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

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

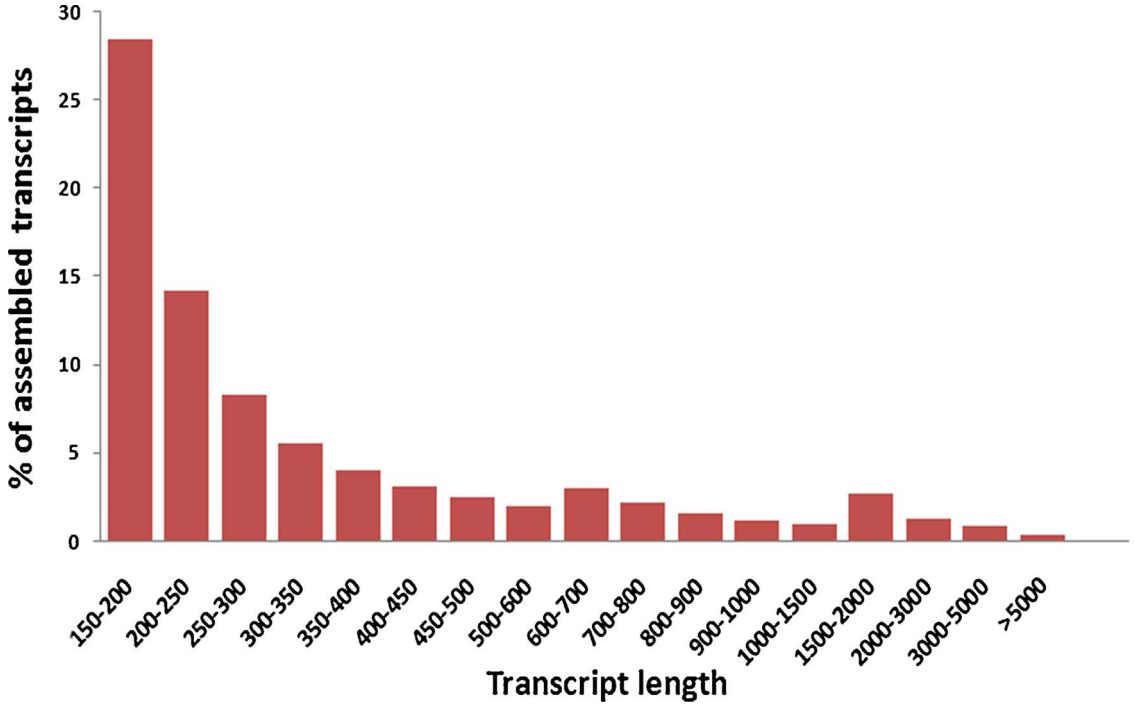

Fig. 1. Length distributions of transcripts in the arecanut leaf transcriptome.

length distribution is depicted in Fig. 1. The average GC content was found to be 46% and the N50 value was ~ 650 bp.

3.2. Transcript expression estimation and functional annotation

During Bowtie2 alignment, 41,437,150 (81%) reads were aligned to the assembled transcripts. Overall, 48,783 transcripts had FPKM value ≥ 1, the average length being 800 bp. The FPKM distributions are shown in Fig. 2. Transcripts with ≥ 150 bp length and ≥ 1 FPKM value were used for the functional annotation. The highest expressed transcript in arecanut genome correspond to a gene involved in flavonoid biosynthesis, leucoanthocyanidin reductase [EC:1.17.1.3], which had a read count of 376,121 and FPKM of 3572, followed by ribulose-bisphosphate carboxylase [EC:4.1.1.39]. The genes involved in flavonoid and terpenoid biosynthesis which are highly expressed (≥100 FPKM) are given in Table 2. The similarity search against NCBI non-redundant protein database using BlastX resulted in 32,485 hits, thus giving annotation for 67% of overall transcripts (Supplementary table S1). The top blast hits of each transcript were studied and the organism name was extracted. Overall, 16.5% of the matches were with *Vitis vinifera* followed by 8.3% with *Oryza sativa,* 5.7% with *Zeya mays,* 5.2% *Theobroma cacao* (Fig. 3). Among the total significant BlastX hit transcripts,

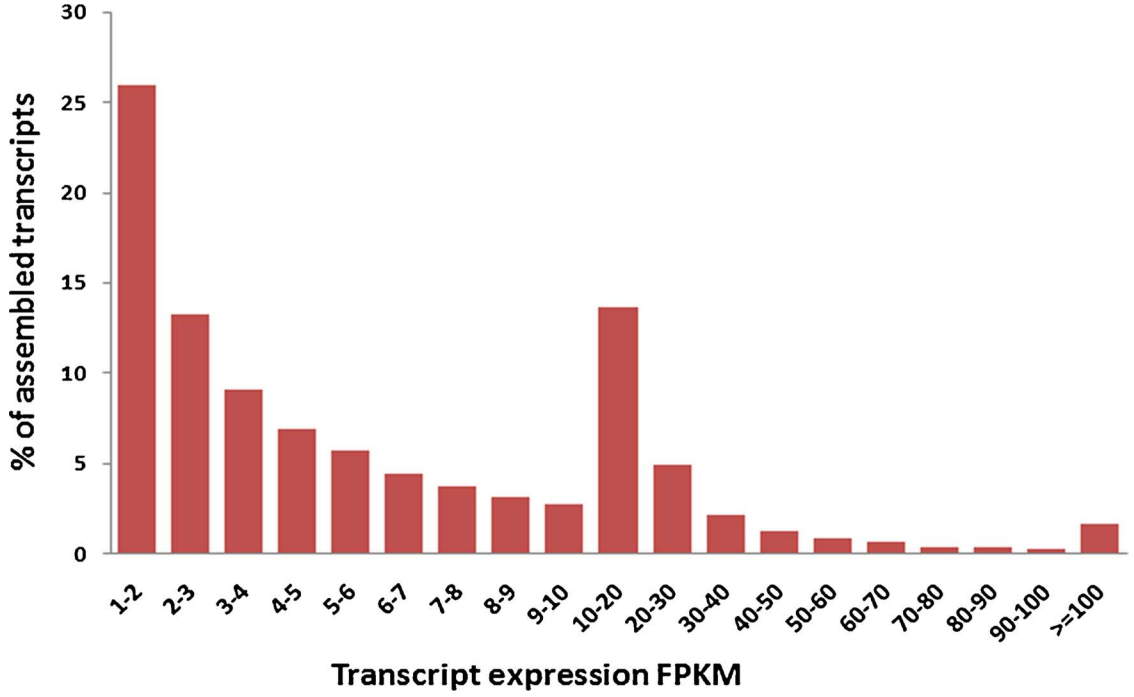

Fig. 2. Percentage of transcripts in arecanut leaf transcriptome based on expression values (Fragments per kilobase million, FPKM).

Table 2
Highly expressed transcripts related to flavonoid and terpenoid biosynthesis based on FPKM values obtained from RNA- seq data of arecanut.

Transcript Id	Gene name	Pathway	Read Count	FPKM
436630	Leucoanthocyanidin reductase [EC:1.17.1.3]	Flavonoid biosynthesis	376,121	3572.9
434102	Enyl diphosphate reductase [EC:1.17.1.2]	Terpenoid backbone biosynthesis	134,099	1483.7
426667	Flavanone 4-reductase [EC:1.1.1.219 1.1.1.234]	Flavonoid biosynthesis	33,462	513.2
439086	1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]	Terpenoid backbone biosynthesis	16,153	123.4
414549	chalcone isomerase [EC:5.5.1.6]	Flavonoid biosynthesis	5299	115.3
423583	Cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]	Phenylpropanoid biosynthesis	6463	109.8
432856	flavonoid 3′-monooxygenase [EC:1.14.13.21]	Flavonoid biosynthesis	8953	105.8

11,680 transcripts were annotated using UniProt database and gene ontology terms were extracted. Based on gene ontology analysis, 9222 biological process, 11268 molecular function and 7574 cellular components terms were identified (Fig. 4a–b) (Supplementary table S2).

3.3. Pathway analysis

Metabolic pathway analysis of the assembled transcripts identified 124 plant related pathways. A total of 1778 enzymes in our transcripts could be matched to the KEGG pathways. We obtained 2250 transcripts involved in metabolic pathways including carbohydrate metabolism (553), energy metabolism (264), lipid metabolism (310), nucleotide metabolism (216), amino acid metabolism (450), glycan biosynthesis and metabolism (99), metabolism of cofactors and vitamins (184), metabolism of terpenoids and polyketides (85) and biosynthesis of other secondary metabolites (89). A total of 1460 transcripts were identified to be involved the genetic information processing pathways including transcription, translation, protein modification, etc. The other major pathways included environmental information processing (145) and cellular processes (206) and organismal systems (117) (Supplementary table S3). As arecanut is a source of carotenoids, tannins and alkaloids, we investigated the enzymes involved in the secondary metabolite production pathway such as biosynthesis of ubiquinone and other terpenoid-quinones; terpenoid backbone; isoquinoline alkaloids, tropane, piperidine and pyridine alkaloids, carotenoids, flavonoids, brassinosteroids, phenylpropanoid, stilbenoid, diarylheptanoid and gingerol. The genes coding for the enzymes in the ubiquinone and other terpenoid-quinone biosynthesis have higher read count, as high as > 6600 for (MPBQ/MSBQ methyltransferase involved in plastoquinone biosysnthesis), and others have high FPKM values for tocopherol O-methyltransferase [EC: 2.1.1.95], naphthoate synthase [EC: 4.1.3.36], homogentisate solanesyltransferase, 4-coumarate–CoA ligase [EC: 6.2.1.12] and aminotransferases. The genes involved in the terpenoid biosynthesis, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase [EC: 1.17.1.2] had high read count (134,099) and FPKM value of 1483. This is the key enzyme involved in terpenoid biosynthesis [18]. The arecanut palm is rich in flavonoids and interestingly the gene involved in flavonoid biosynthesis such as leuco anthocyanidin 4-reductase (LAR) which converts the flavan 3,4-diol to catechins [19] had very high read count and FPKM values of 376,121 and 3572 respectively. This gene had highest expression level in whole of arecanut leaf transcriptome. The other enzymes like dihydroflavonol 4-reductase, flavonoid 3′-monooxygenase [EC:1.14.13.21], chalcone isomerise, chalcone synthase in the pathway also had high read count and FPKM values. The enzymes involved in phenyl propanoid pathway such as oniferyl-aldehyde dehydrogenase [EC: 1.2.1.68], cinnamyl-alcohol dehydrogenase [EC: 1.1.1.195], caffeoyl-CoA O-methyltransferase [EC: 2.1.1.104], COMT; caffeic acid 3-O-methyltransferase [EC: 2.1.1.68] and phenylalanine ammonia-lyase [EC: 4.3.1.24] had high read count (Supplementary table S3). The genes involved in carotenoid biosynthetic pathway could be identified from the transcriptome data.

3.4. Simple sequence repeat (SSR) marker prediction

The SSR regions in assembled transcripts were predicted using modified SEMAT SSR pipeline with default parameters. A total of 6853 SSR regions were identified (Table 3). The distribution of SSRs is shown in Fig. 5. Overall, 3963 di repeats, 2602 tri repeats, 194 tetra repeats, 45 penta repeats and 49 hexa repeats were found in arecanut leaf transcriptome. Possible SSR specific primers were also designed and provided in the Supplementary table S4.

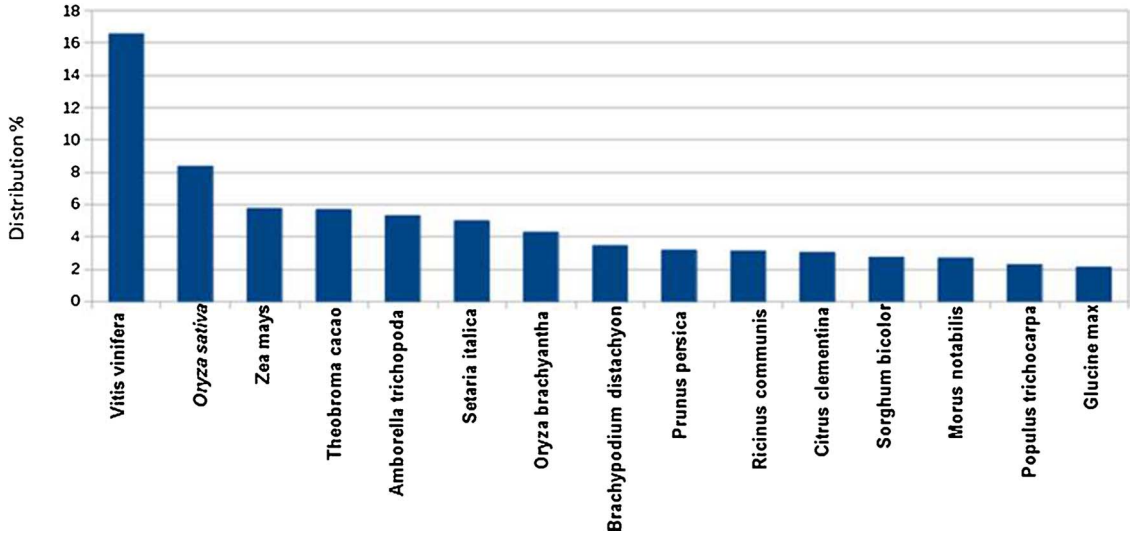

Fig. 3. Species wise distribution of blast hits of arecanut transcripts.
The highest percent identity was observed with *Vitis vinifera* sequences.

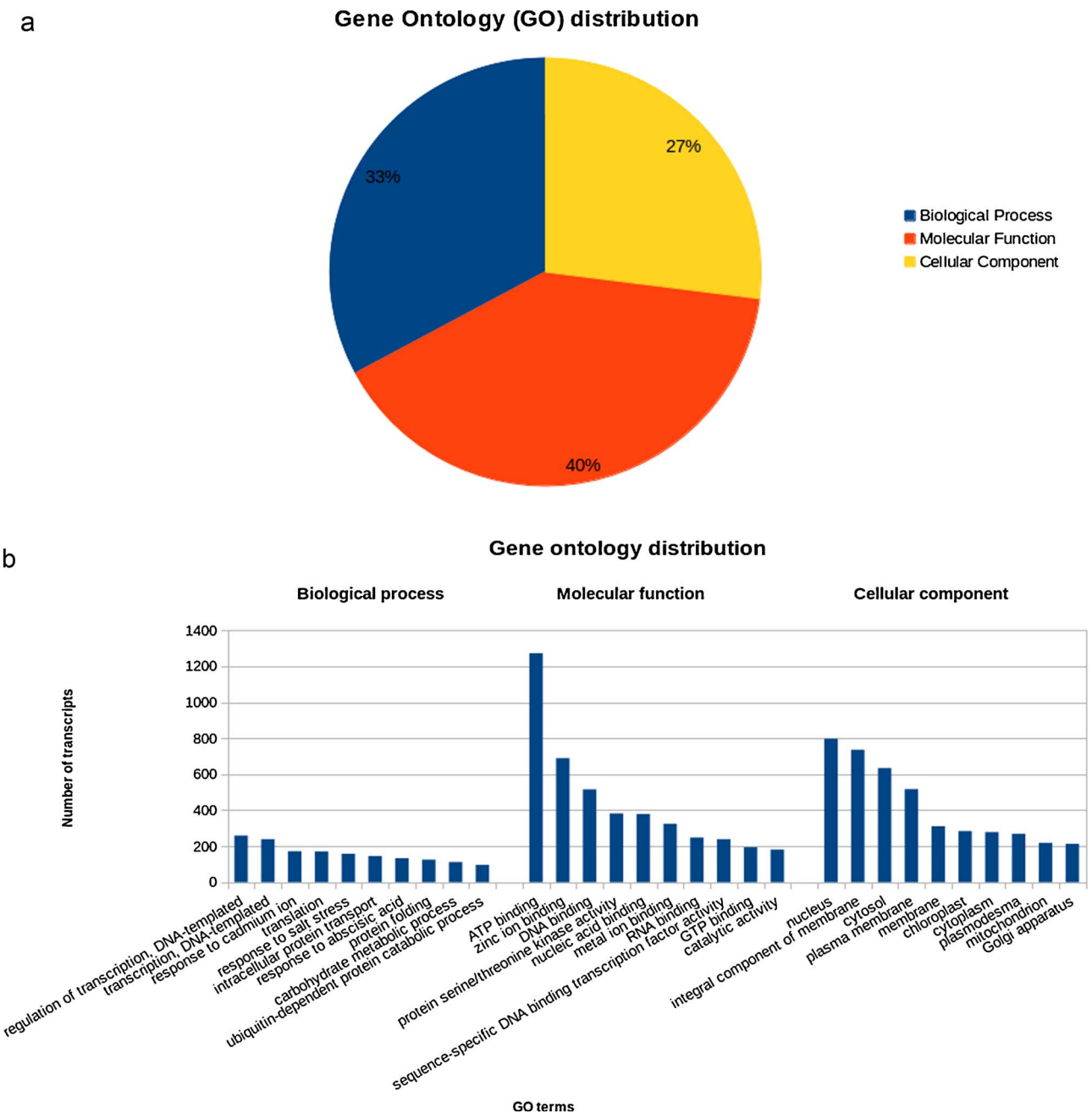

Fig. 4. Gene ontology (GO) classification of assembled transcripts using UniProt database.
a. GO terms in the biological process, molecular function and cellular component.
b. Number of transcripts annotated with GO terms

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

The comparison of the arecanut transcriptome sequences with coconut, oil palm and date palm sequences showed high similarity between the arecanut assembled sequences and other palm sequences. Overall, 54.5% of arecanut sequences aligned with coconut sequences, 44% with date palm sequences and 25.6% with oil palm sequences (Table 4).

4. Discussion

Arecanut palm, *Areca catechu* L. provides the betel nut which is widely used as a masticatory nut and also finds a major role in religious ceremonies. The medicinal use of arecanut is also well known. Arecanut genetic improvement and hybridization have been conducted in the past at Central Plantation Crops Research Institute (CPCRI). Six high yielding varieties (Mangala, Sumangala, Sreemangala, Mohitnagar, Swarnamangala and Kahikuchi) and two hybrids (VTLAH1 and VTLAH2) are available for cultivation in India. But the available information on the genome of arecanut is sparse. There is no data on arecanut genes in public domain databases. Of late, the next generation sequencing techniques enabled generation of ample sequence information within a short span of time. The *de novo* assembly and characterization of bark transcriptome of rubber tree using Illumina sequencing has been reported [20]. We attempted the arecanut transcriptome sequencing and *de novo* assembly to unveil the huge amount of novel genomic information on the palm. We selected the cultivar of

Table 3
Summary of the simple sequence repeat (SSR) types in the arecanut transcriptome.

Description	Number
Total number of identified SSRs	6853
Number of SSR containing sequence	6091
Number of sequences containing more than 1 SSR	673
Number of SSRs present in simple form	6435
Number of SSRs present in compound form	418
Distribution to different repeat type classes –	
Di repeats	3963
Tri repeats	2602
Tetra repeats	194
Penta repeats	45
Hexa repeats	49

Definement of microsatellites (unit size/minimum number of repeats) (2/6) (3/5) (4/5) (5/5) (6/5).

arecanut, South Canara Local which is largely grown in South Canara district of Karnataka and Kasaragod district of Kerala, India. We got 48,783 good quality transcripts (length ≥ 150, FPKM > 1), the average transcript length being 800 bp. In the coconut, RNA seq work reported earlier using Illumina sequencing, Fan et al. (2013) obtained 54.9 million short reads which on *de novo* assembly produced 57,304 unigenes with an average length of 752 base pairs. So, the results are comparable with similar works done on palms. Our initial results on overall transcriptome of arecanut could be further substantiated with comparative transcriptome studies in relation to biotic or abiotic stress conditions. A major biotic challenge faced by arecanut palms in South India is the Yellow Leaf Disease and this could be a problem addressed in future based on further RNA seq studies.

With BlastX, 67% of our transcripts were annotated. The enzymes in the transcripts were mapped in to KEGG pathways. Hence, the data we provided is a backbone for functional genomics studies in arecanut including, but not limited to, the isolation and characterization of enzymes involved in specific metabolic pathways especially the carotenoid biosynthesis. Alignment of the arecanut transcriptome with other palms sequences (ESTs, transcriptome and mRNA) showed an overall high similarity between the arecanut assembled sequences and other palm sequences. Further, 6853 SSR regions were identified in the

Table 4
Comparative study of arecanut transcriptome with coconut, oil palm and date palm sequences.

Name	No. of EST/mRNA/ Transcriptome sequence	No. of Arecanut sequence alignment	Reference representation
Coconut	57175	64781(~54.5%)	28761 (~50.3%)
Oil palm	37492	30473 (~25.6%)	14631 (~39%)
Date palm	28889	52351 (~44%)	17414 (60.2%)

transcriptome. Earlier report on comparative study of date palm linkage groups with oil palm genome, [21] observed that the two genomes maintained high levels of synteny. Hence, our arecanut transcriptome information will also help in mining genes and markers across the palm family. This is the first report of arecanut transcriptome sequencing and analysis and in future this would form the basis for genetic improvement studies in arecanut. It will also contribute greatly in the understanding of palm genomes on the whole.

The application of genomics technologies has expedited the discovery of secondary metabolite biosynthetic pathway genes that encode enzymes and regulatory proteins with novel functions. By large-scale, transcriptomics analyses provide initial hints about the biosynthetic processes. Even though arecanut is a source of alkaloids and tannins, no molecular evidence is available on their biosynthesis pathway genes. Hence this report presents the first information on the genes in the biosynthetic pathways of alkaloids, flavonoids and terpenoids in arecanut. Interestingly, there is high level of transcripts in the carotenoid biosynthesis pathway genes implying that the arecanut palm is a potential source of carotenoids which could be explored commercially.

5. Conclusion

To conclude, we have generated arecanut transcriptome sequence and this first report on arecanut transcriptome assembly. The total clean reads was about 11.3 Gb from which a total of 57,304 unigenes were obtained. The functional annotation and classification were done using BLAST against public databases (Swiss-Prot, GO, KEGG and COG). Genic SSRs identified in the present study would help in genetic

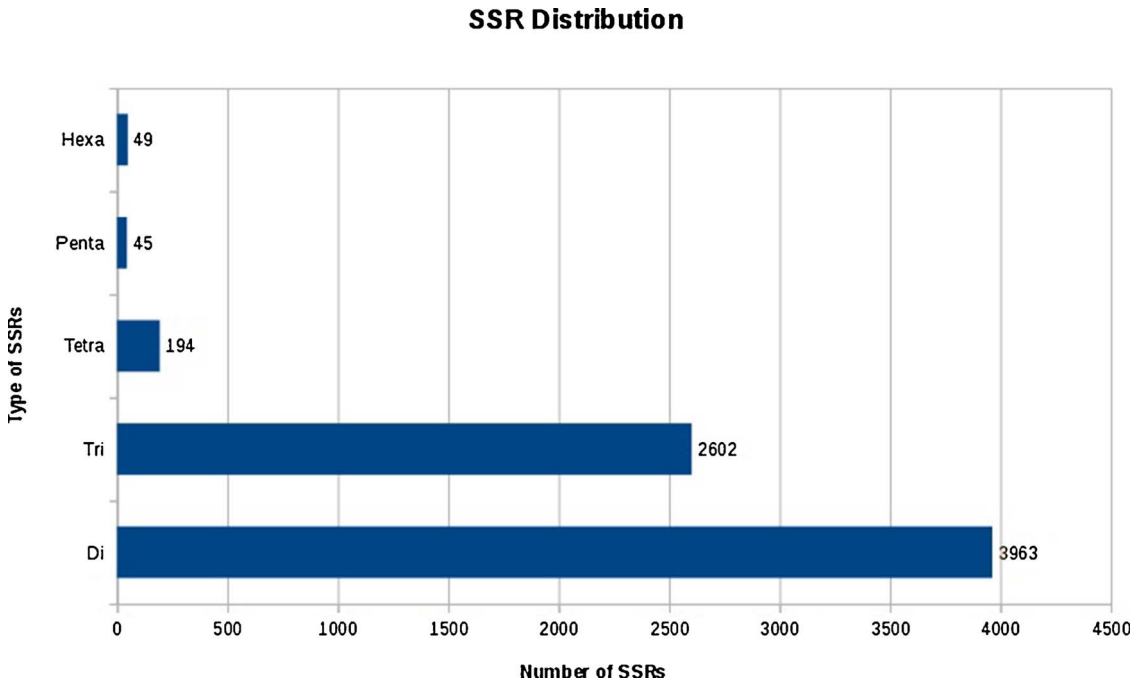

Fig. 5. Simple sequence repeat (SSR) types in the arecanut transcriptome.

characterization of arecanut. The genes in the biosynthetic pathways of alkaloids, flavonoids (> 3500 FPKM) and terpenoids were found to be highly expressed. The information on the KEGG metabolic pathways elucidates the secondary metabolite pathway genes in areca palm.

Data archiving

All the raw reads have been submitted as sequence read archive (SRA) in NCBI (Accession: PRJNA287587 ID: 287587).

Disclosure

All authors have approved the final article should be true and included in the disclosure. There is no conflict of interest.

Author contribution

R.M.: Conceptualization, writing manuscript
S.N.: RNA isolation, preparation of samples, editing the manuscript
A.N.: Data analysis, annotation, assembly and pathway analysis
A.K.: Conceptualization, editing the MS.
S.M.: Data analysis, editing the MS
Hu: Conceptualization, editing the MS

Acknowledgement

The work was supported by ICAR – National Agriculture Science Fund. Grant number: F.No. NFBSFARA/BS-3013/2012-13

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.btre.2017.12.005.

References

[1] G.W. Staples, R.F. Bevacqua, C.R. Elevitch (Ed.), Species Profiles for Pacific Island Agroforestry, Permanent Agriculture Resources (PAR), Holualoa, Hawaii, 2006, <http://www.traditionaltree.org>.
[2] M.M. Rao, Introduction, in: K.V.A. Bavappa, M.K. Nair, T.P. Kumar (Eds.), The Arecanut Palm (*Areca Catechu* Linn.), Central Plantation Crops Research Institute, Kasaragod Kerala, 1982, pp. 1–7.
[3] S. Giri, J.R. Idle, C. Chen, T.M. Zabriskie, K.W. Krausz, F.J. Gonzalez, A metabolomic approach to the metabolism of the arecanut alkaloids arecoline and aracaidine in the mouse, Chem. Res. Toxicol. 19 (2006) 818–827.
[4] M. Kumar, U.R. Moon, A. Mitra, Rapid separation of carotenes and evaluation of their in vitro antioxidant properties from ripened fruit waste of *Areca catechu* –a plantation crop of agro-industrial importance, Ind. Crops Prod. 40 (2012) 204–209.
[5] R. Manimekalai, K.P. Deeshma, K.P. Manju, V.P. Soumya, M. Sunaiba, N. Smitha, K.S. Ananda, Molecular marker-based genetic variability among Yellow Leaf Disease (YLD) resistant and susceptible arecanut (*Areca catechu* L.) genotypes, Indian J. Hortic. 69 (2012) 455–461.
[6] M. Sankaran, P. Chandrasekar, D.R. Singh, V. Damodaran, Assessment of genetic diversity among arecanut accessions by using RAPD markers, Indian J. Hortic. 70 (2013) 428–430.
[7] B.G. Bharath, K.S. Ananda, J. Rijith, N.R. Nagaraja, K.P. Chandran, K. Anitha, M.K. Rajesh, Studies on genetic relationships and diversity in arecanut (*Areca catechu* L.) germplasm utilizing RAPD markers, J. Plant. Crops 43 (2015) 117–125.
[8] E.K. Al-Dous, B. George, M.E. Al-Mahmoud, M.Y. Al-Jaber, H. Wang, Y.M. Salameh, E.K. Al-Azwani, S. Chaluvadi, A.C. Pontaroli, J. DeBarry, V. Arondel, J. Ohlrogge, I.J. Saie, K.M. Suliman-Elmeer, J.L. Bennetzen, R.R. Kruegger, J.A. Malek, *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*), Nat. Biotechnol. 29 (2011) 521–527.
[9] R. Singh, O.M. Abdullah, E.L. Low, M.A.A. Manaf, R. Rosli, R. Nookiah, L. Cheng-Li Ooi, S. Ooi, K.M. Chan, M.A. Halim, N. Azizi, N. Jayanthi, B. Bacher, N. Lakey, S.W. Smith, D. He, M. Hogan, M.A. Budiman, E.K. Lee, R. DeSalle, D. Kudrna, J.L. Goicoechea, R.A. Wing, R.K. Wilson, R.S. Fulton, J.M. Ordway, R.A. Martienssen, R. Sambanthamurthi, Oil palm genome sequence reveals divergence of inter fertile species in Old and New worlds, Nature 500 (2013) 335–339.
[10] H. Fan, Y. Xiao, Y. Yang, W. Xia, A.S. Mason, Z. Xia, F. Qiao, S. Zhao, H. Tang, RNA-seq analysis of *Cocos nucifera*: transcriptome sequencing and subsequent functional genomics approaches, PLoS One (2013) e59997, http://dx.doi.org/10.1371/journal.pone.0059997.
[11] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, J. Wang, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience 1 (2012) 18.
[12] B. Langmead, S. Salzberg, Fast gapped-read alignment with bowtie 2, Nat. Methods 9 (2012) 357–359.
[13] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNASeq, Nat. Methods 5 (2008) 621–628.
[14] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[15] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs, Nucleic Acids Res. 38 (2010) 355–360.
[16] Y. Moriya, M. Itoh, S. Okuda, A. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server, Nucleic Acids Res. 35 (2007) W182–W185.
[17] N.S. Asari, R. Manimekalai, E.A. Subbian, K.P. Manju, S.K. Malhotra, A. Karun, Standalone EST microsatellite mining and analysis tool (SEMAT): for automated EST-SSR analysis in plants, Tree Genet. Genomes (2014), http://dx.doi.org/10.1007/s11295-014-0785-2.
[18] A.A. Ramosa, A.R. Marquesa, M. Rodriguesa, N. Henriquesa, A. Baumgartnerb, R. Castilhoa, Molecular and functional characterization of a cDNA encoding 4-hydroxy-3-methylbut-2-enyl diphosphate reductase from *Dunaliella salina*, J. Plant Physiol. 166 (2009) 968–977, http://dx.doi.org/10.1016/j.jplph.
[19] J. Pfeiffera, C. Kuhnela, J. Brandta, D. Duya, N. Punyasirib, G. Forkmanna, T.C. Fischer, Biosynthesis of flavan 3-ols by leucoanthocyanidin 4-reductases and anthocyanidin reductases in leaves of grape (*Vitis vinifera* L.), apple (Malus x domestica Borkh.) and other crops, Plant Physiol. Biochem. 44 (2006) 323–334.
[20] D. Li, Z. Deng, B. Qin, X. Liu, Z. Men, De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.), BMC Genomics 13 (2012) 192, http://dx.doi.org/10.1186/1471-2164-13-192.
[21] L.S. Mathew, M. Spannagl, A. Al-Malki, B. George, M.F. Torres, E.K. Al-Dous, E.K. Al-Azwani, E. Hussein, S. Mathew, K.F. Mayer, Y.A. Mohamoud, K. Suhre, J.A. Malek, A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms, BMC Genomics 15 (2014) 285.