Biochemical Systematics and Ecology 67 (2016) 7–16

Contents lists available at ScienceDirect

Biochemical Systematics and Ecology

journal homepage: www.elsevier.com/locate/biochemsyseco

Insights into biosynthetic genes involved in the secondary metabolism of *Gardenia jasminoides* Ellis using transcriptome sequencing

Chun-xia Yang [a], Ting Zhang [a], Meng Xu [b], Pei-lin Zhu [a, *], Shao-yong Deng [a]

[a] Jiangxi Academy of Forestry, Nanchang, Jiangxi 330032, China
[b] Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing 210037, China

ARTICLE INFO

Article history:
Received 14 December 2015
Received in revised form 16 May 2016
Accepted 21 May 2016
Available online 30 May 2016

Keywords:
Gardenia jasminoides
Transcriptome
Gene expression differences
Carotenoid biosynthesis

ABSTRACT

Gardenia jasminoides is an economically important woody plant in Asia. It contains multiple secondary metabolites in its fruit that are used in Chinese traditional medicines for treating diseases and as a natural colorant for food. However, little is known at the genetic level about how secondary metabolites are biosynthesized in this plant. Thus, transcriptome sequencing of *G. jasminoides* fruit at two developmental stages was performed using the Illumina sequencing platform. A total of 55,376,454 clean reads were obtained. Assembly of the clean reads resulted in 100,889 transcripts, including 51,908 unigenes. Of these, 23,361 (45.0%) were unigenes with significant similarities to publicly available plant protein sequences, and 6225 unigenes were mapped to 246 KEGG pathways. There were 1757 differentially expressed unigenes between the two developmental fruit stages, including 902 up-regulated and 855 down-regulated unigenes. Metabolic pathway analysis revealed that 29 unigenes were predicted for the biosynthesis of carotenoids, with 10 genes differentially expressed between the two fruit stages. The globally sequenced genes covered a considerable proportion of the *G. jasminoides* transcriptome, and these data provide a valuable resource for genetic and genomic studies on secondary metabolite synthesis in *G. jasminoides*.

1. Introduction

Gardenia jasminoides Ellis is an evergreen shrub with white flowers containing a simple corolla, and belonging to the Rubiaceae. The plant originated from central China and has been cultivated in China for more than 2000 years (since the Han Dynasty). It is now commonly found in some Asian countries and regions (Han et al., 2007). Zhizi (a Chinese herbal name for the dried ripe fruit of *G. jasminoides*) is a traditional Chinese medicine with extensive pharmacological activities (Koo et al., 2004), and is used for treatment of jaundice, inflammation, headache, edema, fever, hepatic disorders and hypertension (Commission, 2010). In addition, gardenia yellow, an extract of gardenia fruit, is a natural pigment widely utilized as a coloring agent in the food industry (Lee et al., 2005; Yamada et al., 2012).

There are many secondary metabolites in gardenia fruit with valuable nutritional and pharmacological effects on humans. Carotenoids, which are important secondary metabolites in plants, play significant roles in protecting the

* Corresponding author.
E-mail address: yczpl@126.com (P.-l. Zhu).

http://dx.doi.org/10.1016/j.bse.2016.05.011
0305-1978/

photosynthetic apparatus from photo-oxidation and have fundamental roles in human nutrition as metabolic precursors and antioxidants (Frank and Cogdell, 1996; Fraser and Bramley, 2004). Furthermore, carotenoids are the most widespread group of pigments in plants' and are associated with the colors of flowers, fruits, leaves, shoots, roots and seeds. Gardenia fruits are deep red and reddish yellow at the fully ripened stage due to a high accumulation of carotenoids. However, almost all of the biosynthetic pathways of these secondary metabolites remain poorly understood, with limited knowledge at the molecular level. The development of next generation sequencing technology (NGS) has enabled progress in this area.

NGS is a highly efficient, rapid and low cost DNA sequencing platform, and a number of secondary metabolism genes have been successfully identified, such as from *Asparagus racemosus* (Upadhyay et al., 2014) and *Uncaria rhynchophylla* (Guo et al., 2014) with little genomic sequence information. SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina and GS FLX Titanium/GS Junior from Roche are the three best-known NGS platforms (Liu et al., 2012). Illumina HiSeq 2000, with the greatest output and lowest reagent cost, has recently been widely used for the deep sequencing of model and non-model organisms (Ge et al., 2014; Zhai et al., 2013). Here, *de novo* transcriptome sequencing of gardenia fruit at two developmental stages was performed using the Illumina HiSeq 2000 sequencing platform. Some functional genes and potential metabolic pathways were determined by bioinformatics analyses, and the transcripts involved in carotenoid biosynthesis were identified. The transcriptome data will provide an insight into the molecular pathways and regulatory mechanism active during *G. jasminoides* fruit development in future studies.

2. Materials and methods

2.1. Plant material

The *G. jasminoides* fruit (Voucher number: 2013010020, JXF) at different ripening stages were harvested from a plantation in Zhangshu City, Jiangxi Province, China. Transcriptome sequencing was applied to fruit at weeks 16 and 36 following flowering. The 16-week fruit were closed with green exocarps and colorless mesocarps, while 36-week fruit were closed with red exocarps and deep ripening mesocarps. The fruit were immediately frozen in liquid nitrogen upon harvesting and stored at −80 °C until further processing.

2.2. RNA isolation, library preparation and Illumina sequencing

Total RNA samples were isolated using the CTAB-LiCl method (Zhang et al., 2013). RNA concentrations were measured using a Qubit RNA Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies, CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA). A total amount of 3 μg of RNA per sample was used as input material for the RNA sample preparations.

Sequencing libraries were generated using a NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's recommendations. The library quality was assessed on the Agilent Bioanalyzer 2100 system. Clustering of the index-coded samples was performed on a cBot Cluster Generation System using a TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq 2000 platform and paired-end reads were generated.

2.3. De novo assembly

Prior to assembly, clean reads were obtained by removing reads containing adapters, reads containing poly-N in excess of 10% and those with more than 50% low-quality base identification (base quality ≤ 5) from raw data. At the same time, Q20, Q30, GC-content and sequence duplication levels of the clean data were calculated. All downstream analyses were based on clean data of high quality. Then, the *de novo* assembly of clean reads was accomplished using Trinity software (v2012-10-05), which was designed specifically for transcriptome assembly from RNA-Seq data (Grabherr et al., 2011). The sequences resulting from the Trinity assembly were termed ‘unigenes’.

2.4. Gene functional annotation

Some databases, including NCBI non-redundant protein sequences (Nr), NCBI non-redundant nucleotide sequences (Nt), Protein family (Pfam), Clusters of Orthologous Groups of proteins (KOG/COG), a manually annotated and reviewed protein sequence database (Swiss-Prot), KEGG Ortholog database (KO) and Gene Ontology (GO), were used to analyze the assembled unigenes. The best alignment results determined the unigenes' directions. When there was a conflict among the database results, the priority order was Nr, Swiss-Prot, KO and then COG.

2.5. Gene expression analyses

Gene expression levels were estimated using RSEM software for each sample (Li and Dewey, 2011). The expected number of fragments per kilobase of transcript sequence per million base-pairs sequenced is currently the most commonly used