Gene 544 (2014) 220–227

Contents lists available at ScienceDirect

Gene

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

De novo transcriptome analysis of the Siberian apricot (*Prunus sibirica* L.) and search for potential SSR markers by 454 pyrosequencing

Shubin Dong [a], Yulin Liu [b], Jun Niu [a], Yu Ning [a], Shanzhi Lin [b], Zhixiang Zhang [a,*]

[a] Lab of Systematic Evolution and Biogeography of Woody Plants, College of Nature Conservation, Beijing Forestry University, Beijing 100083, China
[b] Key Laboratory for Genetics and Breeding of Forest Trees and Ornamental Plants of Ministry of Education, College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 7 January 2014
Received in revised form 10 April 2014
Accepted 17 April 2014
Available online 18 April 2014

Keywords:
454 pyrosequencing
Siberian apricot (*Prunus sibirica* L.)
SSR markers
Transcriptome

ABSTRACT

The Siberian apricot, an economically and ecologically important plant in China, contains seeds high in oil and can grow on marginal land. Although this species has multiple purposes and may be a feedstock of biofuel in China, transcriptome information and molecular research on this species remain limited. RNA-Seq technology has been widely applied to transcriptomics, genomics and the development of molecular markers, and functional gene studies. In this study, we obtained 1,243,067 high-quality reads with a mean size of 425 bp in a single run, totaling 528.4 Mb of sequence data using 454 GS FLX Titanium sequencing. All reads were assembled de novo into 46,940 unigenes with a mean size of 651 bp (range: 45–5566 bp). Assembled unigenes were annotated in multiple public databases based on similarity alignments to genes and proteins. 191 unigenes involving in lipid biosynthesis and metabolism were found, among them, expression patterns of two desaturase enzymes were analyzed by quantitative real-time polymerase chain reaction (qRT-PCR), based on six tissues from Siberian apricot, the seeds had the highest expression. 7304 simple sequence repeats (SSR) were identified from 6509 unigenes, a total of 9930 primer pairs were designed, 50 primer pairs were randomly selected to validate of the usefulness, and 24 (48%) primer pairs produced bands of the expected size. These data provide a base of sequence information to improve agronomic characters and molecular marker-assisted breeding to alter the composition of fatty acids in seeds from this plant, and hence, facilitate its utilization as a future biodiesel feedstock.

1. Introduction

With rapid economic and industrial modernization, a worldwide energy crisis and environmental pollution are increasingly causing concern, with many countries actively seeking substitutes for fossil fuels. Biodiesel is a clean, biodegradable, renewable energy source, but unlike Europe and the United States, China does not have the land available for growing large-scale biofuel crops (e.g., soybean, rape) because maintaining sufficient land for food production is a priority, currently the main biofuel sources are waste kitchen oil and a very limited supply of non-edible plant oil (e.g., *Jatropha curcas*, *Xanthoceras sorbifolia*, *Pistacia chinensis*). Due to rising prices in waste cooking oil and shortage of non-food feedstocks, many biofuel producers in China have stopped operating in recent years. Therefore, to find alternatives to petrochemicals and to develop a diversity of biofuel feedstocks, fully understanding the characteristics and growth habits of any potential biofuel plants in China is vital.

The Siberian apricot (*Prunus sibirica* L.) is widely grown in northern and northeastern China, eastern and southeastern Mongolia, eastern Siberia, and the maritime area of Russia (Wang, 2011). This tree has many excellent characteristics, including resistance to salt, drought, low temperature stress, and alleviation of soil degradation and prevention of wind and water soil erosion (Wang, 2008). The Grain for Green Program and Three North Shelterbelt Project have been planting this species throughout China; the total area of both naturally distributed and planted trees is estimated to be 2,266,700 ha, and the annual seed production is over 192,500 tons (Wang, 2011). Research results from 17 samples taken from different geographic regions indicate that the Siberian apricot seed kernel has a high oil content, ranging from 44.73% to 57.83%, with a mean of 50.18% (Wang, 2012a). The fatty acid composition of the seed kernel oil includes a high percentage of oleic (65.23 ± 4.97%) and linoleic acid (28.92 ± 4.62%) (Wang, 2012b), making it a good potential biofuel crop and a potential edible oil in China.

Despite its promise as a biofuel, the Siberian apricot has mostly been researched for its fatty acid composition (Wang, 2012b), seed oil content, biodiesel traits (Gumus and Kasifoglu, 2010), ecological use, afforestation (Yuan et al., 2010), medicinal uses (Grundy, 1987), other extracts (Harrison and Were, 2007; Wijeratne et al., 2006), production

Abbreviations: ACP, acyl carrier protein; BLAST, basic local alignment search tool; CDD, Conserved Domain Database; COG, cluster of orthologous group; EST, expressed sequence tag; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; NCBI, National Center for Biotechnology Information; SSRs, simple sequence repeats.

* Corresponding author.
E-mail address: zxzhang@bjfu.edu.cn (Z. Zhang).

http://dx.doi.org/10.1016/j.gene.2014.04.031
0378-1119/

of an almond drink (He et al., 2011; Li et al., 2008), and grafting. To date, only 19 primer pairs have been reported for use in biomolecular studies on the Siberian apricot (Liu et al., 2013), and no published data are available on the transcriptome of the Siberian apricot. As it has a longer breeding cycle, a desperate need exists for genetic engineering methods and information on genetic manipulation, which require genomic information to clone important genes. The current lack of Siberian apricot genome sequence data limits the development of gene-based biofuels.

In recent years, transcriptome sequencing can provide additional data on gene expression, biological pathways, and molecular mechanisms, even without reference genome information (Margulies et al., 2005; Metzker, 2009; Sun et al., 2012; Wang et al., 2009; Wheat, 2010). The GS-FLX Titanium, the 454 sequencing platform, can generate more than 1 million reads, with an average length of up to 400 base pairs (bp) at 99.5% accuracy per run (Sun et al., 2010). Continuous development of the GS FLX Titanium chemistry now can offer read lengths up to 1 kb (http://454.com/products/gs-flx-system/index.asp). Because of the increased read length generated from 454 pyrosequencing compared to other platforms (Collins et al., 2008; Parchman et al., 2010), it has been successfully applied in transcriptome sequencing of many non-model species, including olive (Alagna et al., 2009), and chestnut (Barakat et al., 2009), American ginseng (Sun et al., 2010), *J. curcas* (Natarajan and Parani, 2011), lentil (Kaur et al., 2011), *Ammopiptanthus mongolicus* (Zhou et al., 2012), *Ulva linza* (Zhang et al., 2012), *Chlamydomonas* spp. (Kim et al., 2013), *Lycoris aurea* (Wang et al., 2013), tea (H. Wu et al., 2013), *Podophyllum hexandrum* (Bhattacharyya et al., 2013).

In this paper, we first report on and discuss the results of transcriptome sequencing, a search for simple sequence repeats (SSRs), and putative unigenes involved in fatty acid biosynthesis in the Siberian apricot. The transcriptome data generated from our study comprise a useful resource for gene excavation, molecular marker development, transcriptomic assembly, and microarray development. Additionally, the SSR markers identified in this study will contribute to marker-assisted breeding selection, facilitate gene mapping, help with linkage mapping, and enable genetic diversity analysis of the Siberian apricot.

2. Materials and methods

2.1. Material selection and RNA extraction

Buds, leaves, stems, flowers, fruit pulp, and seeds of the Siberian apricot were collected from a single individual tree at the Beijing Forestry University experimental station. All materials were immediately frozen in liquid nitrogen after collection and stored at −80 °C until use. Total RNA was extracted from all tissues using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. RNA quality and quantity were assessed by electrophoresis on a 1% agarose gel and via a spectrophotometer, respectively. Equal amounts of total RNA from each tissue were blended together for cDNA preparation.

2.2. Library preparation and 454 sequencing

Magnetic beads with an oligo (dT) were used to isolate the poly (A^+) RNA from the total RNA mixture. Following this, the purified mRNA was fractionated using fragmentation solution at 70 °C for 30 s. Random primers were used to synthesize the first strand of cDNA, and the second strand cDNA was synthesized using DNA polymerase I and RNaseH (Zhu et al., 2001). Double-stranded cDNA was separated on a 2% agarose gel, and cDNA shorter than 100 bp was cut from the gel. Purification and end repair of these cDNA fragments were carried out with AMPure beads (Beckman Coulter, Brea, CA, USA) and End Repair mix (Invitrogen), respectively, and a specific sequence adapter was then ligated to this repaired cDNA. The cDNA library was then sequenced using a 454 GS FLX Titanium genomic sequencer (Roche, Indianapolis, IN, USA).

2.3. Assembly

Original sequences were pre-processed before 454 pyrosequencing to eliminate sequences shorter than 45 bp and low-quality sequences, using the Lucy (http://lucy.sourceforge.et/) and Newbler (http://454.com/products/analysis-software/index.asp) software programs. SeqClean (http://compbio.dfci.harvard.edu/tgi/software) was used to trim the adapters and SMART primers used for reverse transcription. After performance comparison of five de novo assemblers (CAP3, MIRA, Newbler, SeqMan, and CLC), Newbler 2.5 produced the best contig lengths, better alignment to reference sequences, and was faster and easy to use (Kumar and Blaxter, 2010). So we used Roche's 454 Newbler (Version 2.6) to assemble short reads into long isotigs. However, some reads could not be assembled and remained as single sequences that formed a collection of singlets. All isotigs and singlets sequences with more than 95% similarity were clustered into groups using CD-HIT (Version 4.5.6), the longest sequences in each group represented a group on their own, and these representative sequences formed the unigenes data sets of the Siberian apricot transcriptome.

2.4. Functional annotation and classification

We used BLASTX (E-value < 10^{-5}) to compare the unigenes with several protein databases, such as the NCBI database (http://www.ncbi.nlm.nih.gov), PFAM (http://pfam.sanger.ac.uk), and Swiss-Prot (http://www.expasy.ch/sprot). The BLASTX result was parsed with genes previously identified in the databases, and based on the accession number for each BLASTX hit in NCBI (Meyer et al., 2009), we obtained the gene name and taxonomic information. Uncertain BLASTX results, such as unknown, unnamed, hypothetical, and uncharacterized unigenes, were omitted from our analysis (Sloan et al., 2012). All unigenes were compared with the Conserved Domain Database (CDD) (http://www.ncbi.nlm.nih.gov/cdd) to identify conserved protein domains. These searches were carried out with a Reverse Position Specific Blast (E-value < 10^{-5}), which is a blast-like algorithm for comparing newly identified sequences to a set of known profiles (Marchler et al., 2002). The unigenes were also searched in the TrEMBL database (http://www.uniprot.org), and the resulting output was analyzed with Blast2GO (Conesa et al., 2005) using the default parameter settings to assign gene ontology (GO) terms (molecular function, cellular component, biological process) to each sequence. The unigene sequences were also aligned against the Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/cog) to predict and classify possible functions. Pathway assignments were performed according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa and Goto, 2000).

2.5. Related species comparison analysis

The protein datasets of *Prunus persica*, *Malus domestica* and *Fragaria vesca* which species is Rosaceae family with *P. sibirica* were downloaded from the NCBI EST database, Non-redundant datasets were then generated using CD-HIT-EST as previously described (Li and Godzik, 2006). Sequence similarity comparisons and clustering were performed using tBLASTx in conjunction with OrthoMCL (Li et al., 2003) using a defined E-value < 10^{-10}.

2.6. Relative expression analysis of stearyl-ACP desaturase (SAD) and Δ12-fatty-acid desaturase (DFAD) genes in various tissues

Total RNA from the buds, leaves, stems, flowers, fruit pulp, and seeds was treated with DNase I (Takara, Dalian, China) and purified. Quantified total RNA from each sample was used for first-strand cDNA