Mol Breeding
DOI 10.1007/s11032-014-0128-x

The first Illumina-based de novo transcriptome sequencing and analysis of pumpkin (*Cucurbita moschata* Duch.) and SSR marker development

Tingquan Wu · Shaobo Luo · Rui Wang · Yujuan Zhong · Xiaomei Xu · Yu'e Lin · Xiaoming He · Baojuan Sun · Hexun Huang

Received: 6 January 2014 / Accepted: 2 June 2014

Abstract Pumpkin (*Cucurbita moschata* Duch.) is an important vegetable crop cultivated worldwide. In this study, the pumpkin transcriptome was sequenced by RNA-seq using the Illumina Hiseq 2000. A total of 52,849,316 clean sequencing reads, 66,621 contigs and 62,480 unigenes were postulated. Based on similarity searches with known proteins, 47,899 genes (76.66 % of the unigenes) were annotated: 47,596, 34,368 and 16,700 mapped in Nr, Swissprot and COG classifications, respectively; 21,164 were annotated with 44 gene ontology functional categories; and 13,728 were annotated to 269 pathways by searching the Kyoto Encyclopedia of Genes and Genomes pathway database. A total of 7,814 simple sequence repeats (SSRs) were identified in these unigenes and 4,794 pairs of primers were designed for application of SSRs. To date, 35 SSRs have been validated in 12 pumpkin varieties and can separate the pumpkin varieties into *Cucurbita maxima* and *Cucurbita moschata*. In addition, the expression of eight photoperiod-related unigenes were studied in different pumpkin plants and it was deduced that they may contribute to late flowering and light insensitiveness. This research will provide an important platform to facilitate gene discovery for functional genome studies of pumpkin and to conduct SSR discovery for breeders for use in pumpkin breeding.

Keywords Pumpkin · Transcriptome · SSR primers · Illumina sequencing · PPIS pumpkin plants

Abbreviations

GO	Gene ontology
KEGG	The Kyoto Encyclopedia of Genes and Genomes pathway database
SSR	Simple sequence repeat
PPIS	Photoperiod-insensitive
PPS	Photoperiod-sensitive
FAO	Food and Agriculture Organization of the United Nations
CDS	Coding sequences

Introduction

The Cucurbitaceae are an important plant family, consisting of approximately 125 genera and 960

Tingquan Wu and Shaobo Luo are co-first authors and contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-014-0128-x) contains supplementary material, which is available to authorized users.

T. Wu · S. Luo · R. Wang · Y. Zhong · X. Xu · Y. Lin · X. He · B. Sun · H. Huang (✉)
Vegetable Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China
e-mail: 214333407@qq.com

T. Wu · S. Luo · R. Wang · Y. Zhong · X. Xu · X. He · B. Sun
Guangdong Key Laboratory for New Technology Research of Vegetables, Guangzhou 510640, China

Published online: 13 June 2014

species, which ranks the highest among plant families for number and percentage of species used as human food (http://en.wikipedia.org/wiki/Cucurbitaceae). In the past few years, the whole-genome sequencing of several *Cucurbitaceae* crops has been completed, such as cucumber (Huang et al. 2009), melon (Garcia-Mas et al. 2012) and watermelon (Guo et al. 2012). However, the genomic data of pumpkin (*Cucurbita moschata* Duch.) ($2n = 2x = 40$) is seriously lacking. Pumpkin is an important vegetable and is cultivated extensively in the world. In China alone, in 2011, pumpkins, squash and gourds were cultivated on 377,682 ha (accounting for 21.28 % of world planting area), with a total production of 6,978,167 tonnes (accounting for 28.71 % of world production) (FAO 2011, http://faostat.org). In addition, pumpkin is believed to possess health and functional benefits because it is rich in nutrients such as vitamins, amino acids, flavonoids, phenolics and carbohydrates (Zhang et al. 2000; Wang et al. 2002). The pumpkin fruit can be stored for 4 months or even longer under proper conditions, though it can be grown all year round. In addition, various important medicinal properties of pumpkin have been well documented, including anti-diabetic, antioxidant, anti-carcinogenic and anti-inflammatory activities (Yadav et al. 2010). In contrast, research on pumpkin, especially at the molecular level, remains at a low level, out of proportion to its importance. To date, there is still a notable absence of data in genome research on pumpkin, which has seriously hindered developments in the fields of molecular biology and genetics.

The transcriptome is the complete collection of RNA, including mRNA, tRNA, rRNA and other non-coding RNA, from an organism, a tissue or a few mixed tissues, even a specific cell in a particular environment or a specific developmental stage. Transcriptome sequencing research has important value in new gene discovery, gene functional annotation, gene differential expression and molecular marker development (Lu et al. 2010; Wang et al. 2010; Zhang et al. 2012). Illumina paired-end sequencing technique, one of next-generation sequencing technologies, has provided a novel method for transcriptome analysis (Morozova and Marra 2008) and has been widely used in research on model and non-model plants. In recent years, applications of transcriptome sequencing have also been reported in the Cucurbitaceae family, such as *Citrullus lanatus* (Guo et al. 2011), *Cucumis sativus* (Ando et al. 2012), *Cucumis melo* (Blanca et al. 2012), *Momordica cochnichinensis* (Hyun et al. 2012), *Benicasa hispida* (Jiang et al. 2013) and so on.

Table 1 Summary of sequence assembly after Illumina sequencing

	Total nucleotides (nt)	All numbers	Mean length (bp)	N50 (bp)
Total clean reads	4,738,438,440	52,649,316	90	
Total contigs	52,477,436	66,621	788	1,241
Total unigenes	47,789,321	62,480	765	1,215
Unigenes (5′–3′)	42,827,798	47,714	898	1,359
GC percentage	48.09 %			
N percentage	0.00 %			
Q20 percentage	97.32 %			

In this work, sequencing of the transcriptome for pumpkin was initiated (NCBI BioProject Accession: SRA107523). RNA samples containing leaves, stems and shoots of pumpkin at the 2–4 true leaf stage were sequenced using the most popular Illumina sequencing technology. The purposes of this study are to construct a transcriptome databank of pumpkin and to develop simple sequence repeat (SSR) markers, laying the foundation for future research in the molecular biology, cytology and molecular genetics of pumpkin.

Results

Illumina sequencing and de novo assembly

To obtain transcriptome data of pumpkin, total RNA was extracted from leaves, stems and shoots at the five true leaf stage and the RNA sample was sequenced using the Illumina paired-end sequencing technique. Reads were assembled de novo using Trinity, a novel short-read assembly method for the efficient de novo reconstruction of the transcriptome (Grabherr et al. 2011). In this project, we obtained 52,649,316 clean reads with a total of 4,738,438,440 nucleotides (nt); the mean length of the reads was 90 base pairs (bp) and

the percentage of Q20, N and GC was 97.32, 0.00 and 48.09 % respectively (Table 1). The high-quality reads were assembled into 66,621 contigs with a total of 52,477,436 nt, a mean length of 788 bp and an N50 of 1,241 bp (Table 1). The total number of unigenes yielded in this program was 62,480 with 47,789,321 nt, a mean length of 765 bp and an N50 of 1,215 bp, and of them the total number of unigenes (5′–3′) was 47,714 with 42,827,798 nt, a mean length of 898 bp and an N50 of 1,359 bp (Table 1).

Evaluation of de novo sequence assembly

The sequence quality and coverage depth of the unigenes were evaluated using SOAPaligner, allowing up to two-base mismatches, to realign all the reads to the assembled unigenes (Li et al. 2008). The sequencing coverage depth ranged from 0.0316- to 94,530.4303-fold and the average fold was 44.3097. About 79.10 % (100–20.90 %), 31.33 % (100–20.90–36.35–11.42 %) and 12.66 % (5.20 % + 6.87 % + 0.59 %) of the unigenes were remapped with more than 10 reads, more than 100 reads and more than 500 reads, respectively (Supplementary Fig. 1).

Citrullus lanatus, *Cucumis sativus* and *Cucurbita moschata* Duch. belong to the gourd family. Unigenes were therefore compared with orthologous coding sequences of *Citrullus lanatus* and *Cucumis sativus*, respectively, to evaluate the extent of transcript coverage of unigenes. A total of 38,998 and 37,451 unigenes could be matched to the orthologous coding sequences of *Citrullus lanatus* and *Cucumis sativus*, respectively, and the alignment result showed that 239 (0.06 %) and 290 (0.77 %) unigenes, respectively, had the same length (ratio equal to 1) with the orthologs of *Citrullus lanatus* and *Cucumis sativus*, 3,323 (8.52 %) and 2,739 (7.31 %) unigenes had a greater length than the orthologs of *Citrullus lanatus* and *Cucumis sativus* (ratio greater than 1), and 35,436 (90.87 %) and 34,422 (91.9 %) unigenes, respectively, were shorter than the orthologs of *Citrullus lanatus* and *Cucumis sativus* (ratio less than 1) (Fig. 1A, C). 7,396 (47.91 %) *Citrullus lanatus* orthologs and 6,627 (45.02 %) *Cucumis sativus* orthologs could be covered by unigenes with a percentage greater than 80 %, and the coverage percentage of 3,496 (22.65 %) *Citrullus lanatus* and 3,465 (23.54 %) *Cucumis sativus* orthologs ranged from 50 to 80 % (Fig. 1B, D). Additionally, 1,191 (7.72 %) *Citrullus lanatus* and 1,292 (8.78 %) *Cucumis sativus* orthologs were covered by only 20 % of unigenes or lower (Fig. 1B, D). Although a number of unigenes failed to cover the complete coding sequences of their *Citrullus lanatus* and *Cucumis sativus* orthologs, our unigenes were able to cover most of the *Citrullus lanatus* and *Cucumis sativus* orthologs, and the ratios of unigene length to *Citrullus lanatus* ortholog length or u,igene length to *Citrullus lanatus* ortholog length were distributed around 1 (Fig. 1A, 1C), which indicated that the sequence quality was high.

Characterization of assembled transcriptome

The length distribution of the pumpkin contigs is shown in Supplementary Fig. 2A. A total of 66,621 contigs were assembled and all of the contigs were greater than 200 bp in length; of them, 32,393 contigs, about 48.62 % of the total, were greater than 500 bp in length. A total of 62,480 unigenes were assembled and of these, 29,131, about 48.62 % of the total, were longer than 500 bp and 15,566, about 24.91 % of the total, were longer 1,000 bp (Supplementary Fig. 2B). The length distribution of unigenes was closely consistent with contigs (Supplementary Fig. 2A, B). The coding sequences (CDS) were searched with unigenes against Nr, SwissProt, KEGG and COG protein databases in order, using the BLASTX program (E value $< 10^{-5}$) (Grabherr et al. 2011), and the functions of proteins were predicted from the annotations of the most similar proteins. A total of 45,643 unigenes, about 68.51 % of all assembled unigenes, were significant BLAST hits; the size distribution for the CDS is given in Supplementary Fig. 2C. The CDS of the unigenes without BLAST hits were predicted using ESTScan (Iseli et al. 1999), and 1,574 unigenes, about 2.36 % of the total, were analyzed using the ESTScan method; the size distributions of the expressed sequence tags are shown in Supplementary Fig. 2D.

Characteristics of similarity search of unigenes

Unigene annotations provide information on protein sequence similarities, gene ontology (GO) analysis, COG clusters and the KEGG pathway. We searched the sequences of pumpkin unigenes against the protein databases of Nr, SwissProt, COG and KEGG using BLASTX (E value $< 10^{-5}$) and predicted the functions of these proteins according to the functional annotations of their most similar proteins.

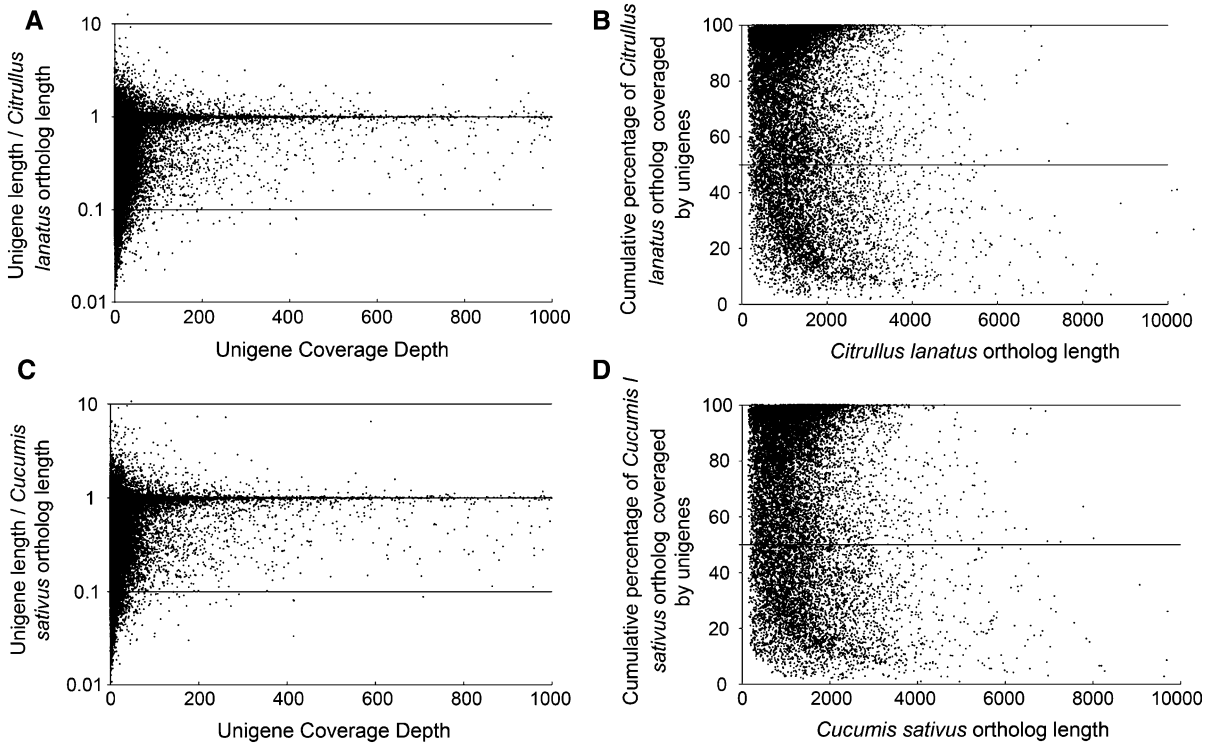

Fig. 1 Comparison of *Cucurbita moschata* Duch. unigenes to orthologies of *Citrullus lanatus*. **A** Ratio of *Cucurbita moschata* unigene length to *Citrullus lanatus* ortholog length plotted against *Cucurbita moschata* unigene coverage depth. **B** Total percent of *Citrullus lanatus* orthologies covered by all *Cucurbita moschata* unigenes. **C** Ratio of *Cucurbita moschata* unigene length to *Cucumis sativus* ortholog length plotted against *Cucurbita moschata* unigene coverage depth. **D** Total percent of *Cucumis sativus* orthologies covered by all *Cucurbita moschata* unigenes

BLASTed against the Nr database, 46,051 unigenes, about 69.12 % of the total, had hits and the *E* value distribution of the top hits revealed that 38,999 unigenes, about 84.69 % of the mapped unigenes, had homologous sequences with *E* values ranging from 1E−150 to 1E−5, while 7,052 unigenes, accounting for 15.31 % of the total of mapped unigenes, had strong similarity with *E* values smaller than 1E−150 (Supplementary Fig. 3A). Similarly, 32,815 unigenes, about 48.26 % of the total, were identified from the database of SwissProt. The *E* value distribution of the top hits showed that 3,621 unigenes, about 11.03 % of mapped unigenes, had strong similarity with *E* values smaller than 1E−150, whereas 29,194 unigenes, accounting for 88.97 % of the total of mapped unigenes, had homologous sequences with *E* values ranging from 1E−150 to 1E−5 (Supplementary Fig. 3B). BLASTed against the COG database, 15,637 unigenes, accounting for 23.47 % of the total, had homologous sequences. There were only 385 unigenes (2.46 % of mapped unigenes) with *E* values smaller than 1E−150, and 15,252 unigenes (97.54 % of mapped unigenes) with *E* values between 1E−150 and 1E−5 (Supplementary Fig. 3C). In addition, 14,455 unigenes, about 21.70 % of the total, were identified from the KEGG database. The *E* value distribution of the top hits showed that 2,998 unigenes, about 20.74 % of mapped unigenes, had strong similarity with the *E* values smaller than 1E−150, whereas 11,457 unigenes, accounting for 79.26 % of the total of mapped unigenes, had homologous sequences with *E* values ranging from 1E−150 to 1E−5 (Supplementary Fig. 3D).

Functional annotations

In order to predict and classify possible functions of the pumpkin unigenes, they were searched against the COG database. A total of 29,889 sequences had COG classifications and were distributed between 25 COG categories (Supplementary Fig. 4A). In the COG database, every protein was assumed to have evolved

from the same ancestor protein. Among the 25 COG categories, "General function prediction only" was the largest group (4,900, 16.39 %), followed by "Transcription" (2,790, 9.33 %), "Posttranslational modification, protein turnover, chaperones" (2,414, 8.08 %), "Replication, recombination and repair" (2,355, 7.88 %) and "Signal transduction mechanisms" (2,241, 7.50 %) followed in size, whereas only eleven and nine unigenes were assigned to "Nuclear structures" and "Extracellular structures," respectively (Supplementary Fig. 4A).

GO is an international standardized gene functional classification system which offers a dynamically-updated controlled vocabulary and a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component and biological process. The basic unit of GO is the GO-term. Every GO-term belongs to a type of ontology. GO functional analysis provides GO functional classification annotation for differentially expressed genes as well as GO functional enrichment analysis for differentially expressed genes. Here, we obtained GO functional annotations of the pumpkin unigenes with Nr annotations using the Blast2GO program (Conesa et al. 2005), and a total of 21,164 pumpkin unigenes obtained GO-term annotations using Blast2GO. The terms were summarized into three main GO groups and 44 sub-functional groups (Supplementary Fig. 4B). In the group of biological processes, "metabolic processes", "cellular processes" and "response to stimulus" were the most frequent terms and contained 10,335, 9,917 and 3,652 unigenes, respectively, whereas "cell killing" (eight unigenes), "biological adhesion" (nine unigenes) and "locomotion" (nine unigenes) were the least frequent (Supplementary Fig. 5B). In the two other main categories, "cellular components" and "molecular functions", the most prominent terms were "cell part" (14,545 unigenes), "cell" (14,545 unigenes), "binding" (10,995 unigenes), "catalytic activity" (10,126 unigenes) and "organelle" (9,945 unigenes) (Supplementary Fig. 4B).

Pathway annotations

The pathway analysis is able to help us better understand the biological functions of genes. In this study, 14,455 unigenes were annotated in 269 pathways (Supplementary Table S1) and most of the unigenes were enriched in 44 pathways (Supplementary Fig. 5). The number of unigenes in each pathway was more than 1 % of the total. The largest enrichment was 688 unigenes (4.76 % of the total unigenes annotated) in the "Ribosome" pathway, followed by "Plant hormone signal transduction", "Protein processing in endoplasmic reticulum", "RNA transport" etc. (Supplementary Fig. 5). In these pathways, we paid much more attention to "Plant hormone signal transduction" because the pumpkin seedlings were in the active period of flower bud differentiation with cross-talk of many kinds of endogenous plant hormones. 530 unigenes (3.67 % of the total unigenes annotated) in the "Plant hormone signal transduction" pathway were found in the pathways of auxin, cytokinine, gibberellins, abscisic acid, ethylene, brassinosteroid, jasmonic acid and salicylic acid (Supplementary Fig. 6).

The detection of photoperiod-related genes in photoperiod-sensitive and photoperiod-insensitive pumpkin plants

Pumpkin is a short-day plant, sensitive to day length, which poses an obstacle to its widespread planting. The material sequenced was a photoperiod-insensitive (PPIS) pumpkin variety obtained after years of breeding, with extensive regional adaptability. After April every year in Guangzhou (China), planted PPIS pumpkin is normally able to blossom and produce fruit; however, photoperiod-sensitive (PPS) pumpkin has few female flowers and fails to fruit or bears few fruit. The eight unigenes, viz. unigene0061604, unigene0042707, unigene0039722, unigene0048891, unigene0020784, unigene0062257, unigene0007796, unigene0053665 and unigene0053528, were found to have photoperiod-related orthologies in *Arabidopsis* and were named *Cm-LWD1*, *Cm-GI*, *Cm-FBH3*, *Cm-FIO1*, *Cm-PIE1*, *Cm-TIC*, *Cm-ELF6* and *Cm-APRR5*, respectively; they were detected by real-time PCR in PPIS and PPS pumpkin plants (Supplementary Table S2). The results of the test showed that, in PPIS pumpkin plants, the gene expression levels of *Cm-LWD1*, *Cm-FIO1*, *Cm-PIE1*, *Cm-TIC* and *Cm-ELF6* were higher than in PPS pumpkin plants; however, those of *Cm-GI*, *Cm-FBH3* and *Cm-APRR5* were lower (Fig. 2).

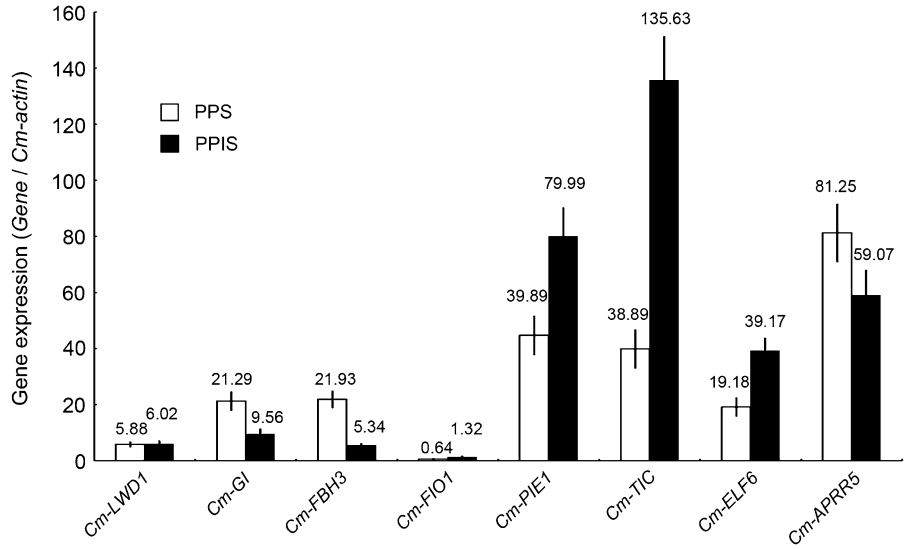

Fig. 2 The detection of photoperiod-related genes in PPS and PPIS pumpkin plants. The experiments were repeated twice and the results were similar

Development and application of SSR markers in pumpkin

In the present study, we identified 7,814 SSRs (Supplementary Table S3) with 4,794 pairs of primers designed for application (Supplementary Table S4), based on 62,481 sequences examined. The total size of the sequences examined was up to 47,789,321 bp (Supplementary Table S5). The number of SSR-containing sequences was 6,868 and the number of sequences containing more than one SSR was 1,094; 493 SSRs were present in compound formation (Supplementary Table S5). A total of 656 motif forms were identified in our transcriptome data. Tri-nucleotide motifs were the most abundant form of SSRs (4,254, 52.21 %), followed by di-nucleotide (2,747, 33.07 %), tetra-nucleotide (602, 7.25 %), hexa-nucleotide (398, 4.79 %) and penta-nucleotide motifs (306, 3.68 %) (Supplementary Table S5). In addition, the number of repeat unit of SSRs was determined and the results showed that the number of repeat units of the di-nucleotide motifs was distributed mainly from 6 to 11, and the tri-nucleotide, tetra-nucleotide, penta-nucleotide and hexa-nucleotide motifs mainly contained 5–7, 4–5, 4 and 5 repeat units, respectively (Supplementary Table S6). Furthermore, the motif types of SSRs were analyzed and their distribution is shown in Supplementary Fig. 7. The AG/CT motif was the most abundant with a frequency of up to 23.5 %, followed by AAG/CTT (19.4 %), AT/TA (6.8 %), AGG/CCT (6.1 %), AGC/CTG (5.6 %), CCG/CGG (4.8 %), ACC/GGT (3.6 %), ATC/ATG (3.5 %), AAC/GTT (3 %), AC/CT (2.8 %), AAT/ATT (2.5 %), ACG/CGT (2.2 %), AAAG/CTTT (1.2 %), AAAT/ATTT (1.1 %), AAAC/GTTT (0.7 %) and others (13.3 %) (Supplementary Fig. 7).

We randomly designed 57 pairs of primers (Supplementary Table S7) based on the SSRs developed in pumpkin and used them in 12 pumpkin varieties (Supplementary Table S8). The results of detection showed that, of the 57 pairs of SSR primers, 35 pairs for PCR were able to produce bands with polymorphism and 14 pairs of primers produced bands without polymorphism in 12 varieties, and eight pairs of primers were able to produce no or blurred bands (Supplementary Table S8). Several different results of PCR amplification are shown as examples in Fig. 3. The 0000087 primers for PCR produced no or blurred bands and the 0000100, 0000190 and 0000194 primers were able to produce non-polymorphic bands, polymorphic bands and polymorphic bands, respectively. In addition, the phylogenetic tree of the 12 pumpkin varieties was constructed based on the results of polymorphism detection and the results indicated that pumpkin varieties with *Cucurbita maxima* and *Cucurbita moschata* ancestry were divided into two main groups (Fig. 4). These experimental results revealed the applicability of the SSR primers developed, indicating that the quality of pumpkin RNA-seq was very high.

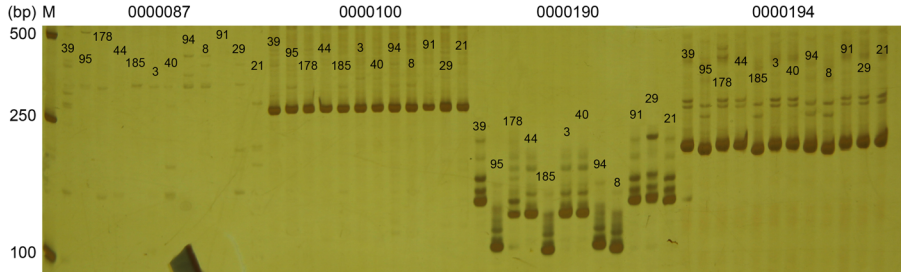

Fig. 3 Examples of polymorphic products amplified by SSR primer pairs. 0000087, 0000100, 0000190 and 0000194 are SSR primers. 39, 95, 178, 44, 185, 3, 40, 94, 8, 91, 29 and 21 represent 12 pumpkin varieties

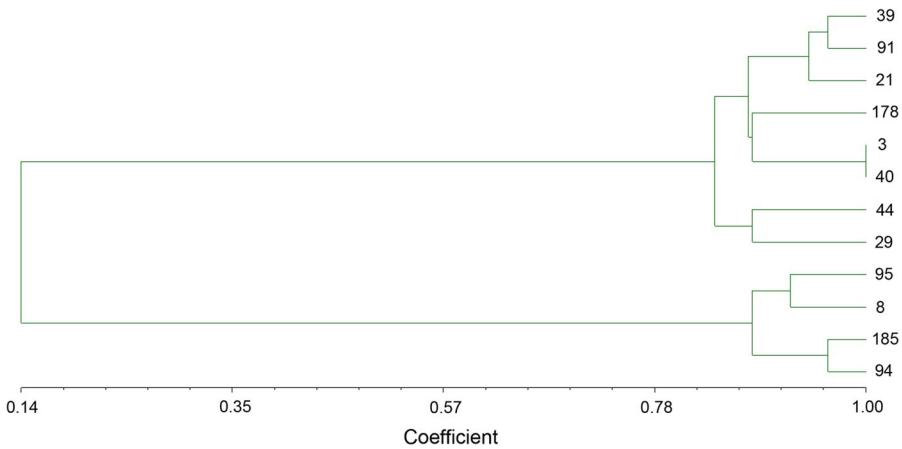

Fig. 4 Phylogenetic analysis of 12 pumpkin varieties. 39, 95, 178, 44, 185, 3, 40, 94, 8, 91, 29 and 21 represent 12 pumpkin varieties; of them, varieties 95, 8, 185 and 94 and *Cucurbita maxima* had a common ancestor and the others came from *Cucurbita moschata*

Discussion

Pumpkin is an important vegetable crop worldwide and is not only rich in nutrients, but also rich in medicinal ingredients such as carotenoids, beta-carotene and alphacarotene, which keep the immune system strong and the body healthy (Isutsa and Mallowa 2013). As we know, studies on pumpkin to date have mainly focussed on genetic breeding and functional analysis of bioactive components; there are few studies, however, on molecular biology. The lack of genomic data has seriously hampered research on the molecular biology of pumpkin. Due to the large and complex genome of pumpkin and the high cost of sequencing, it is very difficult to complete the genome sequencing. Luckily, transcriptome sequencing analysis based on next-generation sequencing technology with availability and low cost is being widely used (Trapnell et al. 2010, Grabherr et al. 2011). In this study, we made a comprehensive analysis of pumpkin transcriptome data, including data assembly, sequencing quality assessment and gene annotation, and we developed large numbers of SSR primers. This is the first exploration of the pumpkin transcriptome through the analysis of massive transcript data and the information will provide a basis for future studies on molecular biology, molecular breeding, physiology and biochemistry of pumpkin.

For sequencing and assembly, we obtained 52,649,316 clean reads with a mean length of 90 bp, which were assembled into 66,621 contigs and eventually produced 62,480 unigenes with an N50 of 1,215 bp. The average length of the unigenes was 765 bp, which was longer than that assembled in recent studies such as those on wax gourd (709 bp) (Jiang et al. 2013), sweet potato (581 bp) (Guo et al. 2011) and safflower (446 bp) (Huang et al. 2012). In addition, the average sequencing depth was up to 44.3-fold. The coverage depth of the unigenes was detected against orthologs of *Citrullus lanatus* and *Cucumis sativus*, and the match rates were both about 60 %. The results indicated that the quality of pumpkin

transcriptome sequencing and de novo assembly was very high.

For gene annotation, the CDS of unigenes were searched against the protein databases of Nr, SwissProt, KEGG and COG. About 68.51 % of the unigenes were able to uniquely match known proteins in these databases, suggesting that some genes may be unique to pumpkin. A large number were assigned into GO categories and COG classification, indicating that the transcriptomic data reflected the extensive diversity of pumpkin transcripts.

Regarding pathway annotation, the unigenes rich in the "Plant hormone signal transduction" pathway were the second largest group. This is our point of interest for future research, because our sequenced pumpkin material is a mutant insensitive to light and the phenotype formation of the mutant may be related to the regulation of many kinds of hormones.

Bud differentiation and flowering time are regulated by photoperiod-related genes (Fabbrini et al. 2012). In the model plant, *Arabidopsis*, we selected 12 reported photoperiod-related genes, viz. *LIGHT-REGULATED WD1* (*LWD1*) (AT1G12910.1), *GIGANTEA* (*GI*) (AT1G 22770.1), *FLOWERING BHLH3* (*FBH3*) (AT1G51 140.1), *ADAGIO 2* (*ADO2*) (AT2G18915.2), *FIONA1* (*FIO1*) (AT2G21070.3), *PHOTOPERIOD-INDEPENDENT EARLY FLOWERING1* (*PIE1*) (AT3G12810.1), *TIME FOR COFFEE* (*TIC*) (AT3G22380.1), *VERNALIZATION 5* (*VRN5*) (AT3G24440.1), *EARLY FLOWERING 6* (*ELF6*) (AT5G04240.1), *ARABIDOPSIS PSEUDO-RESPONSE REGULATOR 5* (*APRR5*) (AT5 G24470.1), *ARABIDOPSIS THALIANA FLOWERING PROMOTING FACTOR 1* (*ATFPF1*) (AT5G24860.1) and *ARABIDOPSIS THALIANA SWC6* (*ATSWC6*). The sequences of these 12 genes were BLASTed against the sequence data of all unigenes to look for orthologies and the results showed that eight genes (*LWD1*, *GI*, *FBH3*, *FIO1*, *PIE1*, *TIC*, *ELF6* and *APRR5*) had orthologies and another four genes (*ADO2*, *VRN5*, *ATFPF1* and *ATSWC6*) had no orthologies in pumpkin unigenes. LWD1/LWD2 are clock proteins involving in photoperiod control and the lwd1lwd2 double mutant with an early-flowering phenotype was contributed by a shift of CONSTANS (CO) and up-regulation of FLOWERING LOCUS T (FT) expression before dusk (Wu et al. 2008). Photoperiodic flowering and circadian rhythms controlled by the *GI* gene (Park et al. 1999) and loss of *GI* function causes late flowering and reduced transcription levels of *CO* and *FT* (Tseng et al. 2004). FBH3 was a transcriptional activator of CO and overexpression of the *FHB3* gene was able to increase *CO* level and caused early flowering, whereas, in the *fbh* mutants, *CO* levels were reduced (Ito et al. 2012). *FIO1* was essential for regulation of photoperiod length in the circadian clock of *Arabidopsis* and the *fio1* mutation caused early flowering and altered daylength-dependent flowering and hypocotyl growth (Kim et al. 2008). *PIE1* was an ISWI family gene whose function in FLOWERING LOCUS C (FLC) activation and floral repression in *Arabidopsis* and the *pie1* mutant caused early flowering in non-inductive photoperiods, independently of *FLC* (Noh and Amasino 2003). The *TIC* gene was a regulator of the circadian clock gene circuit and the *tic* mutant disrupted circadian gating, photoperiodism and circadian rhythms (Hall et al. 2003). *ELF6* acted as a repressor of photoperiod and mutations in *ELF6* genes in *Arabidopsis* led to brassinosteroid (BR)-related phenotypes and controlled flowering time (Noh et al. 2004, Yu et al. 2008). *APRR5* played an important role in circadian clock function and aberrant expression of the *APRR5* gene resulted in early flowering (Sato et al. 2002). In our study, *Cm-LWD1*, *Cm-FIO1*, *Cm-PIE1*, *Cm-TIC* and *Cm-ELF6* with higher transcriptional expression in (Photoperiod-insensitive) PPIS pumpkin plants may cause insensitivity to light and later flowering because the loss of function of these genes led to early flowering. However, *Cm-GI*, *Cm-FBH3* and *Cm-APRR5* with lower transcription expression in PPIS pumpkin plants may also cause later flowering and insensitivity to light because loss of the genes' function caused late flowering and reduced transcription levels of *CO* and *FT*.

SSRs contain DNA repeat sequences of 2–6 base pairs; they are a type of variable number tandem repeat and are typically co-dominant. They are extensively used as molecular markers in genetic mapping, pedigree and molecular breeding in crops (Temnykh et al. 2001, Senior et al. 1998, Neeraja et al. 2007). Development of SSR markers using traditional methods is expensive, and using high-throughput sequencing techniques to develop SSR markers is highly efficient, convenient and low in cost. In the sequences of pumpkin unigenes, we found 656 motif sequence types, far more than the 183 of wax gourd. In addition, a total of 57 pairs of SSR primers from the SSRs developed in pumpkin were used for phylogenetic analysis from 12 pumpkin varieties and obtained ideal results, which indicated that the SSR primers developed by us were very good for future research.

In conclusion, our transcriptome data will provide a very effective platform for pumpkin molecular studies in the future.

Materials and methods

Plant materials and growth conditions

The pumpkin plants were cultivated in a greenhouse set at 26 °C, with 15 h light (5,500 lux) and 9 h dark. The pumpkin is a short-day plant and most pumpkin varieties are photoperiod-sensitive, but the pumpkin variety used for RNA-seq was insensitive to photoperiod.

Tissue collection and RNA isolation for RNA-seq

The leaves and stems of PPIS pumpkin seedlings were harvested at the 2–4 true leaf stage after sowing and total RNA was isolated using TRIzol (Invitrogen) according to the manufacturer's instructions.

Preparation and sequencing

The experiment workflow is shown in Fig. 2a. In brief, RNA was collected and oligo(dT) beads were used to enrich poly(A) mRNA. The mRNA was chopped into short fragments using fragmentation buffer and these short fragments were used as templates to synthesize the first cDNA strand by random hexamers. The second cDNA strand was synthesized by using buffer, dNTPs, RNase H and DNA polymerase I. The paired-end library was synthesized using the Genomic Sample Prep kit (Illumina, Shenzhen, China), according to the manufacturer's instructions. Short fragments were purified with the QIAquick PCR extraction kit (Qiagen, Shanghai, China) and then resolved with EB buffer for end repair and the addition of poly(A). The short fragments were then connected with sequencing adapters and suitable fragments were selected as templates for PCR amplification by agarose gel electrophoresis. Finally, the sequencing library was built and sequenced using the Illumina HiSeq^{TM} 2000 by Guangzhou Gene Denovo Biological Technology Co., Ltd (Guangzhou, China).

Data assembly

Reads were assembled using Trinity (Grabherr et al. 2011). At first, transcriptome reads were cut into *k*-mers which were continuous DNA sequences, and then these high frequency *k*-mers were used to construct contigs according to overlaps and the longest assembled sequences containing the fewest *N*s were called contigs. Finally, sequences were obtained with the fewest *N*s which could not be extended at either end. Such sequences were defined as unigenes.

Bioinformatic analysis

The data assembly workflow is shown in Fig. 2c. The raw data were transformed by base calling into raw reads and adaptor fragments were removed from the raw reads to yield the clean reads. These short reads were assembled into unigenes and unigene sequences were aligned with @blastdb using BLASTX (E value < 0.00001). Sequence orientations were determined according to the best hit in the database. If results from different databases conflicted with each other, we chose one according to this priority: @blastdb. Orientation and CDS of sequences with no hit in BLAST were predicted using ESTScan (Iseli et al. 1999). Original transcript sequences (5′–3′) were provided if we could determine their orientations. Other sequences were provided as the assembler outputs. Functional annotations of unigenes included protein sequence similarity, KEGG pathway, COG and GO. We searched unigene sequences against protein databases (Nr, SwissProt, KEGG and COG) using BLASTX (E value < 0.00001). Protein function information was predicted from the annotation of the most similar protein in those databases. The KEGG pathway database records networks of molecular interactions in the cells, and variants of them specific to particular organisms. Pathway-based analysis helped to further understand the genes' biological functions. Pathway information of unigenes was obtained from KEGG pathway annotations. COG is a database where orthologous gene products are classified. Every protein in COG is assumed to have evolved from an ancestor protein, and the whole database is built on coding proteins with the complete genome as well as system evolutionary relationships of bacteria, algae and eukaryotes. Unigenes were aligned to the COG database to predict and classify possible functions. GO functional annotation was obtained with Nr annotation using the Blast2GO program (Conesa et al. 2005). Blast2GO has been cited by other articles more than 150 times and is a widely recognized GO annotation software. After obtaining GO annotation for every unigene, WEGO software (Ye et al. 2006) was used to perform GO

functional classification for all unigenes and to understand the distribution of gene functions of the species at the macro level.

RNA isolation and real-time quantitative PCR assay

Total RNA was extracted from PPIS and PPS pumpkin plants with 2–4 true leaves using TRIzol reagent (Invitrogen). Real-time quantitative PCR was performed as previously described (Wu et al. 2013).

Genomic DNA extraction and PCR test

Genomic DNA was extracted from the leaves of 12 pumpkin varieties with 2–4 true leaves using the CTAB (cetyltrimethylammonium bromide) method according to Murray and Thompson (1980). The names and codes of these varieties were *Cucurbita moschata*: Changqin No. 2 (39), Changqin No. 3 (91), Qingmi No. 1 (21), Guangmi No. 1 (178), Zaoxuan No. 1 (3), Zaoxuan No. 2 (40), Yuemi No. 1 (29) and Yuemi No. 2 (44); *Cucurbita maxima*: Hongmi No. 2 (95), Hongmi No. 1 (8), Puhong No. 2 (185) and Puhong No. 1 (94).

SSR discovery and validation

In total, 7,814 records were created. Primer modeling was successful for 4,794 sequences with SSR and failed for 3,020 sequences. Of the 4,794 pairs of SSR primers in pumpkin, 57 were used for phylogenetic analysis from 12 pumpkin varieties.

Acknowledgments We thank Guangdong Academy of Agricultural Sciences, Southern China Innovation Center for help in providing instruments and equipment. This work was supported by Guangdong Natural Science Foundation (No. S2012010010722), Guangdong Academy of Agricultural Sciences Dean Fund (No. 201203); Science and Technology Infrastructure Construction Project of Guangdong Key Laboratory for New Technology Research of Vegetables (Grant No. 2013112) and "948″ project from *Ministry of Agriculture of China* (2012-Z55).

References

Ando K, Carr KM, Grumet R (2012) Transcriptome analyses of early cucumber fruit growth identifies distinct gene modules associated with phases of development. BMC Genom 13:518

Blanca J, Esteras C, Ziarsolo P, Pérez D, Fernández-Pedrosa V, Collado C et al (2012) Transcriptome sequencing for SNP discovery across *Cucumis melo*. BMC Genom 13:280

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

Fabbrini F, Gaudet M, Bastien C, Zaina G, Harfouche A, Beritognolo I et al (2012) Phenotypic plasticity, QTL mapping and genomic characterization of bud set in black poplar. BMC Plant Biol 12:47

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González MV et al (2012) The genome of melon (*Cucumis melo L.*). Proc Natl Acad Sci USA 109:11872–11877

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Guo SG, Liu JG, Zheng Y, Huang MY, Zhang HY, Gong GY et al (2011) Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. BMC Genom 12:454

Guo SG, Zhang JG, Sun HH, Salse J, Lucas WJ, Zhang HY et al (2012) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet 45:51–58

Hall A, Bastow RM, Davis SJ, Hanano S, McWatters HG, Hibberd V (2003) The TIME FOR COFFEE gene maintains the amplitude and timing of Arabidopsis circadian clocks. Plant Cell 15:2719–2729

Huang SW, Li RQ, Zhang ZH, Li L, Gu XF, Fan W et al (2009) The genome of the cucumber, *Cucumis sativus L.* Nat Genet 41:1275–1281

Huang LL, Yang X, Sun P, Tong W, Hu SQ (2012) The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLoS ONE 7:e38653

Hyun TK, Rim Y, Jang HJ, Kim CH, Park J, Kumar R et al (2012) De novo transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis. Plant Mol Biol 79(4–5):413–427

Iseli, C., Jongeneel, C.V., 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

Isutsa DK, Mallowa SO (2013) Increasing leaf harvest intensity enhances edible leaf vegetable yields and decreases mature fruit yields in multi-purpose pumpkin. J Agric Biol Sci 8:610–615

Ito S, Song YH, Josephson-Day AR, Miller RJ, Breton G, Olmstead RG (2012) FLOWERING BHLH transcriptional activators control expression of the photoperiodic flowering regulator CONSTANS in Arabidopsis. Proc Natl Acad Sci USA 109:3582–3587

Jiang B, Xie DS, Liu WR, Peng QW, He XM (2013) *De Novo* assembly and characterization of the transcriptome, and development of SSR Markers in wax gourd (*Benicasa hispida*). PLoS ONE 8(8):e71054

Kim J, Kim Y, Yeom M, Kim JH, Nam HG (2008) FIONA1 is essential for regulating period length in the Arabidopsis circadian clock. Plant Cell 20:307–319

Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformation 24:713–714

Lu TT, Lu GJ, Fan DL, Zhu CR, Li W, Zhao Q et al (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20:1238–1249

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–264

Murray MG, Thompson WF (1980) Rapid isolation of high-molecular-weight plant DNA. Nucleic Acids Res 8:4321–4325

Neeraja CN, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard BC et al (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. Theor Appl Genet 115:767–776

Noh YS, Amasino RM (2003) PIE1, an ISWI family gene, is required for FLC activation and floral repression in Arabidopsis. Plant Cell 15:1671–1682

Noh B, Lee SH, Kim HJ, Yi G, Shin EA, Lee M et al (2004) Divergent roles of a pair of homologous Jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time. Plant Cell 16:2601–2613

Park DH, Somers DE, Kim YS, Choy YH, Lim HK, Soh MS et al (1999) Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene. Science 285:1579–1582

Sato E, Nakamichi N, Yamashino T, Mizuno T (2002) Aberrant expression of the *Arabidopsis* circadian-regulated *APRR5* gene belonging to the APRR1/TOC1 quintet results in early flowering and hypersensitiveness to light in early photomorphogenesis. Plant Cell Physiol 43:1374–1385

Senior ML, Murphy JP, Goodman MM, Stuber CW (1998) Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci 38:1088–1098

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, Lipovich L et al (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa L.*): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Tseng TS, Salomé PA, McClung CR, Olszewski NE (2004) SPINDLY and GIGANTEA interact and act in Arabidopsis thaliana pathways involved in light responses, flowering, and rhythms in cotyledon movements. Plant Cell 16:1550–1563

Wang P, Liu JC, Zhao QY (2002) Studies on nutrient composition and utilization of pumpkin fruit. J Inner Mongolia Agric Univ 23:52–54

Wang ZY, Fang BP, Chen JY, Zhang XJ, Luo ZX, Huang LF et al (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). BMC Genom 11:726

Wu JF, Wang Y, Wu SH (2008) Two new clock proteins, LWD1 and LWD2, regulate Arabidopsis photoperiodic flowering. Plant Physiol 148:948–959

Wu TQ, Tang DZ, Chen WD, Huang HH, Wang R, Chen YF (2013) Expression of antimicrobial peptides thanatin(S) in transgenic Arabidopsis enhanced resistance to phytopathogenic fungi and bacteria. Gene 527:235–242

Yadav M, Jain S, Tomar R, Prasad GBKS, Yadav H (2010) Medicinal and biological potential of pumpkin: an updated review. Nutr Res Rev 23:184–190

Ye J, Fang L, Zheng HK, 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 34:293–297

Yu X, Li L, Li L, Guo M, Chory J, Yin Y (2008) Modulation of brassinosteroid-regulated gene expression by Jumonji domain-containing proteins ELF6 and REF6 in Arabidopsis. Proc Natl Acad Sci USA 105:7618–7623

Zhang F, Jiang ZM, Zhang EM (2000) Pumpkin function properties and application in food industry. Sci Technol Food Indus 21:62–64

Zhang JN, Liang S, Duan JL, Wang J, Chen SL, Cheng ZS et al (2012) De novo assembly and characterization of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea L.*). BMC Genom 13:90