Gene 534 (2014) 155–162

Contents lists available at ScienceDirect

Gene

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

De novo transcriptome analysis of *Liriodendron chinense* petals and leaves by Illumina sequencing

Ying Yang [1], Meng Xu [1], Qunfeng Luo, Jie Wang, Huogen Li *

Key Laboratory of Forest Genetics & Gene Engineering of the Ministry of Education, Nanjing Forestry University, Nanjing 210037, China

ARTICLE INFO

Article history:
Accepted 27 October 2013
Available online 14 November 2013

Keywords:
Endangered
Unigenes
Gene expression differences
Carotenoids biosynthesis

ABSTRACT

Liriodendron chinense (Hemsl.) Sarg is an endangered species and occupies a pivotal position in phylogenetic studies of flowering plants, while its genomic resources are limited. In this study, we performed transcriptome sequencing for *L. chinense* petals and leaves using the Illumina paired-end sequencing technique. Approximately 17.02-Gb clean reads were obtained, and *de novo* assembly generated 87,841 unigenes, with an average length of 778 bp. Of these, there were 65,535 (74.61%) unigenes with significant similarity to publically available plant protein sequences. There were 3386 genes identified as significant differentially expressed between petals and leaves, among them 2969 (87.68%) were up-regulated and 417 (12.31%) down-regulated in petals. Metabolic pathway analysis revealed that 25 unigenes were predicted to be responsible for the biosynthesis of carotenoids, with 7 genes differentially expressed between these two tissues. This report is the first to identify genes associated with carotenoid biosynthesis in *Liriodendron* and represents a valuable resource for future genomic studies on the endangered species *L. chinense*.

1. Introduction

Liriodendron chinense (Hemsl.) Sarg is an endangered deciduous tree species belonging to the magnolia family (Magnoliaceae). It is native to southern China and northern Vietnam, occurring in small and widely isolated populations in China, and with dramatically declining numbers (He and Hao, 1999). This species is listed as a protected Chinese plant (Fu and Jin, 1992) and an IUCN endangered plant (http://www.iucnredlist.org/search). The genus *Liriodendron* contains only two species, with the others species, *Liriodendron tulipifera*, widely distributed in eastern North America. The two species are morphologically distinct, with the crucial distinction in petal color, with yellow-green in *L.chinense* and orange-yellow in *L. tulipifera* (Wen and Wu, 1993). Additionally, *L. chinense* leaves have one side lobe on each side of the leaf and dense salient white points on the abaxial leaf surface; correspondingly, *L.tulipifera* has 2–3 side lobes and sparse white points (Fang, 1994; Hao et al., 1995). These two species form a biogeographic pattern of disjunctive distributions in eastern Asia and eastern North America (Wu and Wang, 1983). Understanding the origin and development of *Liriodendron* will help elucidate the relationship between the respective east Asian and North American flora (Nathan et al., 2008; Nie et al., 2006), and the climate and geological changes in Northern Hemisphere history (Gamble et al., 2008; Lemmon et al., 2007). Thus, the genus of *Liriodendron* occupies an ideal position in phylogenetic studies of flowering plants.

The genomic resources of *Liriodendron* are still limited, although several studies have been conducted in *L.tulipifera* (Jin et al., 2011; Liang et al., 2008, 2011). Currently, there are few DNA sequences available in the NCBI database for *L. chinense*. Considering the urgency for germplasm conservation of *L. chinense* and its important role in phylogenetic studies, the advent of transcriptome data of *L. chinense* will contribute to the conservation of this endangered species and related genetic research in *Liriodendron*.

Carotenoids are isoprenoid pigments that are secondary metabolites ubiquitous in nature and essential for life. The importance of carotenoids in human health has been extensively reviewed. For example, they are associated with the human immune system (Chew and Park, 2004) and cancer (Giovannucci, 2002) and many other diseases (Fraser and Bramley, 2004; Giuliano et al., 2008). Carotenoids in plants are associated with the formation of the phytohormones abscisic acid (ABA) and the production of color and flavor that attract insects and animals for pollination and seed dispersal. Orange and yellow are carotenoid-based colors in plants (Fambrini et al., 2010) and as mentioned above, *L. chinense* bears yellow-green flowers, while *L. tulipifera* has orange-yellow flowers. Therefore, analysis of interspecies differences in

Abbreviations: ABA, phytohormones abscisic acid; MYB340, Myb-related protein 340; FBP2, floral binding protein 2; PKS1, phytochrome kinase substrate 1; PSY, phytoene synthase; GGPP, geranylgeranyl pyrophosphate synthase; PDS, phytoene desaturase; Z-ISO, ζ-carotene isomerase; ZDS, ζ-carotene desaturase; CRTISO, carotenoid isomerase; LCY, lycopene cyclases; ZEP, zeaxanthin epoxidase; VDE, violaxanthin de-epoxidase; CCS, capsanthin/capsorubin synthase; NGS, next-generation sequencing; NR, non-redundant protein database; GO, Gene Ontology; Nt, non-redundant nucleotide sequences; COG, Clusters of Orthologous Groups of proteins; KO, KEGG Ortholog database; DEGs, differentially expressed genes.

* Corresponding author at: Key Laboratory of Forest Genetics & Gene Engineering of the Ministry of Education, Nanjing Forestry University, Longpan road 159#, Nanjing 210037, Jiangsu, China. Fax: +86 25 85427760.
E-mail address: hgli@njfu.edu.cn (H. Li).

[1] These authors contributed equally to this work.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.10.073

carotenoid biosynthesis might help disclose the divergence between *L. chinense* and *L. tulipifera*. Thus, unraveling the pathway for carotenoid biosynthetic and the underlying mechanism of gene regulation of *L. chinense* could be of great significance.

In this paper, we characterized the petal and leaf transcriptomes of *L. chinense* using the *de novo* Illumina sequencing platform. Furthermore, we aimed to identify the transcripts involved in carotenoid biosynthesis. To our knowledge, this study is the first transcriptome resource of the endangered species *L. chinense*. This resource and the findings will contribute to studies on functional genomics and biogeography of *Liriodendron*.

2. Results

2.1. Sequencing and assembly

With the purpose of determining the leaf and petal transcriptomes of *L. chinense*, two sequencing libraries were prepared from the two tissues and sequenced with the Illumina paired-end technique. In total, there were 102,526,130 raw reads generated from leaves and 100,572,686 from petals (Table 1). The sequencing raw data have been submitted to the Short Reads Archive (SRA) under the accession numbers SRP029714. Of the raw reads from petals, more than 95.12% bases has a Q value ≥ 20 (an error probability of 0.045%), and for leaves 94.41% (an error probability of 0.06%). The GC-contents were 46.17and 50.61% for petals and leaves respectively. These were used for *de novo* assembly.

There were 81,596,006 clean reads for petals and 90, 630,278 for leaves (All sequences are being submitted to the NCBI Sequence Read Database). The Trinity software generated 129,097 all-transcripts (Table 2) with an average length of 778 bp and an N50 of 1434 bp; and 87,841 all-unigenes were achieved (Supplementary File 1). Of these, 13,170 (14.99%) were 200–500 bp, 13,170(14.99%) were 500–1000 bp, 8060 (9.18%) were 1–2 kb and the remaining 4462 (5.08%) were >2 kb (Table 2).

2.2. Functional annotation and classification

All the 87,841 assembled unigenes were searched against the Nr Nt, Swiss-Prot and COG databases using the BLAST algorithm (E-value < 1E-5). A total of 65,535 unigenes were annotated, accounting for 74.61% (Table 3). Among them 23,188 unigenes (26.40%) showed high homology with sequences in the Nr database, and 16,295 (18.55%) showed homology with unknown genes (hypothetical proteins). The number of unigenes with significant similarity to sequences in COG, KEGG and Nt databases were 8864 (10.09%), 5229 (5.95%) and 8345 (9.50%), respectively (Table 3).

With the GO classification, the 21,560 matched unigenes were classified into 3 functional categories: molecular function, biological process and cellular component (Fig. 1). In molecular function, these matched unique sequences were clustered into 17classifications. The largest subcategory of the molecular function was 'binding' (42.88%) and the second was 'catalytic activity' (36.24%). In terms of biological processes, these unique sequences were classed into 27 classifications: the most represented biological processes were 'cellular process' (25.55%) and 'metabolic process' (25.47%). According to cellular components, these unique sequences were divided into 14classifications: the most represented cellular components were 'cell' (30.31%) and 'cell part' (30.31%).

The 8864 matched unique sequences were clustered into 14 categories (Fig. 2), using COG proteins. The proteins in the COG categories were assumed to have the same ancestor protein, or to be paralogs or orthologs. The largest category was general functional prediction only, with 17.38%; the second categories were post-translational modification, protein turnover and chaperon (together 10.58%); the third category was signal transduction, with 8.12%.

The assembled unigenes were assigned to the biochemical pathways described in KEGG (Fig. 3). The KEGG database contains a systematic analysis of inner-cell metabolic pathways and functions of gene products. Pathway-based analyses help to further determine the biological function of genes. A total of 26,229 unigenes had blast hits, and 5229 were assigned to 5 KEGG biochemical pathways: metabolism (3330 unigenes), genetic information processing (1715), organism system (905), cellular processes (814) and environmental information processing (507). The largest group, the metabolic pathways, was well represented among the 3330 unigenes of *L. chinense*, with most associated with carbohydrate metabolism (762), amino acid metabolism (556), energy metabolism (506), lipid metabolism (366), nucleotide metabolism (297) and metabolism of cofactors and vitamins (180). Those pathways related to genetic information processing were the second largest group, including genes involved in translation (697), folding (479), replication and repair (323) and transcription (216). The third largest group comprised organismal systems, with a majority of the proteins involved in the nervous (190) and endocrine (176) systems. Pathways related to cellular processes and environmental information processing were also well represented by unigenes from *L. chinense*. These results provide a valuable resource for investigating metabolic pathways in petals and leaves of *L. chinense*.

Table 1
Quality of sequencing.

Sample	Raw reads	Clean reads	Clean bases	Error (%)	Q20 (%)	Q 30 (%)	GC (%)
Petal-R1	50286343	40798003	4.0G	0.04	96.61	90.34	47.11
Petal-R2	50286343	40798003	4.0G	0.05	93.60	85.85	45.22
Leaf-R1	51263065	45315139	4.51G	0.05	96.21	89.12	49.46
Leaf-R2	51263065	45315139	4.51G	0.07	92.61	83.63	51.76
Petal-summary	100572686	81596006	8.0G	0.045	95.12	88.10	46.17
Leaf-summary	102526130	90630278	9.02G	0.06	94.41	86.38	50.61
Summary	203098816	172226284	17.02G				48.39

R1: Reads sequencing from the left.
R2: Reads sequencing from the right.
Q20: The percentage of bases with a Phred value >20.
Q30: The percentage of bases with a Phred value >30.

Table 2
Length distribution of assembled transcripts and unigenes.

Nucleotides length (bp)	Transcripts	Unigenes
200–500 bp	76,321	62,149
500–1 kbp	22,703	13,170
1 k–2 kbp	18,518	8060
>2 kbp	11,537	4462
Total	129,097	87,841
Minimal length (bp)	201	201
Maximal length (bp)	24,988	16,943
N50 (bp)	1434	764
Average length (bp)	778	537

Table 3
Summary for the annotation of unigenes of *s*.

	Sequences (n)	Frequency	Functional categories (n)
All assembled unigenes	87,841		
Gene annotations against NR	23,188	26.40%	
Gene annotations against Swiss-Prot	16,295	18.55%	
Gene annotations against NT	8345	9.5%	
GO annotation for NR protein hits	21,560	24.54%	58
Gene annotations against COG	8864	10.09%	26
Gene annotations against KEGG	5229	5.95%	38
All annotated unigenes	65,535	74.61%	