Biotechnology Reports 17 (2018) 63–69

Contents lists available at ScienceDirect

Biotechnology Reports

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

Transcriptome sequencing and de novo assembly in arecanut, *Areca catechu* L elucidates the secondary metabolite pathway genes

Ramaswamy Manimekalai[a,*], Smita Nair[b], A. Naganeeswaran[b], Anitha Karun[b], Suresh Malhotra[c], V. Hubbali[d]

[a] *Sugarcane Breeding Institute, Indian Council of Agricultural Research (ICAR), Coimbatore, 641 007, Tamil Nadu, India*
[b] *Central Plantation Crops Research institute, Indian Council of Agricultural Research (ICAR), Kudlu P.O., Kasaragod 671 124, Kerala, India*
[c] *Indian Council of Agricultural Research (ICAR), KAB II, New Delhi, India*
[d] *Directorate of Arecanut and Cocoa Development, Kera Bhavan, Kochi, India*

ARTICLE INFO

Keywords:
De novo assembly
Areca genome
Flavonoids
Carotenoids

ABSTRACT

Areca catechu L. belongs to the Arecaceae family which comprises many economically important palms. The palm is a source of alkaloids and carotenoids. The lack of ample genetic information in public databases has been a constraint for the genetic improvement of arecanut. To gain molecular insight into the palm, high throughput RNA sequencing and de novo assembly of arecanut leaf transcriptome was undertaken in the present study. A total 56,321,907 paired end reads of 101 bp length consisting of 11.343 Gb nucleotides were generated. De novo assembly resulted in 48,783 good quality transcripts, of which 67% of transcripts could be annotated against NCBI non – redundant database. The Gene Ontology (GO) analysis with UniProt database identified 9222 biological process, 11268 molecular function and 7574 cellular components GO terms. Large scale expression profiling through Fragments per Kilobase per Million mapped reads (FPKM) showed major genes involved in different metabolic pathways of the plant. Metabolic pathway analysis of the assembled transcripts identified 124 plant related pathways. The transcripts related to carotenoid and alkaloid biosynthetic pathways had more number of reads and FPKM values suggesting higher expression of these genes. The arecanut transcript sequences generated in the study showed high similarity with coconut, oil palm and date palm sequences retrieved from public domains. We also identified 6853 genic SSR regions in the arecanut. The possible primers were designed for SSR detection and this would simplify the future efforts in genetic characterization of arecanut.

1. Introduction

The arecanut palm (*Areca catechu* L., Arecaceae family) is an economically important palm species in the Old World tropics providing livelihood options to millions of farmers. Other economically important members of Arecaceae family are coconut, date palm, oil palm, etc. Arecanut is believed to have originated in Malaysia or the Philippines, is grown extensively in much of the tropical Pacific, Asia and East Africa largely for its fruit which is widely used for masticatory and religious purposes. The leaf sheaths are used as plates, bags, and as wrapping and packing material [1]. The medicinal properties of arecanut have been identified long back with regard to its use against leucoderma, leprosy, cough, fits, worms, anemia and obesity. It is also used as a purgative and is also a component in the ointment for treatment of nasal ulcers [2]. Betel nut is a source of alkaloids and flavonoids. The areca alkaloids comprise arecoline, arecaidine, guvacoline, and guvacine while the flavonoid components comprise tannins and catechins [3]. The ripened pericarp tissue of fruit accumulates carotene compounds. The β-carotene constitutes nearly 30% of the total carotenoid content in the pericarp tissues [4]. The total carotenoid content was found to be 11.67 ± 0.62 mg carotene equivalents per 100 g fresh mass of pericarp tissue.

Intense research activities have been carried out to understand the genetic variability and genetic diversity of arecanut palm in the past [5–7]. Despite the economic importance of arecanut, not much work has been done to understand its genomics. At present, sparse amount of sequence information only available for arecanut palm in the public domain databases. However, whole genome sequence information is present for other economically important members of Arecaceae family like date palm and oil palm [8,9]. Recent developments in genomics and bioinformatics have enabled better understanding of plant genomes. Nowadays, the RNA Seq approach based on next generation

* Corresponding author.
E-mail addresses: R.Manimekalai@icar.gov.in (R. Manimekalai), smitabiot@gmail.com (S. Nair), naganeeswaran@gmail.com (A. Naganeeswaran), anithakarun2008@gmail.com (A. Karun), malhotraskraj@yahoo.com (S. Malhotra), dccd@nic.in (V. Hubbali).

https://doi.org/10.1016/j.btre.2017.12.005
Received 21 September 2016; Received in revised form 14 March 2017; Accepted 5 December 2017
Available online 03 January 2018
2215-017X/ © 2018 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

sequencing technologies like Illumina HiSeq, 454 Pyrosequencing, SOLiD sequencing, etc are being widely used for getting the overview of expressed genes in uncharacterized genomes. The RNA Seq analysis of coconut transcriptome using Illumina technology has been reported. Overall, 57, 304 unigenes were reported, of which, 99.9% were novel compared to available coconut EST sequences [10]. With this background, the present work was designed to obtain the RNA Seq information of arecanut palm using Illumina sequencing and de novo assembly. This would generate ample amount of sequence information on *Areca catechu* L. transcriptome. Apart from this, the information generated here would form a basis for further gene expression studies in arecanut palm with regard to stress tolerance or expression studies for flavanoid and alkaloid principles.

2. Materials and methods

2.1. Tissue sampling and RNA isolation

Spindle leaf tissue samples from nine year old arecanut cultivar South Canara Local during fruit development stage were collected from Sullia (12.5° N, 75.3° E), Karnataka, India. This location is endemic for the yellow leaf disease which is a major problem affecting arecanut in South India. We had taken the samples from healthy areca palm from the field belong to Mr. Naik with his permission. The tissue sample was preserved in RNA Later solution (Life Technologies) before RNA isolation. Total RNA was purified from the tissue using Trizol reagent (Life Technologies). The quality and purity of the extracted RNA were assessed spectrophotometrically. The RNA integrity number (RIN) was observed with Bioanalyzer (Agilent Technologies). RIN value of 6.5 is the threshold for Illumina sequencing.

2.2. Paired end library preparation and RNA sequencing

The RNA seq library preparation was performed with 1 μg RNA sample using the TrueSeq Sample Prep Kits (Illumina) as per the protocol. Briefly, the mRNA molecules were purified with poly-T magnetic beads, fragmented and subjected to complementary DNA (cDNA) synthesis. After end repair process with single adenine residue and adapter ligation, final cDNA library was generated using PCR. Bioanalyzer plots were used throughout for quality check. Illumina Hiseq2000 sequencing method was used for paired-end read generation. Sequencing was carried out in Scigenom, Cochin, Kerala, India using HiSeq2000 technology.

2.3. Raw read processing and de novo assembly

Illumina paired end raw reads were checked for quality parameters such as adaptor contamination, base quality score distribution, average base content per read and GC distribution. Adaptor sequence and low quality regions were trimmed from the raw reads to avoid specific sequence bias during assembly. The reads with average quality score less than 20 were filtered out. Reads contaminated with Illumina adapter were soft masked before assembly. First 17 bases and last 2 bases were trimmed from paired end reads to avoid specific sequence bias and low quality bases. After trimming, we obtained 51 million reads of 82 bp × 2 lengths. Trimmed reads were assembled using SOAP *de novo* 31mer program with default parameters [11]. The contigs obtained were then assembled into scaffolds and finally into transcripts. Assembled transcripts with greater than 150 bp lengths were used for further transcript expression estimation and downstream functional analysis.

2.4. Expression analysis

Trimmed reads were aligned to the assembled transcripts (length ≥ 150 bp) using Bowtie2 (mis-match = 1 and seed length = 31 bp) program [12]. The FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values were used for evaluation of the expressed value and quantification of transcripts [13]. For downstream annotation and differential expression analysis, we focused only on those transcripts with length of ≥ 150 bp and expression of ≥ 1 FPKM.

2.5. Functional annotation

The assembled transcripts with significant gene expression values were subjected to similarity search against NCBI non-redundant protein database using BlastX (E-value ≤ 10^{-5} and similarity score ≥ 40%) program [14]. Blast annotations (NCBI id) were mapped back to the Uniprot protein database and Gene Ontology terms (molecular function, biological process and cellular component) were extracted from the Uniprot database (http://www.uniprot.org/).

2.6. Pathway analysis and simple sequence repeats (SSRs) prediction

Pathway annotations were performed using Kyoto Encyclopedia of Genes and Genome (KEGG Automation Annotation Server (KEGG KAAS) program [15]. The transcript sequences were mapped to KEGG pathway database using KAAS (Online) server [16]. In the KAAS annotation, plant models were used as reference for metabolic pathway identification. The SSR prediction and corresponding primer designing were attempted using modified version of SEMAT program using default parameters [17].

2.7. Comparison of arecanut transcripts with other palms sequence (coconut, oil palm and date palm) information

Totally 57,175 coconut transcripts (ref) and 37,492 oilpalm EST sequences were retrieved from NCBI database. Then, 28,889 date palm predicted mRNA sequences were downloaded from Weill Cornell Medical College database, Qatar (http://qatar-weill.cornell.edu/research/datepalmGenome/). BlastN based similarity search was carried out with the E-value 10^{-5}.

3. Results

3.1. Raw read processing and de novo assembly

The illumina sequencing run generated a total of 56,321,907 paired end reads of 101 bp length consisting of 11.3 Gb nucleotides (Accession: PRJNA287587 ID: 287587). The quality check showed the average base quality was above Q20 (error-probability ≥ 0.01) for most of the reads. The raw reads were trimmed before performing the assembly. The first 17 bases and last 2 bases were trimmed from all forward (R1) and reverse (R2) reads. After pre-processing, the trimmed file of 51,175,929 paired end reads consisted of 8.4 Gb with 82 bp average length of reads (Table 1). The trimmed reads were assembled using SOAP *de novo* program to give 220,917 assembled transcripts. To get high quality annotation, we chose the transcripts greater than 150 bp length for the downstream analysis. Totally 118,847 transcripts (length ≥ 150) were obtained from the assembly. The length of transcripts ranged between 150 bp and 7751 bp, the average length being 470 bp. The overall

Table 1
Summary of raw and trimmed reads from sequencing results.

Parameters	Raw read	Trimmed read
Number of paired end reads	56,321,907	51,175,926
Number of bases (Gb)	5.69	4.20
GC%	49.01	46
Read length (bp)	101*2	82*2

Download English Version:

https://daneshyari.com/en/article/7234959

Download Persian Version:

https://daneshyari.com/article/7234959

Daneshyari.com