**De novo** gene set assembly of the transcriptome of diploid, oilseed-crop species *Perilla citriodora*

Ji-Eun Kim · Junkyoung Choe · Woo Kyung Lee · Sangmi Kim · Myoung Hee Lee · Tae-Ho Kim · Sung-Hwan Jo · Jeong Hee Lee

Received: 24 August 2016 / Revised: 24 August 2016 / Accepted: 12 September 2016

© Korean Society for Plant Biotechnology

**Abstract** High-quality gene sets are necessary for functional research of genes. Although Perilla is a commonly cultivated oil crop and vegetable crop in Southeast Asia, the quality of its available gene set is insufficient. To construct a high-quality Perilla gene set, we sequenced mRNAs extracted from different tissues of *Perilla citriodora*, the wild species (2n = 20) of Perilla. To make a high-quality gene set for *P. citriodora*, we compared the quality of assemblies produced by Velvet and Trinity, the two well-known **de novo** assemblers, and improved the **de novo** assembly pipeline by optimizing *k*-mers and removing redundant sequences. We then selected representative transcripts for loci according to several criteria. The improved assembly yielded a total of 86,396 transcripts and 38,413 representative transcripts. We evaluated the assembled transcripts by comparing them to 638 homologous *Arabidopsis* genes involved in fatty acid and TAG biosynthesis pathways. High proportions of full-length genes and transcripts in the assembled transcripts matched known genes in other species, indicating that the *P. citriodora* gene set can be applied in future functional studies. Our study provides a reference *P. citriodora* gene set for further studies. It will serve as valuable genetic resource to elucidate the molecular basis of various metabolisms.

**Keywords** *Perilla citriodora*, Transcriptome, **de novo** assembly, gene set, fatty acid biosynthesis, TAG (triacylglyceride) biosynthesis metabolic pathway, oilseed crop

**Introduction**

The identification of an organism’s full gene set plays an important role as the foundation for comprehensive genomic and transcriptome studies of the organism (Martin and Wang. 2011). Only a few years ago, our understanding of the transcriptome was dependent on traditional methods such as low-throughput expressed sequence tag (EST)-based or chip-based methods (e.g., DNA microarray), which have several limitations (Wang et al. 2009). Recently, whole-transcriptome profiling based on next-generation sequencing (NGS) has started to provide insights into the large-scale landscape and dynamics of the complex world of the transcriptome (Chen et al. 2011a).

The RNA-sequencing (RNA-seq) approach and powerful bioinformatic tools provide more quantitative and precise measurements of gene expression levels than earlier methods (Marguerat and Bahler et al. 2010), with highly reproducible results and few systematic differences among technical replicates (Zhang et al. 2012). Moreover, many of the latest studies have applied RNA-seq to various biological purposes including the identification of all expressed transcripts (Li et al. 2014), the improvement of genome assembly to find missing information and better understand the structure of the reference genome (Chen et al. 2011b), the detection of novel transcribed regions and alternatively spliced forms (Garber et al. 2011), SNP marker discovery (Iorizzo et al. 2011), and others, both with and without a reference genome (Gongora-Castillo and Buell. 2013). Therefore, it is urgent to gain **de novo** assembled transcript sequences from organisms for which high-quality, assembled reference genomes are not yet available (Chen et al. 2011a). Information about the
transcriptome and genome of the Perilla plant has been difficult to obtain, while interest in the plant has begun to increase (Fukushima et al. 2015).

Perilla, an oilseed crop belonging to the Lamiaceae family, is commonly cultivated in Asian countries such as Korea, Japan, China, and Nepal. There are commonly considered to be four Perilla species and one variety (Jung et al. 2005). *Perilla citriodora* is the wild species (2n = 20) and possibly the ancestor of the tetraploid Perilla species (2n = 40) (Ito et al. 2000). Wild Perilla species were first found in Japan and China. In 2003, the Jeju-17 collection became the first reported example of *P. citriodora* on Jeju island (Jung et al. 2005).

Perilla seeds contain large amounts of unsaturated fatty acids, conferring various benefits to human health (Bumblauskiene et al. 2009). Hence, most Perilla studies to date have focused on the characterization and biological activities of the Perilla metabolites. The genes involved in the biosynthesis of anthocyanins, flavones, and monoterpenoids have been described (Lee et al. 2014). Various genomic, transcriptomic, and molecular analyses of the cultivated species *Perilla frutescens*, which is mainly used as an oil crop, have been performed; however, such fundamental analyses are lacking for *P. citriodora*, probably because of insufficient genetic resources (Lee et al. 2014). Considering the importance of Perilla to various nutritional and industrial applications, the construction of molecular materials (e.g., genomes and genes) for further studies of *P. citriodora* is essential.

In this study, we report the production of a high-quality gene set of *P. citriodora* using a modified de novo assembly pipeline. We evaluated the assembled transcripts with 638 homologous Arabidopsis genes involved in fatty acid and triacylglyceride (TAG) biosynthesis pathways, checking the matched gene coverage and proportion of full-length genes among the assembled transcripts. Our research provides useful information and a basis for future genetic studies of Perilla.

**Materials and Methods**

Sample preparation and RNA extraction

*P. citriodora* plants were cultivated at the National Institute of Crop Science, RDA in Miryang. Tissue samples from leaves, buds, inflorescence before and after fertilization, and seeds after 4 weeks of development after flowering were collected, immediately frozen in liquid nitrogen, and stored at -80°C prior to RNA extraction. Total RNA from each sample was extracted using TRIzol Reagent (Invitrogen) and treated with DNase I (Takara) according to the manufacturer’s instructions. The RNA quality was examined using 1% agarose gel, and the concentration was determined using a Nanodrop spectrophotometer (Thermo). Each RNA pool was prepared by mixing equal amounts of three RNA-extraction replicates.

CdNA library construction and mRNA sequencing

Starting with the total RNA from each of the five samples, mRNA was purified using poly(A) selection or rRNA depletion and then chemically fragmented and converted into single-stranded cdNA using random hexamer priming. Next, the second strand was generated to create double-stranded cdNA. The library construction began with the generation of blunt-end cdNA fragments from the double-stranded cdNA. Then, an ‘A’ base was added to the blunt end in order to make the fragments ready for ligation to sequencing adapters. After the size selection of ligates, the ligated cdNA fragments that contained adapter sequences were enhanced via PCR using adapter-specific primers. Paired-end libraries were prepared using the Illumina TruSeq RNA Sample Preparation Kit v2 (catalog #RS-122-2001, Illumina, San Diego, CA). Then, the libraries were quantified using the KAPA library quantification kit (Kapa Biosystems KK4854) following the manufacturer’s instructions. Each library was then loaded onto the Illumina Hiseq2000 platform (100-bp paired-end reads). High-throughput sequencing was performed to ensure that each sample met the desired average sequencing depth. The sequence data from the leaf tissue has been deposited to the National Agricultural Biotechnology Information Center (NABIC) Sequence Read Archive (SRA) with the accession number NN-1473-000001.

*P. citriodora* transcriptome de novo assembly

Before assembly, the raw sequence reads with sufficient quality [Phred score (Q) ≥ 20] were selected. Reads < 25 bp in length were then removed using the SolexaQA package (v1.13) (Cox et al. 2010). The remaining high-quality reads were then used for de novo assembly with the Velvet (v1.2.07; http://www.ebi.ac.uk/zerbino/velvet/) (Zerbino and Birney. 2008) and Trinity (trinityrnaseq-2.2.0; http://trinityrnaseq.sourceforge.net/) (Grabherr et al. 2011) assembler tools, both of which are based on the de Bruijn graph algorithm. Similar parameters were used for each assembler to keep the same conditions so that the performance of the assemblers could be compared.
The de Bruijn-based assemblers have two important parameters: the \textit{k}-mer (hash length by which a read is divided) and the coverage cutoff (Chen et al. 2014; Kim et al. 2015). Only one \textit{k}-mer length, 25 (the default parameter provided by the author of the program), was employed in Trinity. Because different \textit{k}-mer lengths generate different assembly results, multiple hash lengths (51, 59, 61, 63, 67, 69, 71, 73, 75, 77 and 83) were tested during the Velvet assembly. After the optimal \textit{k}-mer for \textit{P. citriodora} was selected, the Velvet assembly was carried out again with the optimal \textit{k}-mer length. To improve the performance of the Velvet assembly with the optimal \textit{k}-mer length, we used the Oases software (v0.2.08; http://www.ebi.ac.uk/zerbino/oases/) (Schulz et al. 2012) with a modified assembly method. The assembly proceeded according to the following steps. First, the draft contig sequences obtained with each \textit{k}-mer length in the primary Velvet assemblies were joined to create extended contigs using the Oases software. Then, the two best \textit{k}-mer lengths were selected based on the number and length of the assembled contig sets. After an error-correction step with ‘n’ sequence split, a re-assembly was carried out with the two primary sets of assembled contigs made with the two optimal \textit{k}-mer lengths. Like the primary assembly steps, the re-assembly was performed with Velvet, and Oases was applied to the selected optimal \textit{k}-mer. A self-BLAST was performed to remove redundancy from the output. Finally, a gene set was made for the \textit{P. citriodora} transcriptome. The total transcripts comprised alternatively spliced transcripts. We selected the representative transcripts among the many alternatively spliced transcripts.

The assembled transcripts were designated as ‘P.citriodora XSLXXXXXXXtXXX’. ‘P.citriodora’ is an abbreviation of the species name. The first digit ‘X’ following ‘P.citriodora’ denotes the assembly version, and ‘S’ means a reference from SEEDERS. The ‘L’ is an abbreviation for “locus,” and the following six digits signify the representative transcript (locus) number. The ‘t’ is an abbreviation for “transcript,” and the last three digits signify the transcript number for the locus.

Evaluation of the assembly quality with genes involved in fatty acid and TAG biosynthesis

The assembled \textit{P. citriodora} transcripts were searched against 638 Arabidopsis genes (amino acid sequences) involved in fatty acid and TAG biosynthesis (Bates et al. 2014) using TBLASTX with minimum cutoffs of \textit{e-value} \textless{} 1e-10 and sequence identity \textgreater{} 50. The quality of the assembled transcript sets was evaluated by the number of positions in the corresponding Arabidopsis genes that were covered by the assembled transcripts and proportion of full-length protein-coding genes to total assembled transcripts according to the Arabidopsis gene coverage. We determined the numbers of transcripts in each set that displayed 70%, 80%, and 90% coverage, respectively, of corresponding Arabidopsis genes. We measured the proportion of predicted full-length protein-coding genes in the major fatty acid and TAG biosynthesis pathways to the assembled transcripts using the following criteria: (1) the predicted open reading frames (ORFs) should contain basic gene features such as a start codon and a stop codon (otherwise, the sequences were assumed to be partial genes); (2) the transcripts should cover more than 85% of a matching region in an Arabidopsis gene (Kim and Chen. 2015).

Functional annotation of the \textit{P. citriodora} gene set

To functionally annotate the assembled transcripts, we assessed their sequence similarity with the \textit{Sesamum indicum} protein sequences from NCBI and with 40 known plants in the Phytozome database (version 9.1) (http://www.phytozome.net/), respectively. We selected the best-matched transcripts with \textit{e-value} < 1e-10 from BLASTX. To identify the putative functions of the \textit{P. citriodora} transcripts, we performed functional enrichment analysis using the Gene Ontology (GO) database and BLASTX (\textit{e-value} \textless{} 1e-30). The parameters were set to a depth of two in the ontology hierarchy and a hit threshold of five (transcripts counts), and the output was sorted by the hit count. The GO consist of terms that provide a global representation of gene function with a controlled vocabulary including the three GO categories: biological processes, cellular components, and molecular functions (Harris et al. 2014). We also screened the transcripts against the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway database via BLASTP (\textit{e-value} \textless{} 1e-30 and identity \textless{} 70) and identified the biological mechanisms and metabolic pathways corresponding to the identified enzyme commission numbers (Kanehisa and Goto. 2000).

Results and Discussion

Transcriptome resource for \textit{P. citriodora}

We obtained a total of 193,666,492 paired-end reads (19,366,649,200 bp) from the five RNA-seq experiments (Table 1). After strictly filtering by Phred quality score and read length, we obtained 173,911,506 (89.61%) cleaned reads with a total length of 15,584,692,752 bp (80.44%) (Table 1).
Table 1 Statistics of the short-read sequencing of Perilla

| Sample description | Raw reads | Cleaned reads |
|--------------------|-----------|---------------|
|                    | Num. of reads | Total length (bp) | Num. of reads | Avg. length | Total length (bp) | % |
| Leaves             | 21,255,739  | 2,125,573,900 | 19,314,024  | 90.28 | 1,743,658,415 | 82.03% |
|                    | 21,255,739  | 2,125,573,900 | 19,314,024  | 88.62 | 1,711,683,894 | 80.53% |
| Buds               | 20,120,794  | 2,012,079,400 | 18,064,027  | 91.21 | 1,647,650,051 | 81.89% |
|                    | 20,120,794  | 2,012,079,400 | 18,064,027  | 87.71 | 1,584,428,254 | 78.75% |
| Inflorescence (before fertilization) | 19,305,406 | 1,930,540,600 | 16,972,378  | 90.48 | 1,535,591,827 | 79.54% |
|                    | 19,305,406 | 1,930,540,600 | 16,972,378  | 87.26 | 1,480,965,843 | 76.71% |
| Inflorescence (after fertilization) | 17,354,075 | 1,735,407,500 | 15,393,832  | 91.11 | 1,402,454,377 | 80.81% |
|                    | 17,354,075 | 1,735,407,500 | 15,393,832  | 87.07 | 1,340,362,115 | 77.24% |
| Seeds              | 18,797,232  | 1,879,723,200 | 17,211,492  | 93.11 | 1,535,353,132 | 81.68% |
|                    | 18,797,232  | 1,879,723,200 | 17,211,492  | 89.21 | 1,535,353,132 | 81.68% |
| Total              | 193,666,492 | 19,366,649,200 | 173,911,506 | 89.61 | 15,584,692,752 | 80.44% |

%: Percentage of the total length of raw reads represented by the total length of the cleaned reads

Table 2 Summary of de novo assembled transcripts of P. citriodora through RNA-seq

| Assembler         | Transcript type | Total count | Total length (bp) | N50 | AVG |
|-------------------|-----------------|-------------|-------------------|-----|-----|
| Improved assembly | Total transcripts | 86,396 | 155,964,376 | 2,675 | 1,805 |
|                   | Representative transcripts | 38,413 | 49,116,021 | 2,233 | 1,278 |
| Velvet assembly   | Total transcripts | 101,855 | 155,276,941 | 2,167 | 1,524 |
|                   | Loci of transcripts | 40,455 | 40,783,832 | 1,651 | 1,008 |
| Trinity assembly  | Total transcripts | 143,535 | 219,451,507 | 2,437 | 1,528 |
|                   | Loci of transcripts | 46,615 | 39,245,806 | 1,564 | 841 |

N50: a weighted median statistic length such that 50% of the assembled transcripts are longer than the N50 length

AVG: average length of all assembled transcripts

Improving the de novo transcriptome assembly for the P. citriodora gene set

To obtain a high-quality gene set for P. citriodora, we performed de novo assembly with the Trinity and Velvet assembly programs and compared the results (Table 2). We measured the number and length (total bases, N50, maximum and average length) of the assembled transcripts as indicators of the assembly performance. Trinity produced 143,535 transcripts with a total length of 219,451,507 bp using only one k-mer of 25. The Velvet outcomes varied, ranging from 95,416 to 46,373 in the number of transcripts and from 154,039,014 bp to 59,012,466 bp in the total length with multiple hash lengths. Trinity produced a longer assembly than Velvet with the optimal hash length of 67 (N50 = 2,437 bp vs. N50 = 2,167 bp). Furthermore, Trinity produced about 40,000 more transcripts than Velvet (143,535 vs. 101,855). Both programs produced many transcripts in the range of 201–500 bp (56.1% in Trinity and 41.9% in Velvet; Fig. 1). We performed scaffolding using Oases, which is compatible with Velvet, to improve the Velvet assembly quality.

Among the hash lengths used (k-mer = 51–83) in the primary assembly with Velvet and Oases, we selected k-mer = 67 (72,283 transcripts) and k-mer = 69 (69,310 transcripts) as the optimal hash lengths based on the number and lengths of the assembled contigs. We used the outcomes of those two primary assemblies for the re-assembly. Meanwhile, Oases filled each position between reads or contigs with ‘n’ for scaffolding. In the re-assembly, Velvet converts repeated ‘n’ sequences to artificial poly ‘a’ sequences. Because the converted ‘a’ sequences could bring problems in the next assembly steps, we split the assembled sequences at the ‘n’ sequences, deleted the ‘n’ sequences, and excluded the fragments shorter than 100 bp from the primary set of assembled contigs. Through those steps, we refined the two primary contig sets from 72,283 contigs to 71,642 contigs (k-mer = 67) and from 69,310 contigs to 68,783 contigs (k-mer = 69), respectively. We performed the re-assembly by combining the two primary contig sets and using the optimal k-mer of 67 to generate one transcript set containing
86,654 transcripts. We then filtered out 258 redundant transcripts by the self-BLAST (Ness et al. 2011). As a result, a total of 86,396 transcripts were built with an N50 length of 2,675 bp and an average length of 1,805 bp (Table 2). Thus, the final transcript set produced by the re-assembly was 40% smaller than that produced by the Trinity assembly and 16% smaller than that produced by the Velvet assembly.

Selection of representative transcripts

The assembled transcripts could be clustered to loci, with an average of 2.2 transcripts per locus (Table 2). An optimized representative transcript for each locus should be recommended for downstream analyses such as gene prediction and gene expression analysis. Velvet produced 40,455 clusters, and Trinity produced 46,615 clusters, but neither assembly program selects representative transcripts for loci. Therefore, we selected representative transcripts via the following steps. We mapped the short reads used in the final assembly back to their respective transcripts using the Bowtie software (Langmead B et al. 2009). We gave the transcripts with the most mapped reads (read depth) at each locus the highest priority for consideration as representative transcripts. Then, we gave the longest transcripts among the translated amino acid sequences containing open reading frames (ORFs) the next priority. If we were unable to select a representative transcript for a given locus based on the first two steps, we selected the candidate transcript with the longest nucleotide sequence as the representative transcript. Thus, we identified 38,413 non-redundant representative transcripts with a mean length of 1,278 bp, an N50 length of 2,233 bp, and a total length of 49,116,021 bp (Table 2).

Evaluation of the assembly quality with genes involved in fatty acid and TAG biosynthesis

To closely look at the quality of the assembled transcripts, we identified the assembled transcripts with matches among 638 Arabidopsis genes involved in fatty acid and TAG biosynthesis pathways in the acyl-lipid metabolism database for Arabidopsis (Bates et al. 2014). First, we checked the quality of the assembled transcripts by looking into the matched coverage of the query genes. Among the three assembly outputs, the improved assembly not only had the highest percentage of \textit{P. citriodora} and Arabidopsis genes with e-value $\leq 1e-10$ and $\geq 50\%$ sequence homology in the BLAST search but also had the highest overall gene coverage by the transcripts. At the level of $>70\%$ matched length, the improved assembled transcript set contained more genes (452, 82.78\%) associated with fatty acid biosynthesis than the Velvet (365, 68.74\%) or Trinity (328, 64.57\%) transcript sets (Table 3). Among the total transcripts, 56.1\% of the Trinity transcripts and 41.9\% of the Velvet transcripts had a size in the range of 201–500 bp (Fig. 1). Those transcripts had lower homology and gene coverage with the Arabidopsis genes compared with the transcripts from the improved assembly. Second, we examined the proportion of full-length genes among the assembled transcripts using the major Arabidopsis genes involved in fatty acid and TAG biosynthesis pathways (Table 4). We considered transcripts full length if they covered more than 85\% of a matching region in an Arabidopsis gene and had a start codon and a stop codon; otherwise, we considered them partial transcripts. As a result, we considered about 82\% of the assembled transcripts to be putative full-length \textit{P. citriodora} genes with homology to a
Table 3 Statistics of sequence homology between *P. citriodora* transcripts and *Arabidopsis* genes involved in fatty acid and TAG biosynthesis

|                | Improved assembly | Velvet assembly | Trinity assembly |
|----------------|-------------------|-----------------|------------------|
|                | Perilla           | Arabidopsis     | Perilla          | Arabidopsis     | Perilla          | Arabidopsis     |
| n              | %                 | n               | %               | n               | %                | n               | %                |
| BLAST ≥C 70%   | 729               | 546             | 750             | 531             | 485              | 508             |
| ≥C 80%         | 416               | 57.06           | 452             | 82.78           | 293              | 82.78           |
| ≥C 90%         | 297               | 40.74           | 342             | 62.64           | 188              | 25.06           |

*BLAST: number of transcripts with e-value ≤ 1e-10 and identity ≥ 50% in sequence homology search by BLAST; C: Arabidopsis gene coverage*

Fig. 2 GO functional classification of the assembled *P. citriodora* transcripts. A bar chart showing the distribution of *P. citriodora* representative transcripts with the percentage of transcripts assigned (x axis) to each GO term (y axis). GO terms are summarized in three main categories: biological process (BP), cellular component (CC), and molecular function (MF).

We directly compared the assembled transcripts to known plant protein sequences using BLASTX (e-value ≤ 1e-10). Of the 38,413 representative transcripts, 23,667 transcripts (61.61%) matched with 16,143 sesame genes, 24,030 transcripts (62.56%) matched with 19,283 proteins in Phytozome, and 17,752 transcripts (73.87% of 24,030) had highest sequence similarity with *Mimulus guttatus* proteins, which is a close species to Perilla at the family level within the same Asterids phylogenetic group. Otherwise, 14,383 (37.44% of 38,413 transcripts) transcripts had no significant matches to any known protein, suggesting some novel genes in *P. citriodora* or a high level of divergence between *P. citriodora* and the other species.

To identify the putative functions of the *P. citriodora* transcripts, we performed a functional enrichment analysis using the GO and KEGG pathway databases. Of the 38,413 representative transcripts, 16,578 (43.16%) transcripts could be assigned at least one GO term with the BLASTX cutoff (e-value ≤ 1e-30) (Fig. 2). A total of 69 sub-classifications were made at the second level of GO depth, including 28...
Table 4 Details of the assembled *P. citriodora* transcripts associated with fatty acid and TAG biosynthesis

| Gene Symbol | AT gene ID | Length\(^a\) (aa) | Perilla transcript ID | Length\(^b\) (aa) | e-value\(^c\) | F or P\(^d\) | transcripts n\(^e\) |
|-------------|------------|------------------|-----------------------|------------------|--------------|-------------|----------------|
| *PDH(E1α)* | AT1G01090 429 | P.citriodora1SL011651t001 438 | 0 F | 1 |
| *PDH(E1β)* | AT2G34590 407 | P.citriodora1SL012138001 308 | 0 F | 4 |
| *EMB3003(E2)* | AT1G44330 466 | P.citriodora1SL006498001 471 | 1e-156 F | 6 |
| *LTA2 (E2)* | AT3G25860 481 | P.citriodora1SL003106001 463 | 6e-148 F | 6 |
| *LPD1 (E3)* | AT3G16950 624 | P.citriodora1SL006166006 611 | 0 F | 2 |
| α-CTa, α-CTb | AT2G38040 770 | P.citriodora1SL009096001 754 | 0 F | 4 |
| β-CT | ATCG00500 489 | P.citriodora1SL024624001 511 | 0 F | 1 |
| *BC* | AT5G35360 556 | P.citriodora1SL003110001 538 | 0 F | 1 |
| *BCCP1* | AT5G16390 281 | P.citriodora1SL018488001 285 | 1e-54 F | 3 |
| *BCCP2* | AT5G15530 256 | P.citriodora1SL004420001 279 | 5e-48 F | 2 |
| *MCMT* | AT2G30200 394 | P.citriodora1SL001628002 408 | 0 F | 1 |
| *KASIIIA, KASIIIB* | AT1G62640 405 | P.citriodora1SL009277001 402 | 0 F | 2 |
| *KAR* | AT1G24360 320 | P.citriodora1SL004327009 320 | 2e-126 F | 6 |
| *HAD* | AT5G10160 220 | P.citriodora1SL017007002 259 | 3e-102 F | 4 |
| *ER* | AT2G05990 391 | P.citriodora1SL015253001 394 | 0 F | 4 |
| *FATA* | AT3G25110 363 | P.citriodora1SL015493001 238 | 1e-176 P | 2 |
| *FATB* | AT1G08510 413 | P.citriodora1SL000470003 422 | 0 F | 2 |
| *FAB2(SAD)* | AT2G43710 402 | P.citriodora1SL024454001 397 | 0 F | 8 |
| *DES6 (SAD)* | AT1G43800 392 | P.citriodora1SL024454001 397 | 0 F | 8 |
| *KASI* | AT5G46290 490 | P.citriodora1SL026266001 475 | 0 F | 10 |
| *KASI(FAB1)* | AT1G74960 542 | P.citriodora1SL001363001 542 | 0 F | 10 |
| *LACS8* | AT2G04350 721 | P.citriodora1SL028662001 697 | 0 F | 12 |
| *LACS9* | AT1G77590 692 | P.citriodora1SL028662001 697 | 0 F | 12 |
| *Endoplasmic reticulum-desaturase* | | | | |
| *FAD2* | AT3G12120 384 | P.citriodora1SL000613001 383 | 0 F | 2 |
| *FAD3* | AT2G29980 387 | P.citriodora1SL002476004 439 | 2e-174 F | 8 |
| *FAD8* | AT5G05580 436 | P.citriodora1SL002476004 439 | 0 F | 16 |
| *Acyl-CoA-dependent TAG synthesis in Kennedy pathway* | | | | |
| *GPAT9* | AT5G60620 377 | P.citriodora1SL004403009 372 | 0 F | 6 |
| *LPAT2* | AT3G57650 390 | P.citriodora1SL004485001 383 | 0 F | 6 |
| *DGA1* | AT2G19450 521 | P.citriodora1SL006978006 450 | 2e-176 P | 2 |
| *DGA2* | AT3G51520 315 | P.citriodora1SL006419008 511 | 7e-74 P | 6 |
| *DGA3* | AT1G48300 285 | P.citriodora1SL003849001 407 | 3e-24 P | 1 |
| *PC-mediated TAG synthesis* | | | | |
| *LPCAT* | AT1G12640 463 | P.citriodora1SL029356001 466 | 0 F | 6 |
| *PDA1* | AT5G13640 672 | P.citriodora1SL004897001 662 | 0 F | 3 |
| *PDA2* | AT3G44830 666 | P.citriodora1SL004897001 662 | 0 F | 3 |
| *DAG-CPT1* | AT1G13560 390 | P.citriodora1SL004437003 390 | 0 F | 12 |
| *DAG-CPT2* | AT3G25585 390 | P.citriodora1SL004437003 390 | 0 F | 12 |
| *PDCT* | AT3G15820 302 | P.citriodora1SL012582001 285 | 2e-109 P | 1 |
| *Oil-body protein* | | | | |
| *OLN-La, OLN-Lb* | AT3G01570 184 | P.citriodora1SL023891001 183 | 2e-32 F | 4 |
| *OLN-Sa, OLN-Sb* | AT5G40420 200 | P.citriodora1SL002844001 176 | 3e-20 P | 4 |
| *OLN-16KD* | AT3G18570 167 | P.citriodora1SL002918001 157 | 1e-32 F | 1 |
| *OLN-18KD* | AT4G25140 174 | P.citriodora1SL002850001 143 | 1e-35 P | 3 |
| *Transcription factor* | | | | |
| *WRI1* | AT3G54320 439 | P.citriodora1SL001076003 611 | 1e-87 P | 32 |

\(^a\)the length of the Arabidopsis gene

\(^b\)the length of the *P. citriodora* transcript

\(^c\)the e-value matched between the two sequences (Arabidopsis and *P. citriodora*)

\(^d\)the full-length or partial-length transcripts in *P. citriodora*

\(^e\)the number of *P. citriodora* transcripts that were homologous to each Arabidopsis sequence
Fig. 3 KEGG classification of the assembled *P. citriodora* transcripts. A bar chart showing the distribution of *P. citriodora* representative transcripts with the number of transcripts assigned (x axis) to each KEGG biological pathway (y axis). M, Metabolism; GIP, Genetic Information Processing; EIP, Environmental Information Processing; CP, Cellular Processes; OS, Organismal Systems.

biological processes, 22 cellular components, and 19 molecular functions. In the category of biological process, cellular metabolic process (18.9%), primary metabolic process (18.8%), and biosynthetic process (12.5%) were the top three subcategories among 11,876 transcripts. In the category of cellular components, intracellular (39.4%) and intracellular organelle (34.7%) were the top subcategories among 13,865 transcripts. In the molecular function category, transferase activity (29.3%) and hydrolase activity (21.5%) were the top subcategories among 8,962 transcripts. The KEGG pathway results were divided into five major classes; Organismal Systems, Cellular Processes, Environmental Information Processing, Genetic Information Processing, and Metabolism; and seventeen sub-classes (Fig. 3). We mapped a total of 2,474 and 3,148 representative transcripts to 100 KEGG pathways in *A. thaliana* (AT) and 107 KEGG pathways in *Solanum lycopersicum* (SL), respectively. Among those, Carbohydrate Metabolism was the most enriched category within Metabolism (377 and 467 transcripts to the AT and SL pathways, respectively), and Translation was the most enriched category within Genetic Information Processing (436 and 519 transcripts to the AT and SL pathways, respectively). One hundred sixty-three and 243 transcripts were classified as AT and SL pathways, respectively, in Lipid Metabolism within Metabolism.

Conclusions

This study highlights the utility of next-generation sequencing (RNA-seq) as a basis for gene set assembly and functional annotation in non-model species. By comparing assembly programs and modifying the assembly pipeline, we assembled 86,396 total transcripts and 38,413 representative transcripts and evaluated the quality of the assembled transcripts based on the Arabidopsis gene coverage and the proportion of full-length genes among the assembled transcripts. Our transcriptome analysis of *P. citriodora* provides a valuable genetic resource to elucidate the molecular basis of various metabolic pathways and to enhance the molecular breeding of Perilla species.

Acknowledgements

This work was carried out with the support of the National Agricultural Genome Program (PJ010408) of the RDA
(Rural Development Administration), Republic of Korea.

References

Bates PD, Johnson SR, Cao X, Li J, Nam JW, Jaworski JG, et al. (2014) Fatty acid synthesis is inhibited by inefficient utilization of unusual fatty acids for glycerolipid assembly. Proc Natl Acad Sci USA 111(3), 1204-1209

Bumblauskiene L, Jakstas V, Janulis V, Mazzieriene R, Ragazinskiene O. (2009) Preliminary analysis on essential oil composition of Perilla L. cultivated in Lithuania. Acta Pol Pharm 66(4), 409-413

Chen G, Yin K, Wang C, Shi T. (2011a) De novo transcriptome assembly of RNA-Seq reads with different strategies. Sci China Life Sci 54(12), 1129-1133

Chen G, Li R, Shi L, Qi J, Hu P, Luo J, et al. (2011b) Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics 12, 590

Cox MP, Peterson DA, Biggs PJ. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485

Fukushima A, Nakamura M, Suzuki H, Saito K, Yamazaki M. (2015) High-Throughput Sequencing and De Novo Assembly of Red and Green Forms of the Perilla frutescens var. crispa Transcriptome. PLoS One 10(6), e0129154

Garber M, Grabherr MG, Guttmann M, Trapnell C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8(6), 469-477

Gongora-Castillo E, Buell CR. (2013) Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat Prod Rep 30(4), 490-500

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(7), 644-652

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue), D258-261

Iorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, Matveenko M, et al. (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics 12, 389

Ito M, Kiuchi F, Yang LL, Honda G. (2000) Perilla citriodora from Taiwan and its phytochemical characteristics. Biol Pharm Bull 23(3), 359-362

Illumina (http://www.illumina.com/products/truseq_ma_library_prep _kit_v2.html)

Jung CS, Lee MY, Oh KW, HK Kim, Park CB, Sung JD, Suh DY. (2005) Discovery of New Diploid Perilla Species in Korea. Korean J. Breed 37(3), 152-154

Kanehisa M, Goto S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1), 27-30

Kim HA, Lim CJ, Kim S, Choe JK, Jo SH, Baek N, et al. (2014) High-throughput sequencing and de novo assembly of Brassica oleracea var. Capitata L. for transcriptome analysis. PLoS One 9(3), e92087

Kim HU, Chen GQ. (2015) Identification of hydroxy fatty acid and triacylglycerol metabolism-related genes in lesquerella through seed transcriptome analysis. BMC Genomics 16, 230

KAPA biosystems (http://www.kapabiosystems.com/product-applications/products/next-generation-sequencing-2/library-quantification/)

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 15(12), 553

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3), R25

Lee SC, Lee JK, Kim NH, Park JY, Kim HU, Lee HO, et al. (2014) Analysis of expressed sequence tags from a normalized cDNA library of perilla (Perilla frutescens). Journal of Plant Biology 57(5), 312-320

Marguerat S, Bahler J. (2010) RNA-seq: from technology to biology. Cell Mol Life Sci 67(4), 569-579

Martin JA, Wang Z. (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10), 671-682

Ness RW, Siol M, Barrett SC. (2011) De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. BMC Genomics 12, 298

Schulz MH, Zerbino DR, Vingron M, Birney E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8), 1086-1092

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

Zerbino DR, Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5), 821-829

Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, 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 Genomics 13, 90