the significantly increased level of beta-carotene has been found during fruit ripening.

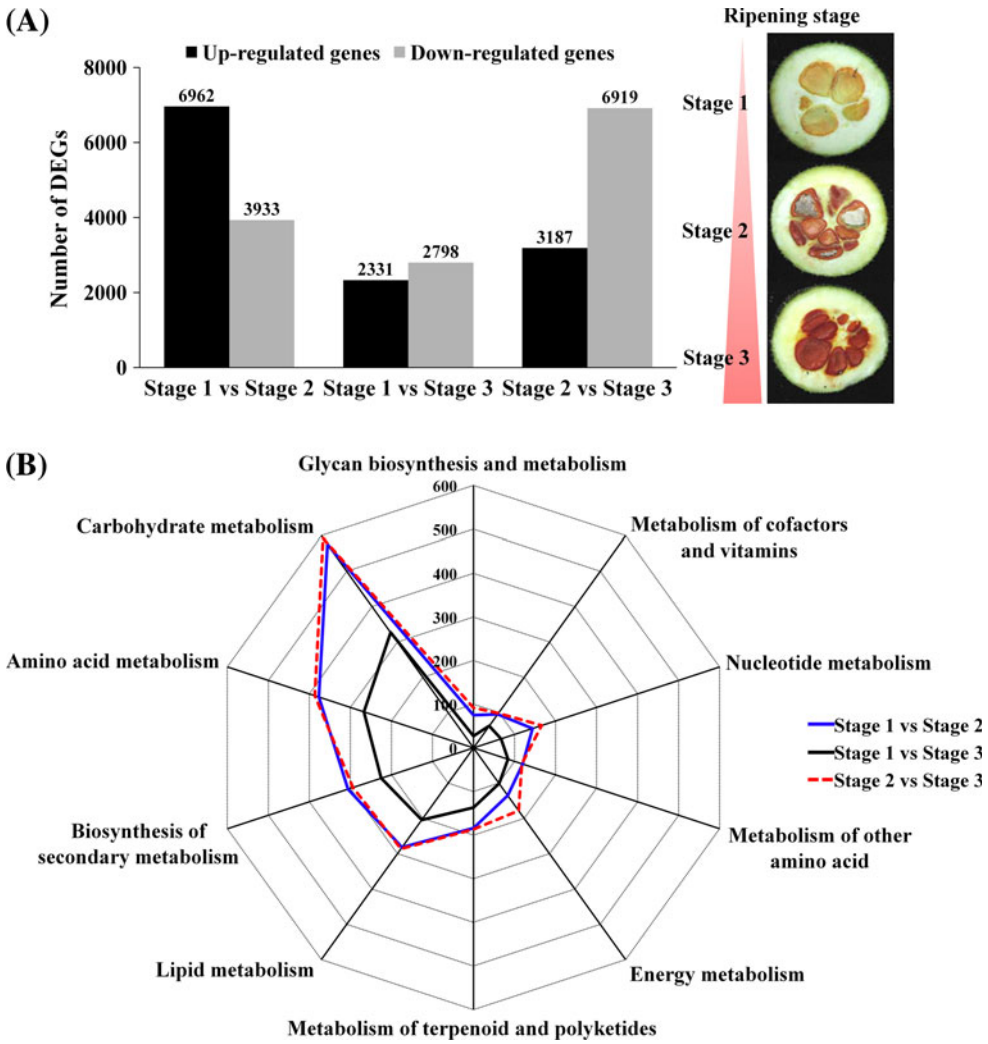

Fig. 3 Analysis of gene expression changes among the different aril ripening stages. **a** The profiling of differentially expressed genes. RNA was extracted from the aril of each stage of fruit. The number of differentially expressed genes was obtained from comparisons in Stage 1 versus Stage 2, Stage 1 versus Stage 3 and Stage 2 versus Stage 3, respectively. **b** Classification of the differentially expressed genes based on metabolism categories

Table 4 Unigenes potentially related to carotenoid biosynthesis including Mevalonate and MEP pathways

	Gene	EC number	Gene Name[a]	NU	MNPU	NGSN	Arabidopsis gene[b]
Mevalonate pathway	Acetyl-CoA C-acetyltransferase	2.3.1.9	ACAT	4	2	–	AT5G47720; AT5G48230
	Hydroxymethylglutaryl-CoA synthase	2.3.3.10	MVA	2	1	–	AT4G11820
	Hydroxymethylglutaryl-CoA reductase	1.1.1.34	HMG	9	3	–	AT1G76490; AT2G17370
	Mevalonate kinase	2.7.1.36	MK	7	2	–	AT5G27450
	Phosphomevalonate kinase	2.7.4.2	PMK	4	3	–	AT1G31910
	Diphosphomevalonate decarboxylase	4.1.1.33	MVD	2	1	–	AT2G38700; AT3G54250
MEP pathway	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	DXPS	2	1	–	AT3G21500; AT4G15560; AT5G11380
	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	DXR	3	1	–	AT5G62790
	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60	ISPD	2	1	–	AT2G02500
	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	2.7.1.148	CDPMEK	4	1	–	AT2G26930
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12	ISPF	2	1	–	AT1G63970
	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1	HDS	5	1	–	AT5G60600
	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2	HDR	3	2	–	AT4G34350
	Isopentenyl-diphosphate delta-isomerase	5.3.3.2	IPP	5	2	–	AT3G02780; AT5G16440
	Geranyl diphosphate synthase	2.5.1.1	GPS	2	2	–	AT2G34630[c]
	Farnesyl diphosphate synthase	2.5.1.10	FPS	2	2	–	AT5G47770[c]
	Geranylgeranyl pyrophosphate synthase	2.5.1.29	GGPS	7	3	–	AT1G49530[c]
Carotenoid biosynthesis	Phytoene synthase	2.5.1.32	PSY	3	1	–	AT5G17230
	15-cis-phytoene desaturase	1.3.5.5	PDS	5	2	–	AT4G14210
	15-cis-zeta-carotene isomerase	5.2.1.12	Z-ISO	–	–	–	AT1G10830
	Zeta-carotene desaturase	1.3.5.6	ZDS	1	1	–	AT3G04870
	Carotenoid isomerase (prolycopene isomerase)	5.2.1.13	CRTISO	3	1	–	AT1G06820
	Lycopene epsilon cyclase	5.5.1.18	LcyE	1	1		AT5G57030
	Lycopene beta-cyclase	5.5.1.19	LcyB	2	1	–	AT3G10230
Total				80	36	–	

NU number of unigenes, *MNPU* maximum number of proteins encoded by unigenes, *NGSN* number of genes stored in NCBI

[a] Following the *Arabidopsis thaliana* genes nomenclature

[b] Gene list from KEGG pathway database

[c] Gene list from Plant Metabolic Network (http://www.plantcyc.org/)

To identify key genes responsible for differences in carotenoid amounts in different ripening stages of arils, we generated a heatmap for gene expression ratios for each pair-wise comparison. Then the heatmaps generated were inserted into the terpenoid backbones biosynthetic and carotenoid biosynthetic pathways. In MVA pathway, the expression level of *HMG* (hydroxymethylglutaryl-CoA reductase) unigenes was more abundant than other genes. Specifically, the transcription level of *HMG* unigenes decreased from 35 DAF (stage 1) to 45 DAF (stage 2), but then some of unigenes increased significantly when the aril turned red (stage 3, 55 DAF) (Fig. 5a). In addition, the increased level of phosphomevalonate kinase (*PMK*), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (*CDPMEK*) and (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*HDS*) unigenes was detected from stage 2 in comparison to stage 1 sample, whereas the expression level of other genes involved in terpenoid backbone biosynthesis did not change during ripening (Fig. 5a). In case of carotenoid biosynthetic pathway, the unigenes of 15-cis-phytoene desaturase (*PDS*), zeta-carotene desaturase (*ZDS*) and carotenoid isomerase (*CRTISO*) were up-regulated from 35 DAF to 45 DAF, but then they were down-regulated when the aril turned red (55 DAF) (Fig. 5b).

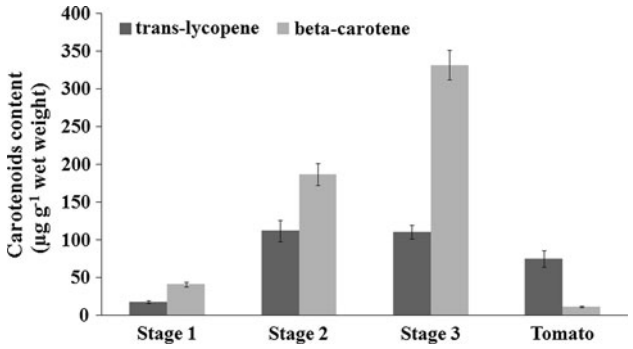

Fig. 4 Comparison of trans-lycopene and beta-carotene content from different ripening stages of gac aril

These findings suggest that the accumulation of lycopene in stage 2 aril (Fig. 4) might be due to increased expression level of carotenoid biosynthetic genes like *PDS*, *ZDS* and *CRTISO*. Furthermore, the accumulation of beta-carotene in stage 3 aril (Fig. 4) might be mediated by down-regulation of lycopene epsilon cyclase (*LcyE*) expression, which leads to the formation of α-carotene (Giuliano et al. 2008), rather than up-regulation of lycopene beta-cyclase (*LcyB*) (Fig. 5b and Supplemental Fig. 7b). In order to verify DGE data, the expression of 12 genes involved carotenoid biosynthetic pathway was analyzed by real time-PCR. As shown in Supplemental Fig. 7, the expression of these genes has varied during the three stages of fruit ripening, and expression pattern of these genes is similar to DGE.

Discussion

The presence of large amount of beta-carotene and lycopene in fruit of *M. cochinchinensis* indicates that it is a potentially valuable source of vitamin A, and could be a suitable model herbal plant for investigating terpenoid biosynthesis pathway, although the genomic information is not available in the public database, so far. In non-model plants, EST analysis is one of most valuable tools for the analysis and identification of novel genes (Tang et al. 2011). However, it has been used for identification of only a few candidate genes involved in complex biosynthetic pathways, because of the limited coverage of the transcriptome by EST analysis (Kim et al. 2006; Tang et al. 2011). With development of NGS technology, this limitation has been overcome with the major advantage that can construct cDNA library without the bacterial cloning step (Sun et al. 2010). Therefore, we applied RNA-seq technology for *M. cochinchinensis*, in which the poly (A) transcriptome was sequenced on the Illumina HiSeq^{TM}2000 sequencing platform. Based on de novo sequencing and assembly, we have established gene pool containing 81,404 unigenes (Table 2). In addition, 39,385 (43.38 % of all unigenes) unigenes returned significant hits from BLAST comparisons with the public databases, although no genomic information of *M. cochinchinensis* and little genomic information in *Cucurbitaceae* family are available. Comparison of assembled genes to gene catalogs of other plants by BLAST and functional annotation (Supplemental Figs. 3 and 4) and classification (Fig. 2) indicates that the aril of *M. cochinchinensis* fruit contains an extensive and diverse expressed gene pattern. DGE profiling by a tag-based transcriptome sequencing approach revealed quantitative changes in transcript abundance on a genome-wide scale through counting the number of individual mRNA molecules produced from each gene (Eveland et al. 2010; Tang et al. 2011). Results from our DGE analyses in arils of three ripening stages showed that we could effectively identify differentially expressed genes across a wide range of transcripts (Fig. 3). Taken together, the sequencing data and global observation on gene expression in the different ripen stages of arils we have made available a large database and unique information for gene discovery playing key roles in fruit development and ripening and for marker development for molecular breeding. In addition, this data would be useful for the genomic study and the identification of functional genes in *Cucurbitaceae* family.

Carotenoids are widely distributed isoprenoid pigments executing remarkable variety of functions in plants (Frank and Cogdell 1996). In most photosynthetic organisms, carotenoids play a vital role in protecting the photosynthetic machinery from photo-oxidation sensitized by excited molecules of chlorophyll (Cunningham and Gantt 2007), and accumulate as secondary metabolites in chromoplasts of flower, fruits, seeds or roots (Devitt et al. 2010). During fruit ripening, the accumulation of carotenoids like lycopene and beta-carotene has been observed in many of plants (Daqiu et al. 2011; Ronen et al. 1999; Tuan et al. 2011). Similarly, the significantly increased level of lycopene and beta-carotene contents has been found in *M. cochinchinensis*, when the aril turned from orange to red color (Fig. 4). It is well known that the arils from fully ripe fruit of gac (orange peel) contain higher level of lycopene content compared to other carotenoids (Aoki et al. 2002; Ishida et al. 2004), although the content of beta-carotene is higher than lycopene in the green gac fruit (Kubola and Siriamornpun, 2011), which is similar to our observation (Fig. 4). As shown by Ronen et al. (1999) in tomato fruit, high level of lycopene in fully ripe gac fruit might be due to the completely disappeared mRNA level of *LcyB* and *LcyE* that involve the alpha- and beta-carotene biosynthesis, whereas the expression of these genes resulted in the increasing alpha- and beta-carotene content in green gac fruit (Fig. 5b and Supplemental Fig. 7b).

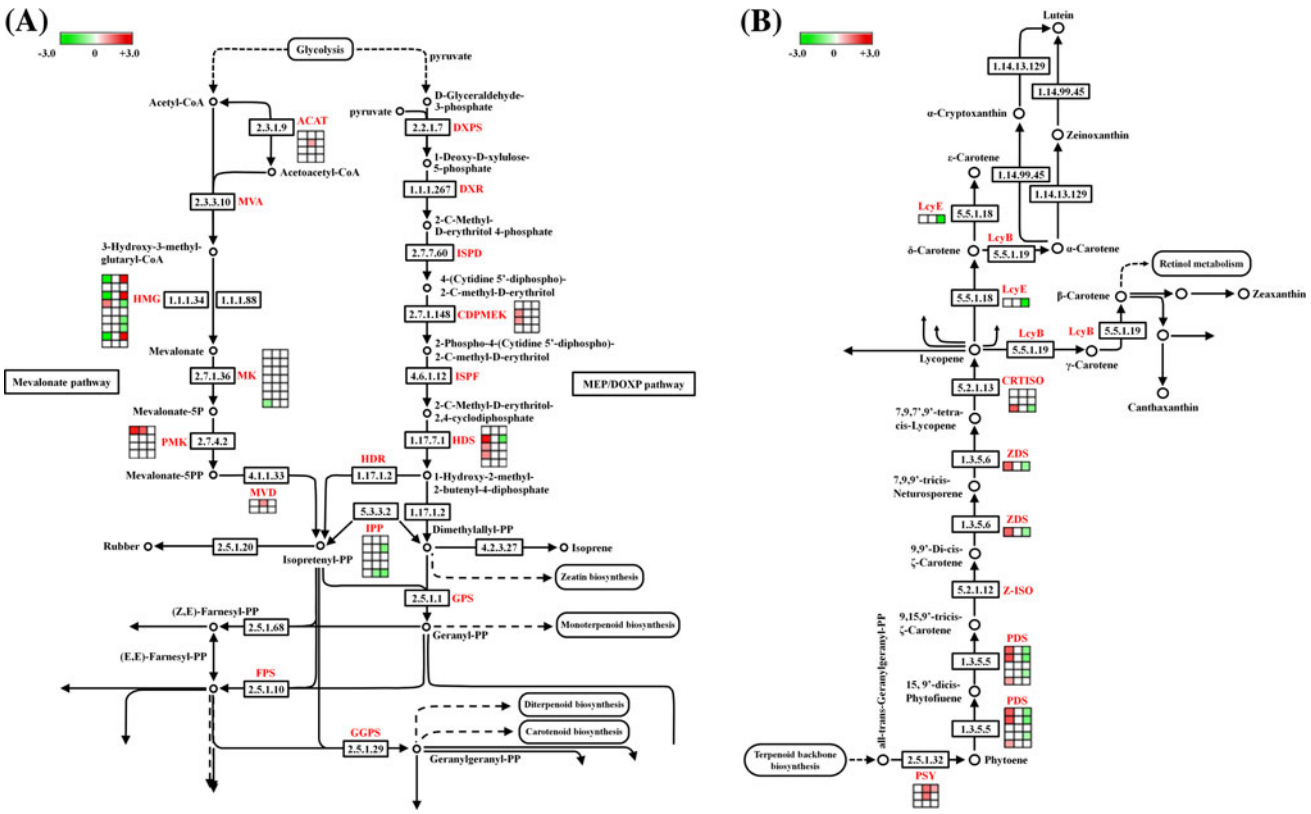

Fig. 5 Carotenoid biosynthetic genes differentially expressed in different ripening stages of gac aril. Pair-wise comparisons were made between stage 1, stage 2 and stage 3. **a** Expression pattern of genes involved in terpenoid backbone biosynthesis pathway. **b** Expression pattern of genes involved in carotenoid biosynthesis. Columns in each heatmap from *left* to *right*: Log2(stage 2/stage 1), Log2(stage 3/stage 1) and Log2(stage 3/stage 2). The Log2 expression ratio values were false color-coded using a scale of −3 to 3. The intensity of *green* and *red* indicates the degree of down- and up-regulation of the corresponding carotenoid biosynthetic genes in the denominator in each column mentioned above, respectively. The green and red color saturates at −3 and 3, respectively. The heatmaps generated were inserted next to the corresponding gene in the carotenoid biosynthesis pathway diagram referenced from the KEGG pathway database (map00900 and map00906)

The carotenoid biosynthesis pathway has been extensively studied in many photosynthetic and non-photosynthetic organisms using molecular genetics and biochemical genomics approaches. Carotenoid biosynthesis has been proposed via pathways through IPP synthesis by two independent pathways, MEP pathway in plastids and MVA pathway in cytosol (Chaudhary et al. 2010; Dellapenna and Pogson 2006; Devitt et al. 2010; Ronen et al. 1999; Wenping et al. 2011). In our *M. cochinchinensis* transcriptome database, we identified 80 unigenes encoding 36 putative enzymes involved in terpenoid backbones biosynthetic pathways (MEP and MVA pathways) and carotenoid biosynthetic pathway (Table 4). Comparative analysis between DGE libraries of different ripening stages demonstrated that the expression levels of *HMG*, *HDS*, *PSY*, *PDS*, *ZDS*, *CRTISO* and *LcyE* were more abundant than other genes involved in carotenoid biosynthesis (Fig. 5), indicating that the transcriptional regulation of these genes might be important for the alteration of carotenoid contents in *M. cochinchinensis* during fruit ripening. In fact, it has been shown that the increased expression of *PSY* and *PDS* is correlated with accumulation of lycopene and total carotenoid content during tomato fruit maturation (Bramley 2002; Cunningham 2002; Fraser et al. 2002). In *Arabidopsis*, it has been shown that loss-of-function of *LcyE* results in the accumulation of beta-carotene with xanthophylls cycle pigments (DellaPenna 1999; Pogson et al. 1996). The down-regulation of *LcyE* was detected in stage 3 compared with stage 2 DGE library (Fig. 5b and Supplemental Fig. 7b), when the highest level of beta-carotene was detected in stage 3 samples (Fig. 4). These findings indicate that LcyE is a key factor for controlling beta-carotene and lutein synthesis pathways during fruit ripening. The labeled precursors of the MVA pathway are incorporated into plastid isoprenoids (Kasahara et al. 2002), suggesting that MVA-derived IPP can be used for the synthesis of isoprenoids in the plastid (Kasahara et al. 2002; Rodriguez-Concepcion 2006). However, the overexpression of *HMG* in *Arabidopsis* does not leads to increased plastid chlorophyll and carotenoid contents (Wang et al. 2012), indicating that MVA pathway could

not account exclusively for the formation of plastid-derived isoprenoids (Fraser and Bramley 2004; Rodriguez-Concepcion 2006). The analysis of the isoprenoid gene network using sparse graphical Gaussian modeling suggested that HMG is a good candidate for interpathway crosstalk between MVA pathway and MEP pathway, and negatively correlates to some genes in MEP pathway (Wille et al. 2004). In addition, the inhibition of the cytosolic MVA pathway by treatment lovastatin (HMG inhibitor) resulted in an increased production or alternatively reduced catabolism of carotenoids (Laule et al. 2003). Similarly, we have detected the low level of *HMG* expression in stage 2 compared to stage 1 (Fig. 5a and Supplemental Fig. 7a), whereas the significantly increased level of lycopene and beta-carotene has been observed in stage 2 (Fig. 4). Taken together, these findings indicate that crosstalk between the cytosolic MVA pathway and the plastidal MEP pathway of isoprenoid biosynthesis might be a common phenomenon in higher plants.

The identification of enzymes and its genes involved in carotenoid biosynthesis pathways not only facilitates the understanding of physiological functions in higher plants but also provides useful resource for metabolic engineering. Based on de novo transcriptome, functional annotation of novel genes and DGE analyses, we provided most of the known genes encoding enzymes involved in terpenoid backbones and carotenoid biosynthesis in *M. cochinchinensis*. To the best of our knowledge, this is the first attempt using Illumina sequencing platform for de novo sequencing and assembly of gac fruit aril without reference genome. We strongly believe that this transcriptome dataset will provide a solid foundation for further functional and genomics research in *M. cochinchinensis* with *Cucurbitaceae* family, and will support the gene resources for improving carotenoid contents using biotechnological applications and metabolic engineering.

Acknowledgments This work was supported by Science and Technology and a grant from the Next-Generation BioGreen 21 Program (SSAC grant PJ008109, PJ008184), Rural Development Administration, Republic of Korea, and funded by the Industrial Strategic technology development program, 10040231, "Bioinformatics platform development for next generation bioinformation analysis", the Ministry of Knowledge Economy (MKE), Republic of Korea.

References

Angeloni F, Wagemaker CAM, Jetten MSM, Op den Camp HJM, Janssen-Megens EM, Francoijs K-J, Stunnenberg HG, Ouborg NJ (2011) De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. Mol Ecol Resour 11:662–674

Aoki H, Kieu NT, Kuze N, Tomisaka K, Van CN (2002) Carotenoid pigments in GAC fruit (*Momordica cochinchinensis* SPRENG). Biosci Biotechnol Biochem 66:2479–2482

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7:986–995

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Bramley PM (2002) Regulation of carotenoid formation during tomato fruit ripening and development. J Exp Bot 53:2107–2113

Burke DS, Smidt CR, Vuong LT (2005) *Momordica cochinchinensis*, *Rosa roxburghii*, wolfberry, and sea buckthorn-highly nutritional fruits supported by tradition and science. Curr Top Nutraceutical Res 3:259–266

Busch M, Seuter A, Hain R (2002) Functional analysis of the early steps of carotenoid biosynthesis in tobacco. Plant Physiol 128:439–453

Butowt R, Granot D, Rodriguez-Garcia MI (2003) A putative plastidic glucose translocator is expressed in heterotrophic tissues that do not contain starch, during olive (*Olea europea* L.) fruit ripening. Plant Cell Physiol 44:1152–1161

Castillo R, Fernandez JA, Gomez–Gomez L (2005) Implications of carotenoid biosynthetic genes in apocarotenoid formation during the stigma development of *Crocus sativus* and its closer relatives. Plant Physiol 139:674–689

Chaudhary N, Nijhawan A, Khurana JP, Khurana P (2010) Carotenoid biosynthesis genes in rice: structural analysis, genome-wide expression profiling and phylogenetic analysis. Mol Genet Genomics 283:13–33

Chew BP, Park JS (2004) Carotenoid action on the immuneresponse. J Nutr 134:257S–261S

Clotault J, Peltier D, Berruyer R, Thomas M, Briard M, Geoffriau E (2008) Expression of carotenoid biosynthesis genes during carrot root development. J Exp Bot 59:3563–7353

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

Cunningham FX Jr (2002) Regulation of carotenoid synthesis and accumulation in plants. Pure Appl Chem 74:1409–1417

Cunningham FX Jr, Gantt E (2007) A portfolio of plasmids for identification and analysis of carotenoid pathway enzymes: *Adonis aestivalis* as a case study. Photosynth Res 92:245–259

Daqiu Z, Chunhua Z, Jun T (2011) Carotenoid accumulation and carotenogenic gene expression during two types of persimmon fruit (*Diospyros kaki* L.) development. Plant Mol Biol Rep 29:646–654

Del Villar-Martinez AA, Garcia-Saucedo PA, Carabez-Trejo A, Cruz-Hernandez A, Paredes-Lopez O (2005) Carotenogenic gene expression and ultrastructural changes during development in marigold. J Plant Physiol 162:1046–1056

DellaPenna D (1999) Carotenoid synthesis and function in plants: Insights from mutant studies in Arabidopsis. Pure Appl Chem 71:2205–2212

Dellapenna D, Pogson BJ (2006) Vitamin synthesis in plants: tocopherols and carotenoids. Annu Rev Plant Biol 57:711–738

Devitt LC, Fanning K, Dietzgen RG, Holton TA (2010) Isolation and functional characterization of a lycopene β-cyclase gene that controls fruit colour of papaya (*Carica papaya* L.). J Exp Bot 61:33–39

Dixon RA (2001) Natural products and plant disease resistance. Nature 411:843–847

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Eveland AL, Satoh-Nagasawa N, Goldshmidt A, Meyer S, Beatty M, Sakai H, Ware D, Jackson D (2010) Digital gene expression signatures for maize development. Plant Physiol 154:1024–1039

Flores-Pérez U, Pérez-Gil J, Closa M, Wright LP, Botella-Pavía P, Phillips MA, Ferrer A, Gerhenzon J, Rodríguez-Concepción M

(2010) PLEIOTROPIC REGULATORY LOCUS 1 (PRL1) integrates the regulation of sugar responses with isoprenoid metabolism in Arabidopsis. Mol Plant 3:101–112

Frank HA, Cogdell RJ (1996) Carotenoids in photosynthesis. Photochem Photobiol 63:257–264

Fraser PD, Bramley PM (2004) The biosynthesis and nutritional uses of carotenoids. Prog Lipid Res 43:228–265

Fraser PD, Romer S, Shipton CA, Mills PB, Kiano JW, Misawa N, Drake RG, Schuch W, Bramley PM (2002) Evaluation of transgenic tomato plants expressing an additional phytoene synthase in a fruit-specific manner. Proc Natl Acad Sci USA 99:1092–1097

Garg R, Patel RK, Tyagi AK, Jain M (2011) De Novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res 18:53–63

Giovannucci E (2002) Lycopene and prostate cancer risk: methodological considerations in the epidemiologic literature. Pure App Chem 74:1427–1434

Giuliano G, Tavazza R, Diretto G, Beyer P, Taylor MA (2008) Metabolic engineering of carotenoid biosynthesis in plants. Trends Biotechnol 26:139–145

Hao DC, Ge G, Xiao P, Zhang YY, Yang L (2011) The first insight into the tissue specific taxus transcriptome via Illumina second generation sequencing. PLoS ONE 6:e21220

Harborne JB (2001) Twenty-five years of chemical ecology. Nat Prod Rep 18:361–379

Hyun TK, Hoffmann A, Sinha AK, Roitsch T (2009) Tomato mitogen activated protein kinases regulate the expression of extracellular invertase *Lin6* in response to stress related stimuli. Funct Plant Biol 36:1088–1097

Isaacson T, Ronen G, Zamir D, Hirschberg J (2002) Cloning of tangerine from tomato reveals a carotenoid isomerase essential for production of β-carotene and xanthophylls in plants. Plant Cell 14:333–342

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In: Proceedings of international conference, Intell Syst Mol Biol pp 138–148

Ishida BK, Turner C, Chapman MH, Mckeon TA (2004) Fatty acid and carotenoid composition of Gac (*Monordica cochinchinensis* Spreng) fruit. J Agric Food Chem 52:274–279

Iwamoto M, Okabe H, Yamauchi T, Tanaka M, Rokutani Y, Hara S, Mihashi K, Higuchi R (1985) Studies on the constituents of *Momordica cochinchinensis* Spreng. I. Isolation and characterization of the seed saponins, momordica saponins I and II. Chem Pharm Bull 33:464–478

Kasahara H, Hanada A, Kuzuyama T, Takagi M, Kamiya Y, Yamaguchi S (2002) Contribution of the mevalonate and methylerythritol phosphate pathways to the biosynthesis of gibberellins in Arabidopsis. J Biol Chem 277:45188–45194

Kim J-Y, Mahe A, Brangeon J, Prioul JL (2000) A maize vacuolar invertase, IVR2, is induced by water stress. Organ/tissue specificity and diurnal modulation of expression. Plant Physiol 124:71–84

Kim MK, Lee BS, In JG, Sun H, Yoon JH, Yang DC (2006) Comparative analysis of expressed sequence tags (ESTs) of ginseng leaf. Plant Cell Rep 25:599–606

Kubola J, Siriamornpun S (2011) Phytochemicals and antioxidant activity of different fruit fractions (peel, pulp, aril and seed) of Thai gac (*Momordica cochinchinensis* Spreng). Food Chem 127:1138–1145

Laule O, Furholz A, Chang H-S, Zhu T, Wang X, Heifetz PB, Gruissem W, Lange MB (2003) Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in Arabidopsis thaliana. Proc Natl Acad Sci USA 100:6866–6871

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Logacheva MD, Kasianov AS, Vinograd DV, Samigullin TH, Gelfand MS, Makeev VJ, Penin AA (2011) De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12:30

Markovic K, Hruskar M, Vahcic N (2006) Lycopene content of tomato products and their contribution to the lycopene intake of Croatians. Nutr Res 26:556–560

Marty I, Bureau S, Sarkissian G, Gouble B, Audergon JM, Albagnac G (2005) Ethylene regulation of carotenoid accumulation and carotenogenic gene expression in colour-contrasted apricot varieties (*Prunus armeniaca*). J Exp Bot 417:1877–1886

Moehs CP, Tian L, Osteryoung KW, DellaPenna D (2001) Analysis of carotenoid biosynthetic gene expression during marigold petal development. Plant Mol Biol 45:281–293

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

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349

Nantachit K, Tuchinda P (2009) Antimicrobial activity of hexane and dichloromethane extracts from *Momordica cochinchinensis* (Lour.) Spreng leaves. Thai Pharm Health Sci J 4:15–20

Nguyen-Guoc B, Foyer CH (2001) A role for 'futile cycles' involving invertase and sucrose synthase in sucrose metabolism of tomato fruit. J Exp Bot 52:881–889

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34

Park H, Kreunen SS, Cuttriss AJ, DellaPenna D, Pogson BJ (2002) Identification of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar body formation, and photomorphogenesis. Plant Cell 14:321–332

Pogson B, McDonald KA, Truong M, Britton G, DellaPenna D (1996) Arabidopsis carotenoid mutants demonstrate that lutein is not essential for photosynthesis in higher plants. Plant Cell 8:1627–1639

Rodriguez-Concepcion M (2006) Early steps in isoprenoid biosynthesis: multilevel regulation of the supply of common precursors in plant cells. Phytochem Rev 5:1–15

Romer S, Hugueney P, Bouvier F, Camara B, Kuntz M (1993) Expression of the genes encoding the early carotenoid biosynthetic enzymes in *Capsicum annuum*. Biochem Bioph Res Co 196:1414–1421

Ronen G, Cohen M, Zamir D, Hirschberg J (1999) Regulation of carotenoid biosynthesis during tomato fruit development: expression of the gene for lycopene epsilon-cyclase is down-regulated during ripening and is elevated in the mutant Delta. Plant J 17:341–351

Saitou N, Nei N (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Saldanha AJ (2004) Java Treeview–extensible visualization of microarray data. Bioinformatics 20:3246–3248

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S (2010) De novo sequencing and analysis of the American ginseng root transcriptome using GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary and genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tang Q, Ma X, Mo C, Wilson IW, Song C, Zhao H, Yang Y, Fu W, Qiu D (2011) An efficient approach to finding *Siraitia*

grosvenorii triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics 12:343

Tuan PA, Kim JK, Park NI, Lee SY, Park SU (2011) Carootenoid content and expression of phytoene synthase and phytoene desaturase genes in bitter melon (*Momordica charatia*). Food Chem 126:1686–1692

Vuong LT, Dueker SR, Murphy SP (2002) Plasma beta-carotene and retinol concentrations of children increase after a 30-d supplementation with the fruit *Momordica cochinchinensis* (gac). Am J Clin Nutr 75:872–879

Wang H, Nagegowda DA, Rawat R, Bouvier-Nave P, Guo D, Bach TJ, Chye M-L (2012) Overexpression of Brassica juncea wild-type and mutant HMG-CoA synthase 1 in Arabidopsis up-regulates genes in sterol biosynthesis and enhances sterol production and stress tolerance. Plant Biotechnol J 10:31–42

Wenping H, Yuan Z, Jie S, Lijun Z, Zhezhi W (2011) *De vono* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics 98:272–279

Wille A, Zimmermann P, Vranová E, Fürholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, Zitzler E, Gruissem W, Bühlmann P (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. Genome Biol 5:R92

Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JFS, Jung HJ, Vance CP, Gronwald JW (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. BMC Genomics 12:199

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34(Web Server issue):W293–W297

Zhang L, Kai G-Y, Lu B-B, Zhang H-M, Tang K-X, Jiang J-H, Chen W-S (2005) Metabolic engineering of tropane alkaloid biosynthesis in plants. J Integr Plant Biol 47:136–143

Zhu C, Naqvi S, Breitenbach J, Sanmann G, Christou P, Capell T (2008) Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. Proc Natl Acad Sci USA 105:18232–18237