De novo characterization of the *Lycium ruthenicum* transcriptome and analysis of its digital gene expression profiles during fruit development and ripening

Yong Peng¹, Huiqin Ma² and Shangwu Chen¹,³*

¹ Beijing Laboratory for Food Quality and Safety, College of Food Sciences and Nutrition Engineering, China Agricultural University, Beijing, China
² College of Horticulture, China Agricultural University, Beijing, China
³ Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing, China

*Corresponding author: swchen@cau.edu.cn

Received: January 23, 2016; Revised: March 16, 2016; Accepted: April 4, 2016; Published online: November 11, 2016

Abstract: *Lycium ruthenicum* Murr., which belongs to the family Solanaceae, is a resource plant for Chinese traditional medicine and nutraceutical foods. In this study, RNA sequencing was applied to obtain raw reads of *L. ruthenicum* fruit at different stages of ripening, and a *de novo* assembly of its sequence was performed. Approximately 52.45 million 100-bp paired-end raw reads were generated from the samples by deep RNA-seq analysis. These short reads were assembled to obtain 164814 contigs, and the contigs were assembled into 84968 non-redundant unigenes using the Trinity method. Assembled sequences were annotated with gene descriptions, gene ontology, clusters of orthologous group and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway terms. Digital gene expression analysis was applied to compare gene-expression patterns at different fruit developmental stages. These results contribute to existing sequence resources for *Lycium* spp. during the fruit-ripening stages, which is valuable for further functional studies of genes involved in *L. ruthenicum* fruit nutraceutical quality.

Key words: *de novo* characterization; digital gene expression profiles; fruit development; *Lycium ruthenicum*; transcriptome

INTRODUCTION

*L. ruthenicum* Murr., from the genus *Lycium* of the family Solanaceae, is mainly distributed in the desert area of northwestern China and Central Asia. In China, it is mainly found in Qinghai, Ningxia and Xinjiang provinces. It is an ideal plant for preventing soil desertification and alleviating soil salinity and alkalinity because of its specific physiological characteristics of drought- and salt-resistance, which are very important for the ecosystem and agriculture in arid areas [1].

The fruit of *L. ruthenicum*, a Chinese traditional herb, is used in Tibetan medicine for the treatment of heart disease, abnormal menstruation and menopause, as well as a nutraceutical food [2]. Studies on *L. ruthenicum* have mainly focused on chemical components such as anthocyanins [3-5], essential oils [6], polysaccharides [7] and their pharmacological properties [8,9]. Recently, an increasing number of studies has aimed to reveal the genetic information of *L. ruthenicum* in shrub population and structure, sequence-related amplified polymorphism (SRAP) markers [10] and simple sequence repeats (SSRs) markers [11]. These researches have enriched the genetic knowledge of *L. ruthenicum*. SSRs are variable and ubiquitous in eukaryotic genomes, and are useful tools for evaluating genetic diversity, population structure and fingerprint mapping [12,13]. Recently, next-generation sequencing (NGS) technology was created as an innovative approach for high-throughput sequence determination and it has dramatically improved the speed and efficiency of gene discovery.

Peel color is a very important fruit characteristic that is mainly determined by the content of anthocyanins, which also confer beneficial properties to many fruits and vegetables [14]. *L. ruthenicum* contains abundant acylated anthocyanins such as petunidin diglucoside, petunidin glucoside and malvidin glu-
copyranoside [15]. The anthocyanin pathway is regulated by the MYB-bHLH-WD40 transcription factor complexes comprising of R2R3-MYB, basic-helix-loop-helix (bHLH) and WD-repeat (WDR) proteins. Meanwhile, the R3-MYB and R2R3-MYB repressors play a role in reducing anthocyanin production [16].

The aim of this study was to uncover transcriptome information for *L. ruthenicum* fruit and compare its digital gene expression (DGE) profiles at various developmental stages. Initially, we sequenced the transcriptome of *L. ruthenicum* fruit with Illumina sequencing technology. Furthermore, we compared the gene expression profiles of fruits among different fruit ripening stages by the digital gene expression system, aiming to characterize the transcriptome related to biosynthesis of anthocyanins of *L. ruthenicum*. The assembled, annotated transcriptome sequences and gene expression profiles provide valuable information for a better understanding of gene expression patterns at different fruit developmental stages of *L. ruthenicum*.

**MATERIALS AND METHODS**

**Plant material**

*L. ruthenicum* fruits were collected in a semi-wild, semi-desert area near Yinchuan, Ningxia province, China. The fruits were harvested at different developmental stages: the green fruit peel period (P1), initiation of peel color change (P2) and the fully mature, black peel period (P3) (Fig. 1). Collected fruits were cleaned, frozen by liquid nitrogen and stored at -80°C.

**RNA extraction and RNA sequencing**

Total RNA was extracted from the whole fruit by the CTAB method [17] and treated with DNaseI (Takara Biotech Incorporation) to remove contaminating DNA. The cDNA library for reference transcriptome was prepared by pooling the extracted RNA from the different fruit-ripening stages (P1, P2 and P3). Beijing Genome Institute (BGI, Shenzhen, China) operated the subsequent enrichment of mRNAs, fragment interruption, addition of adapters, PCR amplification and RNA sequencing according to the next-generation RNA-seq protocol and manufacturer’s instructions [18-20]. The quality and integrity of the RNA samples were examined using the Agilent 2100 Bioanalyzer, and samples with RNA integrity number (RIN) values higher than 7.0 were used. Poly (A) mRNA was isolated using magnetic beads (Illumina) bonding with oligo (d’T), the mRNA was fragmented, and cDNA was synthesized using random hexamer primers (Sangon Biotech) and reverse transcriptase (Invitrogen). The libraries were sequenced using the Illumina HiSeq 2000 platform after end-repair, adaptor ligation and PCR amplification. Libraries that gave reads unevenly distributed among the gene regions were discarded and replaced.

**Assembly and annotation of the reference transcriptome**

The raw sequencing reads were first filtered by removing invalid and low-quality reads, which included reads with adaptor contamination, with N percentages greater than 5% and reads with more than 50% bases with quality lower than Q20 level (an error probability of 1%) in one sequence. The short reads were first assembled into longer but gapless contigs using Trinity software [20], and the contigs were further assembled to obtain the unigenes that could not be extended on either end.

We used Blast2GO to obtain gene ontology (GO) annotations of the unigenes [21] and classified the unigenes’ GO function by using WEGO software [22] and connecting some unigenes with metabolic pathways using the KEGG database [23] (E-value ≤1.0E-5). The sequence direction of the unigenes was determined according to the best alignment results. When the results were conflicting among databases, the direction was determined successively by the NCBI non-redundant protein sequence database (NR), Swiss-Prot, KEGG and clusters of orthologous groups (COG). When a unigene would not align to any database, we applied ESTScan to predict coding regions and determine sequence direction [24]. SSR was screened by the
SSR detecting program MicroSatellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/misa.html). The criteria for SSR identification were as follows: 12 nt for mono-nucleotide, di-nucleotide and tri-nucleotide SSRs; 16 nt for tetra-nucleotide SSRs; 20 nt for penta-nucleotide SSRs; and 24 nt for hexa-nucleotide SSRs.

**Preparation and sequencing of DGE library**

After extracting total RNAs from the samples, mRNAs of samples were enriched by using the oligo (dT) magnetic beads (Illumina) from the total RNAs. By using the fragmentation buffer (Life Technologies), the mRNAs were fragmented into short fragments (about 200 bp), then the first-strand cDNAs were synthesized by random hexamer-primer (Sangon Biotech) using the mRNA fragments as templates. Buffer, dNTPs, RNase H and DNA polymerase I (Invitrogen) were added to synthesize the second strand. The double-strand cDNAs were purified with a QiaQuick PCR extraction kit and washed with ethidium bromide (0.5 μg/ml) buffer for end repair and poly(A) addition. Finally, sequencing adaptors were ligated to the fragments. The fragments were purified by agarose gel electrophoresis and enriched by PCR amplification using adapters as primers. The library products were sequenced using the Illumina HiSeq 2000 platform.

**DGE analysis and mapping**

The original image data were transferred into sequence data via base calling. Filtering and quality control checks (base composition analysis, per base sequence quality and per sequence quality scores) were performed using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). We filtered the reads with adapters, with unknown bases more than 5% and low-quality reads with more than 50% of bases with quality lower than Q20 level.

Clean reads were mapped to reference sequences by using the SOAP2 [25] aligner. No more than 2 mismatches were allowed in the alignment. Quality control of alignment included the statistics of alignment analysis, sequencing saturation analysis, distribution of reads on reference genes analysis and coverage analysis. For gene expression analysis, we used the RPKM (Reads Per Kilo base of exon model per Million mapped reads) [26] method to calculate the gene expression level. Based on gene expression, we selected differentially expressed genes (DEGs) and carried out further functional analysis.

**Evaluation of DGEs**

We analyzed the expression levels to compare the gene expression in three different fruit ripening periods. We used “false discovery rate (FDR) ≤0.001 [27] and the absolute value of log2 Ratio >1” as the threshold to judge the significance of the differences in gene expression. In the pathway analysis, we mapped the differentially expressed genes in the KEGG database and identified significantly enriched genes related to metabolic or signal-transduction pathways.

**RESULTS**

**Sequencing and assembly of clean reads**

We obtained an overview of the gene-expression profiles of *L. ruthenicum* fruit at different stages of ripening. Approximately 52.45 million 100-bp paired-end raw reads were generated from the *L. ruthenicum* fruit by deep RNA-seq analysis. After filtering out the repetitive, low-complexity and low-quality reads, we obtained 51.39 million clean 100-bp reads. These short reads were assembled into 164814 contigs ranging from 200 to over 3000 bp, with a mean size of 323 bp. Q20 level and GC content were 99.01% and 45.78%, respectively. The contigs were assembled into 84968 non-redundant unigenes with a mean size of 610 bp using the Trinity method for paired-end joining and gap-filling. The 84968 unigenes’ total consensus sequences were divided into 9377 distinct clusters and 64432 distinct singletons after clustering.

**Functional annotation of unigenes using public databases**

The functions of a protein can be predicted from annotation of known proteins in the NCBI NR, Swiss-Prot and KEGG databases. We obtained 33868 unigenes after mapping unigene sequences to the Swiss-Prot database (Table 1). BLASTx was used to annotate the
sample’s unigenes against the NCBI NR nucleotide database with an E-value of less than 1.0E-5, yielding 50880 genes. The E-value distribution statistics revealed that 46.4% of the mapped sequences have strong homology (smaller than 1.0E-50), and 53.6% of the sequences have homology ranging between 1.0E-5 and 1.0E-60. The similarity distribution showed that 63.6% of the sequences have higher than 80% similarity, and the highest percentage of unigenes mapped to *Solanum tuberosum* sequences (Table 2).

The unigenes were annotated with Enzyme Commission (EC) numbers from BLASTx alignments against the KEGG database (E-value ≤1.0E-5) to determine the biological pathways in *L. ruthenicum* fruit. KEGG analysis assigned 26269 unigenes to approximately 38635 items and 128 known metabolic or signaling pathways, of which 6330 (24.1%) unigenes were related to metabolic pathways, 3380 (12.87%) to biosynthesis of secondary metabolites, 1479 (5.63%) to plant-pathogen interactions, 1404 (5.34%) to RNA transport, 1304 (4.96%) to plant hormone signal transduction and 497 (including 420 individual unigenes) to flavonoid-biosynthesis pathway categories, closely related to flavonoid, flavone, flavonol, iso flavonoid and anthocyanin biosynthesis for nutraceutical quality formation.
We aligned the unigenes to the COG database to predict their possible functions and analyzed their functional classification. A total of 17133 unigenes were assigned to 25 COG classification groups (Fig. 2). Among these groups, the “general function prediction only” cluster (5019; 29.29%) ranked as the largest group, followed by “transcription” (2552; 14.9%), “posttranslational modification, protein turnover, chaperones” (2524; 14.73%) and “replication, recombination and repair” (2458; 14.34%). Secondary metabolites and related transporters were run in the third group. However, only a few unigenes were matched to the “nuclear structure” (4, 0.02%) and “extracellular structures” (21; 0.12%) categories. In addition, 1588 (9.27%) unigenes were related to “function unknown” and could be undiscovered genes in *L. ruthenicum*.

**GO and COG classification**

We aligned the unigenes to the COG database to predict their possible functions and analyzed their functional classification. A total of 17133 unigenes were assigned to 25 COG classification groups (Fig. 2). Among these groups, the “general function prediction only” cluster (5019; 29.29%) ranked as the largest group, followed by “transcription” (2552; 14.9%), “posttranslational modification, protein turnover, chaperones” (2524; 14.73%) and “replication, recombination and repair” (2458; 14.34%). Secondary metabolites and related transporters were run in the third group. However, only a few unigenes were matched to the “nuclear structure” (4, 0.02%) and “extracellular structures” (21; 0.12%) categories. In addition, 1588 (9.27%) unigenes were related to “function unknown” and could be undiscovered genes in *L. ruthenicum*.

**GO is the system for sorting gene descriptions according to three ontology categories and functional subcategories.** A total of 30874 unigenes of the
A total of 5172 SSRs were identified in 4653 unigenes. The frequency and distribution of SSRs are shown in Figure 4. The trinucleotide repeats were the most abundant (38.5%) followed by mononucleotide (31.11%) and dinucleotide repeats (26.84%) in various groups of SSRs. The hexanucleotide (1.82%), pentanucleotide (0.95%) and tetratricopeptide (0.79%) repeats represented a smaller proportion of the motif classes. Among the dinucleotide repeats, AG/CT was the most abundant (59.8%) followed by AC/GT (25.07%) and AT/AT (14.7%). The trinucleotide repeats, including AAG/CTT, AAC/GTT, ATC/ATG and ACC/GGT, were predominant in this part (Fig. 4).

Sequences mapping to the reference transcriptome database

We mapped the tag sequences of the three fruit-stage libraries to the fruit-reference library of *L. ruthenicum* with the DGE method. Approximately 72544, 68100 and 75212 unigenes were mapped to the reference database in the P1, P2 and P3 fruit stages, respectively. For all fruit-stage DGEs, the number of detected genes vs. the sequenced read amount were displayed as a saturation function curve, when the clean reads number reached above 10 million. In the 20 most abundantly expressed genes of the P3 stage of *L. ruthenicum*, genes such as cytochrome P450, 2S sulfur-rich seed storage protein, lipid transfer protein LTP1 precursor, flower-specific defensin, chitin-binding lectin 1 and pathogenesis-related proteins were expressed in all three fruit stages.

Differential expression profiles at the different fruit stages

To obtain the digital expression profile of the DEGs among the three fruit development stages of *L. ruthenicum* we used the RPKM method to perform gene expression analysis pairwise among the P1, P2 and P3 libraries. In the “P2 vs. P1,” “P3 vs. P2,” and “P3 vs. P1” comparisons we identified 9479, 15953 and 10344 DEGs, respectively, under FDR ≤0.001 and absolute value of log2-ratio >1 (Fig. 5). In total, 6761 upregulated and 2718 downregulated genes were detected in P2 vs. P1, 6178 upregulated and 9775 downregulated genes in P3 vs. P1, and 1760 upregulated and 8584 downregulated genes in P3 vs. P2 (Table 3). In addition, 2943, 286 and 6155 genes were uniquely expressed in the P1, P2 and P3 stages, respectively, while 62232 genes were expressed during all three stages.

The differential expression profiles obtained by pairwise comparison were further delineated by analyzing the DEGs for enriched GO terms on GO on-
tology domains of BP, CC and MF and for enriched KEGG pathway ontology (KO) terms illustrating the fruit’s developmental changes. For fruit stages P2 vs. P1, a total 156 functional catalogs were found changed for GO terms by DEGs, including 81 in BP, 41 in CC and 34 in MF. For P3 vs. P1 fruit stages, a total of 246 functional groups were found changed, and in the P3 vs. P2 comparison, 333 functional groups changes were found, including 188, 79 and 66 groups for BP, CC and MF, respectively. The parallel DEGs signed by KO terms for enriched KEGG pathway annotation with pairwise comparisons of the three fruit ripening stages of *L. ruthenicum* indicated significant changes in biosynthesis of secondary metabolites and metabolic pathways during fruit development and ripening.

A total of 458 DEGs related to “flavonoid biosynthesis” (126), “phenylpropanoid biosynthesis” (156) and “starch and sucrose metabolism” (176) categories were detected during the ripening stages according to gene annotation and biosynthesis pathway analysis (P ≤0.05). Hierarchical cluster analysis of the DEGs at the different ripening stages was performed to examine the similarity and diversity of gene expression. We clustered the three categories of DEGs mentioned above of the different fruit ripening stages. The expression of genes related to flavonoid biosynthesis was downregulated in the P2 stage and upregulated in the P3 stage (Fig. 6A). In addition, the overall trend in expression level of genes related to phenylpropanoid biosynthesis was ascending (Fig. 6B), while the overall trend in expression level of genes related to starch and sucrose metabolism successively declined during the fruit ripening (Fig. 6C).

The expression levels of leucoanthocyanidin dioxygenase, flavonoid 3-glucosyltransferase precursor, anthocyanidin 3-O-glucoside 5-O-glucosyltransferase and flavonoid 3’,5’-hydroxylase [28,29] (Fig. 7A-D), genes related to pigment formation and color, were considerably higher at the fully ripe stage. These increments were positively correlated with the expression of R2R3-Myb [30,31] (Fig. 7 E), the plant transcrip-
tion factor which is highly correlated to the anthocyanin accumulation, while the expression of translation elongation factor 1, used as a housekeeping gene, was nearly equal during the all three stages (Fig. 7 F).

**DISCUSSION**

Next-generation sequencing technology has emerged as a timesaving approach in genetic analysis, comparative genomics and functional genomics of non-model organisms. The genome information on *L. ruthenicum* fruit or related species was absent. In this study, the RNA-seq data from different ripening stages of *L. ruthenicum* fruit provides a comprehensive approach for DEG searches and anthocyanin synthesis molecular mechanism researching.

The annotation results imply that the Illumina HiSeq 2000 sequencing project in this study produced a substantial number of assembled transcripts of *L. ruthenicum* fruits. The species distribution returned from the NR database demonstrates that the transcriptome of *L. ruthenicum* is more closely related to that of *Solanum tuberosum* than to other plant genomes present in current public databases.

Overall, “biosynthesis of secondary metabolites” accounted for almost half of unigenes related to metabolic pathways in KEGG data mapping, suggesting that the secondary metabolic processes are active pathways in *L. ruthenicum* fruit development. It is noticeable that 903 unigenes (5.27%) were classified into the group of “secondary metabolite biosynthesis, transport and catabolism” in COG categories, implying that those secondary metabolite processes play crucial roles in *L. ruthenicum* fruit development.

Among the dinucleotide SSRs, the most abundant motif was GA/CT among the dinucleotide repeats SSRs, similar to the situation in a broad array of species such as *Capsicum annuum* [32,33], *Solanum melongena* [13], wheat [34] and barley [35]. AAG/CTT was the most common trimeric motif present in *L. ruthenicum*, a situation that almost matches that in peppers (*Capsicum*) [36], tomato [37], peanut [38] and radish [39]. The SSRs recognized in the present research provide an abundant resource for *L. ruthenicum* molecular marker studies.

Based on the DGE data of *L. ruthenicum*, multiple genes exhibited different expression patterns in different ripening stages. A total 2943, 286 and 6155 genes were uniquely expressed in stages P1, P2 and P3, respectively, implying that in the fully ripe stage the expression level of genes is higher than in previous stages.

In *L. ruthenicum*, nearly all the structural and regulatory genes involved in anthocyanin pathway were significantly upregulated at the fully ripe stage [29]. The same trend was documented in this study, where the anthocyanin accumulation-related genes were increasingly expressed during fruit ripening, reaching significantly increased expression at full ripe stage.

Fruit skin coloration is a unique pattern in the life cycle of fruiting plants and is mainly attributed to anthocyanin pigments [40]. The content of anthocyanins and color of the fruit skin are related; the increase in anthocyanin biosynthesis-related gene expression causes the anthocyanin accumulation and results in
fruit color deepening. In an anthocyanin biosynthesis-related study, anthocyanin profiling confirmed that anthocyanins were increasingly accumulated during fruit ripening in *L. ruthenicum*, and sharply increased in fully expanded mature fruit [29]. On the other hand, for the starch and sucrose metabolism-related DEGs, the gene expression decreased during *L. ruthenicum* fruit development. This suggests that during fruit ripening sucrose synthesis shows a trend of initial increase followed by a decrease [41] and would be reduced in fully ripened fruit.

**Availability of supporting data**

All clean reads generated by Illumina sequencing have been deposited in the Sequence Read Archive (SRA) data base (http://www.ncbi.nlm.nih.gov/sra) under the accession no. ID SRP071022.

**Acknowledgments:** We thank the Agriculture and Animal Husbandry Bureau of Yinchuan City for providing *L. ruthenicum* fruit samples.

**Authors’ contribution:** PY, MHQ and CSW conceived and designed the experiments, contributed reagents/materials/analysis tools and wrote the paper. PY performed the experiments and analyzed the data.

**Conflict of interest disclosure:** The authors declare that no competing interests exist.

**REFERENCES**