Physiological and Molecular Plant Pathology 100 (2017) 255–263

Contents lists available at ScienceDirect

Physiological and Molecular Plant Pathology

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

Deciphering genes associated with root wilt disease of coconut and development of its transcriptomic database (CnTDB)

Sandeep Kumar Verma [1], Rahul Singh Jasrotia [1], M.A. Iquebal, Sarika Jaiswal, U.B. Angadi, Anil Rai, Dinesh Kumar*

Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012, India

ARTICLE INFO

Article history:
Received 1 March 2017
Received in revised form
29 March 2017
Accepted 29 March 2017
Available online 30 March 2017

Keywords:
Coconut
Differentially expressed genes
GO
Markers
RWD
Transcriptome

ABSTRACT

Coconut (*Cocos nucifera* L.) has global significance in agriculture and industries due to its nutritional and medicinal properties. Coconut Root Wilt Disease (RWD) causes huge economic loss, thus molecular approach for improved varieties is needed. Since whole genome sequence is unavailable, transcriptomic approach is imperative for deciphering pathways as well as genic region marker discovery from contrasting genotypes. This is the first report of RWD transcriptome database having candidate genes and pathway discovery along with genic simple sequence repeats, SNPs, indels to be used as functional domain markers. A relational database, CnTDB (http://webtom.cabgrid.res.in/cntdb/), based on three-tier architecture has been developed having 285235 transcripts with all blast information, annotations, pathways, 22021 DEGs, transcriptional factors, 10126 and 97117 SSR markers mined from DEGs and *de novo* transcriptome assembly, respectively. Putative markers with primers can be valuable genomic resource in endeavor of RWD resistant variety development for higher coconut productivity.

1. Introduction

The coconut palm, an economically important crop, scientifically known as *Cocos nucifera* L., belongs to the family Arecaceae (or Palmae) and is widely grown in tropical and sub-tropical zones having varying climatic and soil conditions. It is the only species from genus *Cocos* of subfamily *Cocoideae* that includes 27 genera and 600 species [1]. The coconut palm also known as *kalpavriksha* is one of the five major *devavrikshas*. It is a diploid with 32 chromosomes (2n = 32) and falls under two groups. i.e., tall and dwarf [1]. The fruit of this plant is used in social and religious functions in India and is also called as *lakshmiphal* [2].

Cocos nucifera has global significance and is grown in around 95 countries worldwide, where Indonesia, Philippines and India are the top countries with maximum production, i.e. 1910213 tonnes, 1469628 tonnes and 11078873 tonnes, respectively (FAOSTAT 2014). In India, it is grown over area 2140000 hactares giving yield of 51770 Hg/Ha.

Coconut has wide application in agriculture as well as industries and is known to support livelihood of poor people as well as helps in environmental sustainability. The coconut kernel, coconut milk, copra, coconut oil, coconut leaves etc. are the beneficial parts of it. Products are in the form of food, fibre, oil, fertilizers, garments, construction, spa ingredients, furniture, fashion accessories, building materials, oleochemicals and biofuels [3]. The oil has nutritional and medicinal properties like it does not contain cholesterol, easy to digest and absorb, has vitamin E, medium chain triglycerides and low content of omega 6 fatty acid. It also helps in the absorption of vitamins, minerals and amino acids along with its antimicrobial properties [4].

Biotic stresses in coconut palm is caused by various pathogens like bacteria, fungi, viruses, phytoplasma etc. leading to disease like bacterial bud rot, algal leaf spot, anthracnose, damping-off, leaf blight, leaf spot, root rot, root wilt disease, lethal decline etc. Of these, Root Wilt Disease (RWD), is the most devastating biotic stress caused by phytoplasma leading to major economic loss around 968 million nuts, annually, which is about 3000 million rupees [5]. In advanced stage of the disease, the yield drop up to 80% has also been reported. The symptoms of this disease are wilting, drooping and flaccidity in leaves along with ribbing, paling and necrosis of leaflets [3].

Genetic enhancement for the productivity and tolerance to abiotic and biotic stress in coconut palm is impeded by long pre-

* Corresponding author.
E-mail address: dineshkumarbhu@gmail.com (D. Kumar).
[1] Contributed equally.

http://dx.doi.org/10.1016/j.pmpp.2017.03.011
0885-5765/

bearing period and slow growth of the palm. Gene pool enrichment has been one of the major concerns of coconut improvement programs. A high degree of variability is accessible in landraces in agriculturalists fields, particularly for adaptation to features of agro-ecosystem, which needs to be exploited. Landraces of plants were found to be invaluable for providing sources of resistance against abiotic and biotic stress and also development of agronomic traits.

Candidate gene discovery from the contrasting genotypes of coconut palm plays important role in variety improvement. In absence of whole genome sequence, study of contrasting genotypes is the ideal source to knowledge discovery, as taken up in this study. Also, there is need of enriched molecular information of coconut palm for better breeding and trait improvement. Very limited studies on the genomic information of coconut have been reported in literature out [1,6]. Not much attempts have been made in molecular biotechnologies for improvement of important traits in *Cocos nucifera* due to the absence of genome sequence information, though the genome sequence of closest crop, i.e., oil palm is already reported [7].

To minimize the yield losses by the root wilt disease, identification of candidate genes responsible for this disease can be a great aid in development of coconut cultivars resistant to RWD. The present work aims at discovery of differentially expressed genes (DEGs) using extreme genotypes associated with RWD in coconut along with discovery of Simple Sequence Repeat (SSR), genic SNPs and Indels. Identification and annotation the differentially expressed genes in root wilt of coconut can be used to delineate pathway associated with plant immunity. Identification of novel genes may play role in providing resistance to palm against RWD. Also, the transcriptome database of coconut developed in this work is the sole genomic resource of coconut and can be valuable for future endeavor of coconut improvement.

2. Materials and methods

2.1. Dataset, its preprocessing and de novo assembly

Paired End Illumina data of *Cocos nucifera* from two extreme genotypes, i.e., resistant (Chowghat Green Dwarf) and susceptible (West Coast Tall) for root wilt disease available at SRA at National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov//) with project IDs SRX436961 (CGD) and SRX437650 (WCT) were used in the study. SRAtoolkit was used to convert SRA data into Fastq, resulting in separate files for forward and reverse data for each sample. FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for visualization of reads quality before and after pre-processing. Trimmomatic tool version 0.33 was used for removal of low quality reads, trimming of bases from 3′ and 5′ end, maintaining the phred-score $\leq$ 20 [8,9]. Using these two preprocessed extreme genotypes, *viz*., resistant and susceptible for root wilt disease transcriptome data, the reads were used for *de novo* transcriptome assembly using *Trinity* assembler v2.0.6 [10,11], that uses *De Bruijn* graph algorithm with a default *k*-mer value 25 and CAP3 assembler used for removal redundant sequences [12].

2.2. Abundance estimation and identification of differential expressed genes

Alignment and mapping of reads from RWD resistant and susceptible coconut transcriptome data to reference *de novo* transcriptome assembly was performed using bowtie tool [13]. For mapping of reads onto *de novo* transcriptome assembly only two mismatches were allowed. RNA-Seq by Expectation-Maximization (RSEM) tool [14] was used to calculate the expression values in the form of fragments per kilo base of exon per million mapped reads (FRKM). Identification of differential expressed genes was performed using edgeR package (Empirical analysis of Digital Gene Expression in R) [15] of Bioconductor with standard parameters i.e. FDR 0.05 and log fold change of 2 (i.e., Log FC).

2.3. Functional annotation and enrichment analysis

Homology search of *de novo* transcriptome assembly as well as differentially expressed genes was performed with Blastx algorithm of standalone local ncbi-blast-2.2.31+ [16] against the NCBI non-redundant (version nr.36) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using threshold E-value 1e-3. We also used some other blast parameters for significant results such as max_target_seqs and num_alignments. Annotation, functional characterization and pathway analysis of DEGs and full transcripts were performed using Blast2Go Pro version 3.1 software [17]. The Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Enzyme Codes are reported in the study. The genes were categorized into molecular, cellular and biological functions. Transcriptional factors identification was done at PlantTFDB 3.0 (http://planttfdb.cbi.pku.edu.cn/download.php) database [18].

2.4. Markers' identification

The microsatellite markers, SNPs and Indels were identified from coconut transcriptomic data. Microsatellite SSR putative markers from differential expressed genes as well as *de novo* transcriptome assembly of *Cocos nucifera* were mined using Perl script of MISA-MIcroSAtellite identification tool [19]. In order to get the putative SSR markers, 10 repeating units for mononucleotides, 6- repeating units for dinucleotide and 5 repeating units for tri-, tetra-, penta- and hexa-nucleotides were used [20,21]. SSRs or putative microsatellites are generated during replication, especially with sequence having more than 8 repeat units because of slippage event [22]. PRIMER3 tool [23] was used for designing of primers from these markers. Burrows-Wheeler Aligner (BWA) tool [24] was used to align resistant and susceptible reads of *Cocos nucifera* to *de novo* transcriptome assembly separately. SNPs and Indels were discovered by SAMtools package [25].

2.5. Database architecture

Coconut transcriptome database (CnTDB) is an online relational database with “three-tier architecture” with a client tier, middle tier and database tier. In client tier, web pages have been developed for browsing the database and defining queries by user. The database has been developed with various tables for assembled transcripts/contigs and mined markers in MYSQL. Server side scripting has been done using PHP in middle tier for database connectivity, executing query and fetching data from the database. The database is available at http://webtom.cabgrid.res.in/cntdb/.

3. Results

3.1. Quality assessment and de novo assembly

Paired end reads of two extreme genotypes of coconut, *i.e.*, RWD resistant (CGD) and RWD susceptible (WCT) transcriptome data was generated using Illumina HiSeq 2000. A total of 119333177 and 111025986 raw reads, respectively with read length 101 were generated. After preprocessing of these data sets, 1028864 (i.e., 0.93%) and 1036718 (i.e., 0.87%) poor quality reads of RWD resistant and susceptible varieties, respectively were removed. Finally

Table 1
Assembly statistics of *Cocos nucifera* transcriptome.

Assembly	CAP3 assembler
Total Contig	285235
Contig >1000	111707
Contig >500	59979
Contig <499	113549
Largest Contig Size	16870
N50	1901
GC	42.15%

228293581 high quality cleaned reads with a phred score, Q ≥ 20 were obtained and these cleaned reads were used further for *de novo* transcriptome assembly.

A total of 395554 contigs were generated using *trinity* assembler with N50 value of 1717 basepair. CAP3 assembler was run on *trinity* assembled sequences to remove the redundant sequences. Finally CAP3 generated assembly had 285235 contigs with 42.15% of GC content and N50 value of 1901. In the assembly, the minimum and maximum read lengths were 201 basepair and 16870 basepair, respectively (Table 1). This was further used for transcriptome analysis. 61471 sequences had the read length between 200 basepair to 299 basepair, followed by 31408 and 20670 sequences that fall in between 300 basepair to 399 basepair and 400 basepair to 499 basepair, respectively. Total nine sequences had the read length greater than 15000 basepair.

3.2. Abundance estimation and differential genes expression analysis

Reads of resistant and susceptible varieties were mapped to *de novo* transcriptome assembly to calculate the expression values of each unigene in the form of FPKM (fragments mapped per kilo base of exon per million reads mapped) values (Fig. 1). Using edgeR tool, under stringent parameters, total 22021 differential expressed genes were identified with a FDR <0.05 and $\log_2$fold change value ± 2. Out of 22021 DEGs, 11067 and 10954 genes were upregulated and downregulated, respectively. Hierarchical clustering, heat map, MA plot and volcano plots were generated to represent the graphical representation of upregulated and downregulated genes (Fig. 2A and B).

3.3. Annotation and functional categorization

Homology search of *Cocos nucifera* differential expressed genes was performed using standalone blast package against the NCBI non-redundant (NR) database. Out of 22021 differential expressed genes, 18005 unigenes were found to show the match with other known genes in the database. Among 18005 unigenes, 14892 and 11220 unigenes were involved in mapping and annotation, respectively. 14240 unigenes were found to show maximum similarity with *Elaeis guineensis*, followed by 2838, 114, 57, 50 in *Phoenix dactylifera, Musa acuminate, Vitis vinifera* and *Cocos nucifera*, respectively. Detailed information of blast results are given under the supplements heading of the database link CnTDB (http://webtom.cabgrid.res.in/cntdb/).

After mapping, annotation and gene ontology, these unigenes were grouped into three categories i.e., biological process, molecular process and cellular component. Under biological process, "metabolic process" were shown in 7388 unigenes followed by "cellular activities" represented in 6376 unigenes. For the molecular functions, "catalytic activity" was followed by "binding activity" with 6252 unigenes and 4903 unigenes involved in these activities, respectively. Under the cellular component, "cell" and "organelle" were seen in 5891 unigenes and 4622 unigenes, respectively (Fig. 3).

Using Blast2GO Pro software, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis and found that total 133 pathways were involved in 6214 unigenes. "Biosynthesis of antibiotics pathway" was found in 96 unigenes which is maximum, followed by "starch and sucrose metabolism" and "purine metabolism" in 35 and 33 unigenes, respectively (Fig. 4).

Transcriptional factors (TF) play very important role to increase the expression of genes. For identification of TFs, Blastx algorithm was used against the PlantTFDB 3.0. A total of 7240 transcriptional factors were identified from differential expressed genes of *Cocos nucifera* when searched against the all species. Transcriptional factors like "Protein kinase superfamily", "b3 family protein", "TCP family protein", "abc-2 type transporter family" and "RLK (receptor-like protein kinase)" were found in maximum number in differential expressed genes. List of all transcriptional factors can be downloaded from CnTDB database link under the supplement heading.

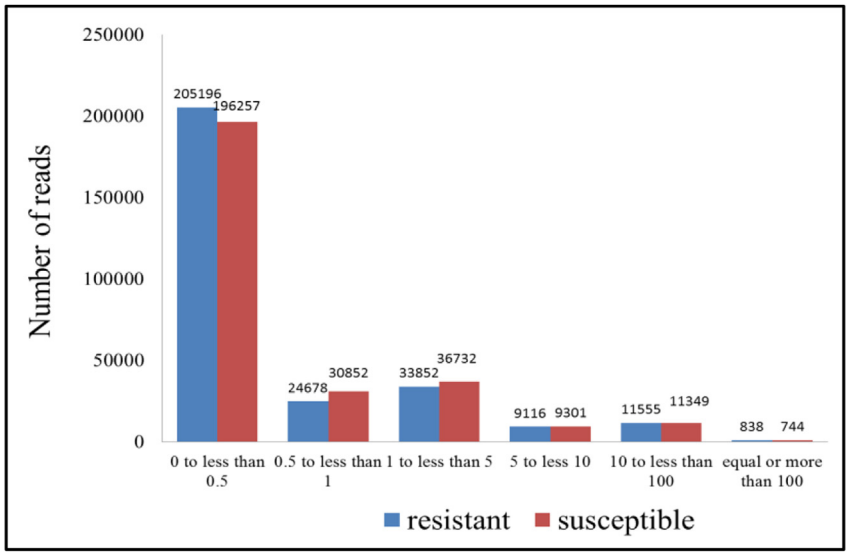

Fig. 1. Graphical view of FPKM (fragments mapped per kilo base of exon per million reads mapped) values in resistant and susceptible cultivars.

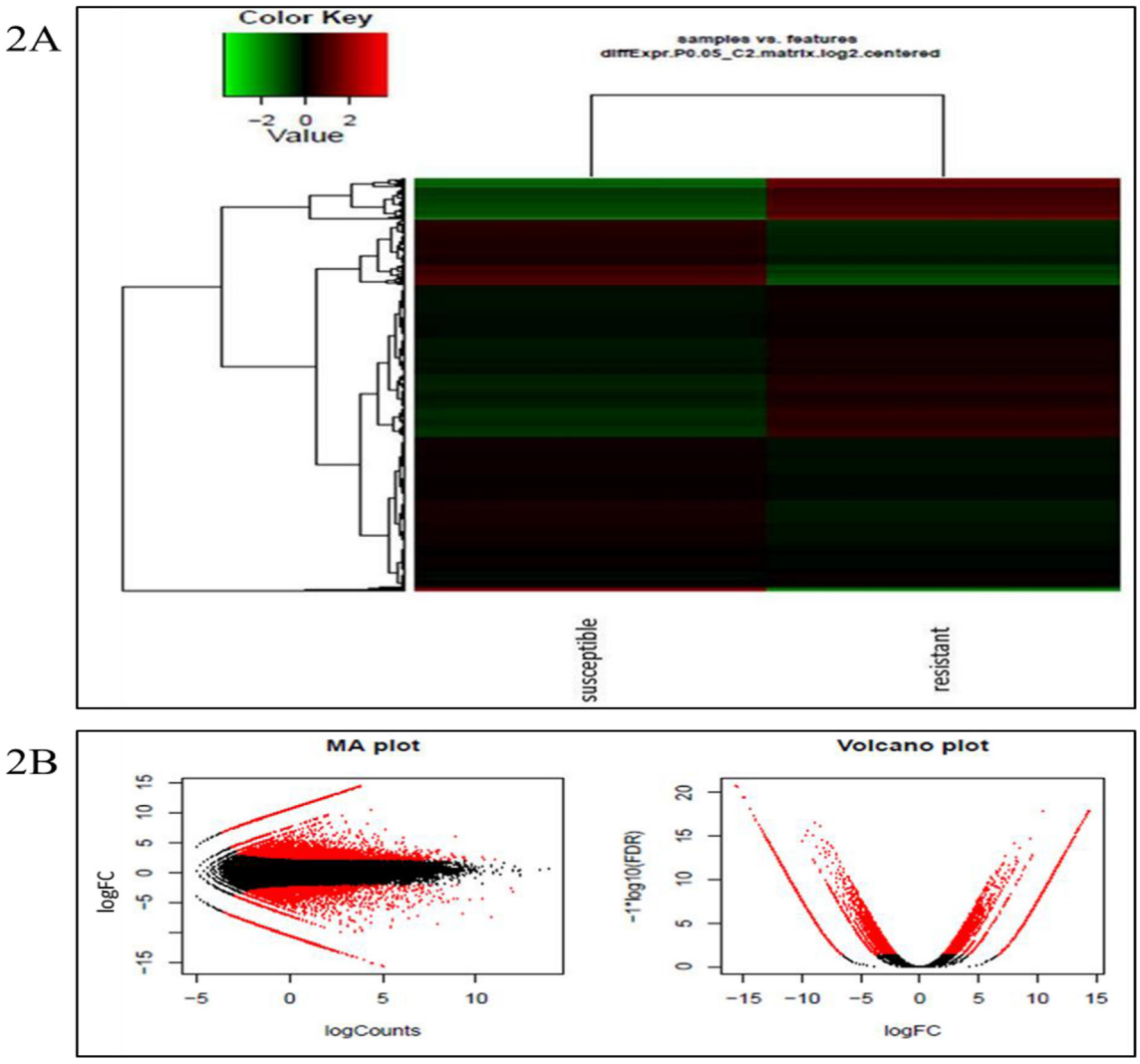

Fig. 2. Graphical representation of DEGs. (2A) Heat map of differential expressed genes (red colour shows upregulated and green colour represents downregulated genes. (2B) MA plot and volcano plot of DEGs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Also, homology searching of full *de novo* transcriptome assembly of *Cocos nucifera* against the NR database was performed. Out of 285235 unigenes, 154150 unigenes showed significant match with known sequences present in the database. It was observed that, maximum number of unigenes matched with *Elaeis guineensis*, i.e., in 113804 unigenes, followed by *Phoenix dactylifera* and *Vitis vinifera* in 26150 and 1341 unigenes, respectively.

3.4. Simple sequence repeats identification

In *Cocos nucifera*, vast numbers of simple sequence repeat (SSR) markers were found from DEGs as well as full *de novo* transcriptome assembly. A total of 10126 and 97117 SSR markers were mined from DEGs and *de novo* transcriptome assembly, respectively. Table 2 provides the information of mononucleotides, dinucleotides, trinucleotides, tetra nucleotides, penta nucleotides and hexa nucleotides markers in *de novo* assembly as well as DEGs. Since genic region SSR markers have several advantages like known gene function, more transferability, suitability in linkage mapping, functional diversity, thus they are preferred in molecular breeding program [26]. Such geneic SSR discovery and its potential use have been reported in several other crops like tomato and pepper [27], sugarcane [28], holy basil (tulsi) [29], sesame [30], African oil palm [31] and tea [32]. All simple sequence repeats and their primers can be downloaded from supplement heading of CnTDB link (http://webtom.cabgrid.res.in/cntdb/).

3.5. Detection of SNPs and indels

A total of 32370 and 39507 SNPs were found in resistant and susceptible varieties, respectively. Out of these, 22664 SNPs were similar in both resistant and susceptible. Also, 2709 and 3178 Indels were found resistant and susceptible, respectively, and 1067 Indel were seen common. Since large number of SNP and Indel were mined, to narrow down the count, quality value 20 and read depth of 4 as filtering criteria were applied to identify the signification SNPs and Indels (Fig. 5A and B). SNPs and Indels can be downloaded from CnTDB link under supplement heading section.

3.6. Database development

The coconut transcriptome database (CnTDB) catalogues 285235 assembled contigs from the two contrast genotypes considered under study. It houses 22021 DEGs and transcriptional

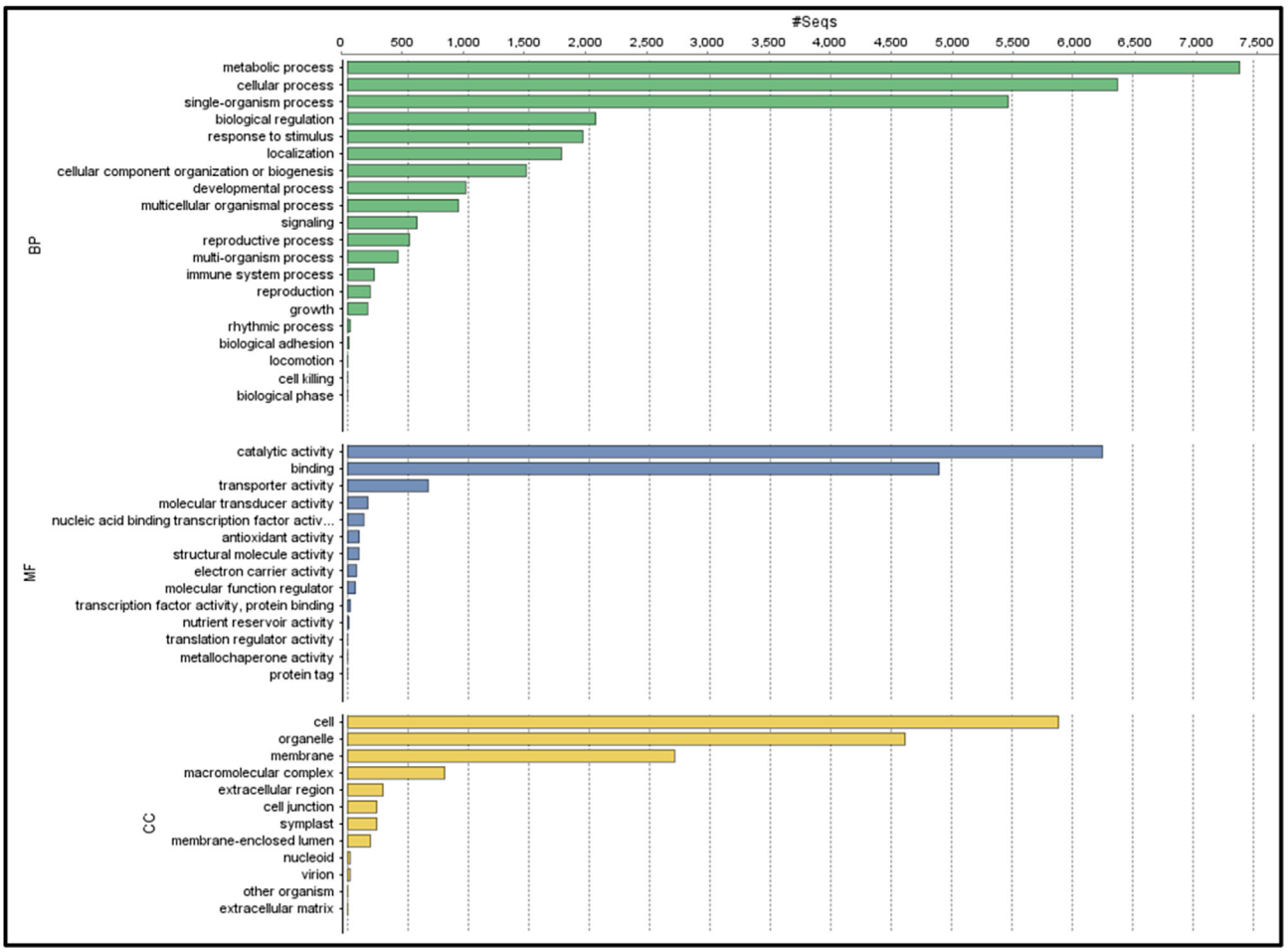

Fig. 3. Functional categorization of differential expressed genes.
(BP=Biological Process; MF = Molecular Function; CC= Cellular Components.

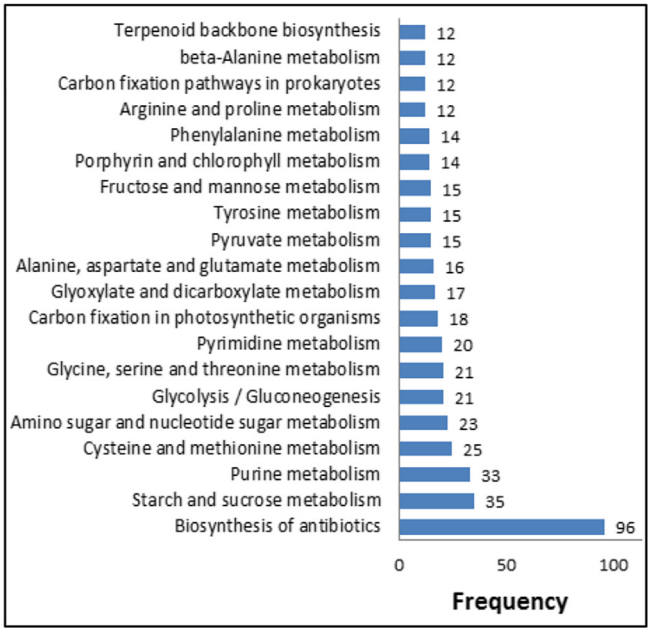

Fig. 4. Top 20 KEGG pathways identified in *Cocos nucifera*.

factors in its database. About 10126 and 97117 SSR markers mined from DEGs and *de novo* transcriptome assembly respectively are catalogued. Beside this, 71877 SNPs and 5887 indels from both, resistant and susceptible varieties are populated along with the description of gene ontologies and pathways of the assembled contigs.

The Coconut Transcriptome Database has five tabs *viz*., Home, Transcripts, Variants, DEG and supplements. "*Home*" page has the general information and utility of the database. Under the "*Transcript*" tab, user can search for expression profile, transcription factor families, domain and family and pathways along with the BLAST option. Under the "*Expression profile*" tab, information of expression values in the form of FPKM of resistant and susceptible varieties along with blast results are shown. The "*Transcription Factor families*" search tab provides the information of transcriptional factors. The tab "*Domain and family search*" shows the information of domains and families of transcripts, to which the hyperlinks are provided for direct link to EMBL-EBI Interpro database. Under the tab "*Pathways search*", information of differential expressed genes which are involved in pathways are provided with the enzyme name and enzyme ID. The maps are directly link to KEGG database.

The next tab, i.e., "*Variants*" has three features of search like SSR, SNPs and Indels. These tabs provide information of identified putative simple sequences repeats along with three sets of primers, single nucleotide polymorphism and indels, respectively. The tab

Table 2
SSR markers in *de novo* transcriptome assembly and differential expressed genes.

	de novo transcriptome assembly	DEGs
Sequences examined	285235	22021
Identified SSRs	97117	10126
SSR containing sequences	73884	7495
Sequences containing more than 1 SSR	17810	1975
SSRs present in compound formation	5903	581
Mono-nucleotide	61566	6063
Di- nucleotide	21713	2474
Tri- nucleotide	12646	1453
Tetra- nucleotide	1117	127
Penta- nucleotide	52	8
Hexa- nucleotide	23	1

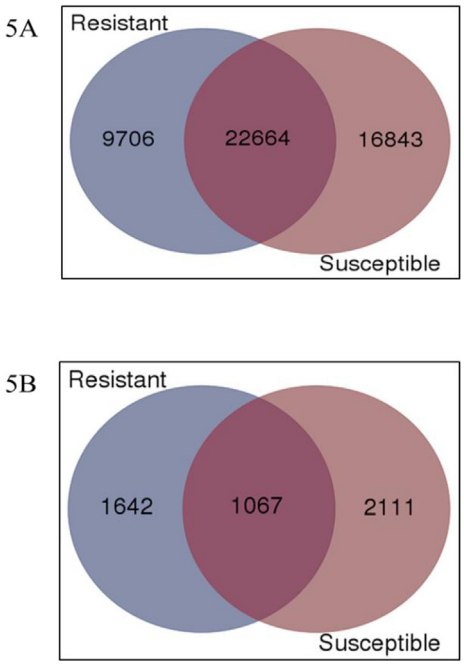

Fig. 5. Venn diagram representing common and unique SNPs and Indels. (**5A**) represents the common and unique SNPs. (**5B**) represents the common and unique Indels.

"*DEG*", provides all information of identified differential expressed genes of resistant and susceptible cultivars. Besides, it also shows the blast results of each transcript. Fig. 6 shows the schematic diagram of CnTDB for search.

4. Discussion

Transcriptome analysis of two extreme coconut genotypes with RWD resistance and susceptibility revealed presence of 22021 DEGs along with variant analysis to discovery of SSR, SNP and Indel that can be valuable genomic resource for coconut breeders.

Since the whole genome of *Cocos nucifera* (coconut) is yet to be sequenced, thus this study was done to develop genomic resource from differential expressed genes from contrasting genotypes of coconut palm associated with root wilt disease available transcriptome data. After assembly, it was found that around 54% unigenes showed similarity with known sequences against searched with non-redundant database, which implies that the unknown, uncharacterized and hypothetical genes may be the novel transcripts.

The identified up and down regulated differential expressed genes discovered in our study in RWD resistant and susceptible coconut cultivars may be a good genomic resource for candidate gene discovery against RWD. The result showed some important disease resistant genes such as NBS-LRR domain, PR1, PR4, pathogenesis-related genes transcriptional activator PTI5-like gene, thaumatin-like protein, HSP70 and glutathione S-transferase in our study, which supports the defence mechanism against RWD. In plants, most of the disease resistant genes encodes nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins and involved in activation of kinases and play important roles in signal transduction which provides the plant defense against pathogen attack. This domain is divided into TIR-domain-containing (TNL) and CC-domain-containing (CNL), both of which play role in recognition of pathogens [33]. We also found NBS-LRR domain highly upregulated in our study.

Previously, a study was done on Chinese jujube and identified the resistant genes which were associated with phytoplasma. In our results, few differentially expressed genes showed similarity with one of disease resistant TLP gene (thaumatin-like protein) which is highly downregulated [34]. HSP70, a chaperone molecule plays vital role in hypersensitive response in plant defence mechanism [35]. Some of the transcripts in our results showed similarity with HSP70, few of which were highly upregulated.

Interestingly, pathogenesis-related (PR) genes transcriptional activator PTI5-like gene was found to be highly upregulated in our study. PTI5 belongs to plant transcription factor ethylene-response factor (ERF) family and this gene transcriptional factor binds to promoters region of pathogenesis-related genes which involves in plant protection [36]. Pathogenesis-related genes plays role in innate immune response in plant defence against pathogens such as bacterial, fungal or virus infections. In present study, we found several pathogenesis-related genes with various isoforms such as PR-1, PR-4 with lower expression values. PR-1 gene was found to be upregulated in tomato when plant got infected with potato purple top phytoplasma [37], but our result show this gene to be downregulated.

Peroxisomal membrane proteins are very active in both, metabolism and morphological. In plants, it is known for their role in response to abiotic and biotic stresses, metabolism and development of crops [38]. In our study, we found different isoforms of peroxisomal membranes proteins such as PEX11, PEX13 and PEX14. The result also showed isoforms of Mitogen-activated protein kinase (MAPK) such as MAPK-3, MAPK-5 and MAPK-13a which supports their role in plant defence. MAPK is known to play an important role at the time of pathogen invader in plant and provide resistant to plant disease [39]. We also found glutathione S-transferase genes with different isoforms being highly upregulated. Down regulation of this gene in susceptible cultivar is known to be responsible for phytoplasma [40].

A total 237 enzymes were involved in 133 pathways in our investigation. Some components of signal transduction pathway

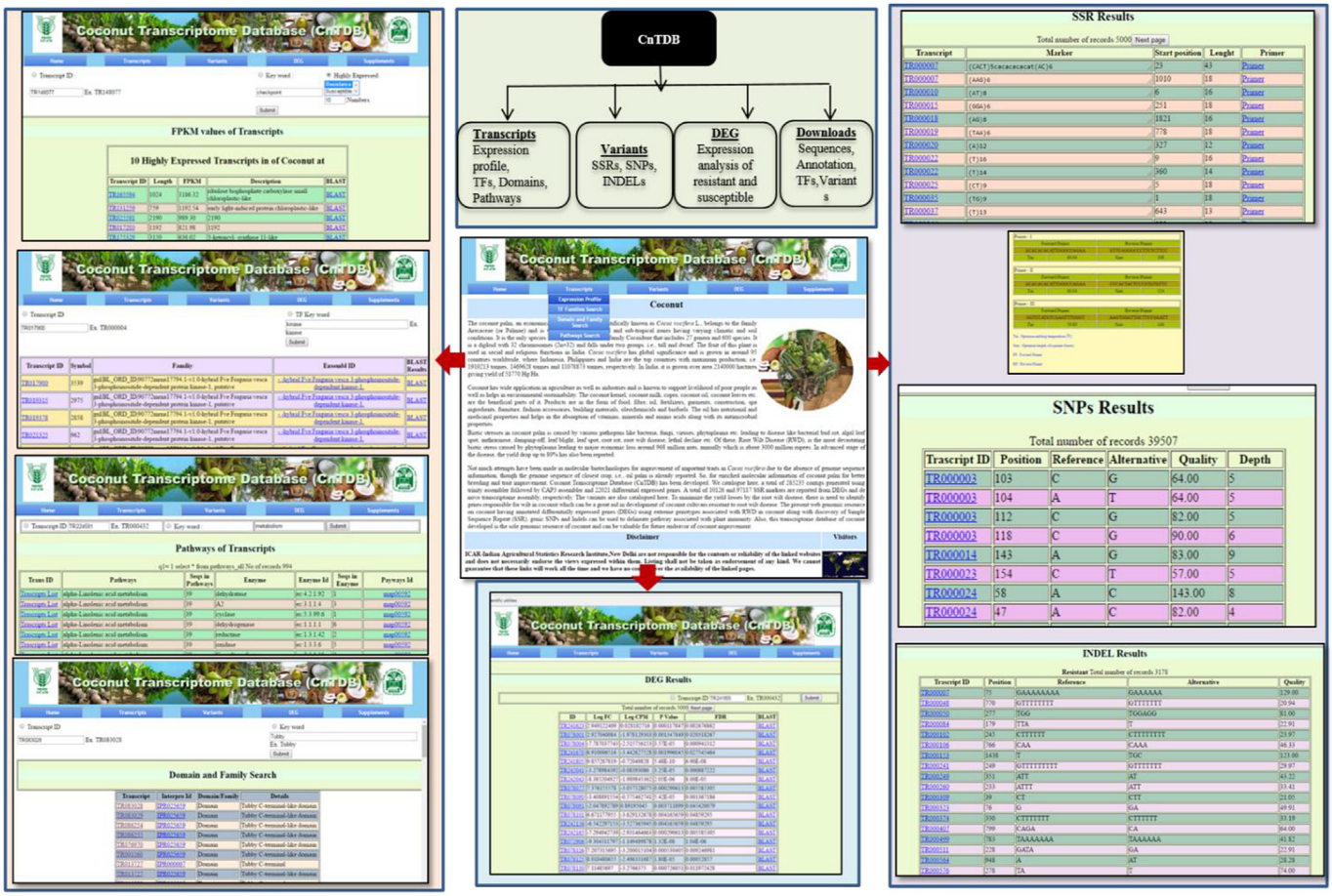

Fig. 6. Schematic diagram of CnTDB.

such as salicylic acid, jasmonic acid were also seen in our result. Phenolic molecule i.e. salicylic acids, is a plant hormone involved in the plant defence mechanism against various biotic stress and abiotic stress. Salicylic acid triggers the generation of reactive oxygen species (ROS) and other defence related mechanism. It activates the peroxidase (POD), which plays an important role in plant defence [41]. In our study we found peroxidases such as glutathione peroxidase and peroxiredoxin which were upregulated.

Phenylpropanoid biosynthesis is another important pathway involved in plant defence against various biotic and abiotic stresses. Phenylpropanoid pathway is responsible for producing other defence related pathways such as lignin biosynthesis and salicylic acid pathways [42]. In our study, we found that 149 genes belong to phenylpropanoid biosynthesis pathways strongly supporting the defence against pathogen.

Transcription factors plays cardinal role in regulation of gene expression. Since, transcription factors play important role in plant defence mechanism against biotic and abiotic stresses, we aimed at this analysis too. There are many reports where transcription factors has been found as candidate gene for important crop trait, for example; trait climbing ability in cucumber [43], NFYB1 transcription factor in crop abiotic stress tolerance [44]. Transcription factor families such as CCAAT, homeodomain, bHLH, NAC, AP2/ERF, bZIP, and WRKY are reported to be associated with drought tolerance, thus very valuable genomic resource in crop improvement program [45]. We found some very important transcriptional factors such as WRKY, basic leucine zipper (bZIP), MYB, TCP and RLK which are involved in plant protection in our study. WRKY plays crucial role in plant development and growth in response to various abiotic and biotic stresses [46]. We found large number of WRKY isoform in differential expressed genes analysis and most of the isoform having high fold change and most of them may be involved in defence signalling pathways. Basic leucine zipper (bZIP) plays an important role in the developmental and physiological processes such as biotic stress and abiotic stresses, morphogenesis. In plants, bZIP transcriptional factors are involved in the plant defence mechanism against the pathogen attacks and it is the key component of signal transduction pathways [47]. In our results, we found bZIP transcriptional factors with different type of isoforms and some are found highly expressed in resistant cultivars. These upregulated genes may be involved in plant defence mechanism. Myeloblastosis related proteins (MYB) transcriptional factors are critically involved in plant defence against pathogen invasion. It is lower in phytoplasma-infected plants than in non-infected plants [48]. In our study, we found that few isoform of MYB were highly upregulated. A pathogen recognition receptors (PRRs) family member, RLK (receptor-like protein kinase) is another transcriptional factor which is encoded by many plant genomes to protect from various biotic stresses such as bacterial, virus, fungal. It provides immunity to plant when it gets infected with foreign invaders [49]. TCP family protein too is an important transcriptional factor which was found in our study. It regulates the response of defence, growth, and development of plants via stimulating bioactive metabolites such as jasmonic acid (JA), brassinosteroid (BR) flavonoids [50].

Simple sequence repeat markers provide useful information

such as linkage mapping, genetic diversity and QTL. In our study, we identified SSR markers from genic regions and developed sets of primers which may be useful for further validation. A previous study has been done on *Cocos nucifera* to find out the genetic diversity of Chinese coconuts and germplasm of coconut in Southeast Asia [51]. Another study was done to identify the genetic variability of tall coconut among the populations using marker techniques [52]. Using five varieties of coconut, the genetic variation can also be identified [53]. In our study, we were found 97117 putative SSR markers which can be useful for studying genetic diversity and variations.

A total of 32370 and 39507 SNPs and 2709 and 3178 Indel from resistant and susceptible coconut cultivars respectively were found in our study. In resistant cultivar, we found maximum number of SNP and Indels in TR027830 (retrotransposon ty1-copia subclass), *i.e.*, 198, followed by 135 and 91 in TR027749 (transformation transcription domain-associated) and TR026985 (hypothetical protein) respectively. Where as in susceptible cultivar, we found 198, 135 and 113 SNP and indels in TR027830 (retrotransposon ty1-copia subclass), TR027749 (transformation transcription domain-associated) and TR026999 (transposon enspm sub-class) respectively. These *in silico* investigated transcription factors and putative marker has advantage in cost reduction and time of *in vitro* discovery but requires further validation. These putative markers may be useful in mapping population and validation for development of new improved cultivars as well as can be valuable genomic resource for future studies in *Cocos nucifera*.

5. Conclusion

This is the first report of the RWD transcriptome database of the two extreme coconut genotypes. Out of 22021 DEGs, 11067 and 10954 genes were upregulated and downregulated, respectively. A total of 10126 and 97117 SSR markers were mined from DEGs and *de novo* transcriptome assembly, respectively. We also report 32370 and 39507 SNPs in resistant and susceptible varieties, respectively and 2709 and 3178 Indels in resistant and susceptible respectively. The GO terms indicated “metabolic process” to be dominant under biological process, while “catalytic activity” and “cell” were maximum reported under molecular functions and cellular component, respectively. The important transcriptional factors found in our study are WRKY, basic leucine zipper (bZIP), MYB, TCP and RLK which are reported to be involved in plant protection. The present findings are the baseline information and is of immense use as genomic resource in endeavor of higher coconut productivity.

Conflict of interest

Authors declare that there is no competing interest.

Acknowledgments

We are thankful to Director, ICAR-IASRI, Indian Council of Agricultural Research, Ministry of Agriculture, Government of India, India for providing financial and infrastructural support to carry out this research. The grant of fellowship to SKV by Indian Agriculture Research Institute, New Delhi, is duly acknowledged.

References

[1] Y.-Y. Huang, C.-P. Lee, J.L. Fu, B.C.-H. Chang, A.J.M. Matzke, M. Matzke, De novo transcriptome sequence assembly from coconut leaves and seeds with a focus on factors involved in RNA-directed DNA methylation, G3 (Bethesda) 4 (2014) 2147–2157, http://dx.doi.org/10.1534/g3.114.013409.

[2] L. Perera, J.R. Russell, J. Provan, W. Powell, Studying genetic relationships among coconut varieties/populations using microsatellite markers, Euphytica 132 (2003) 121–128, http://dx.doi.org/10.1023/A:1024696303261.

[3] R. Ramjegathesh, G. Karthikeyan, L. Rajendran, I. Johnson, T. Raguchander, R. Samiyappan, Root (wilt) disease of coconut palms in South Asia – an overview, Arch. Phytopathol. Plant Prot. 45 (2012) 2485–2493, http://dx.doi.org/10.1080/03235408.2012.729772.

[4] 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 de novo assembly for subsequent functional genomics approaches, PLoS One 8 (2013) e59997, http://dx.doi.org/10.1371/journal.pone.0059997.

[5] K.A.G. Gopala, G. Raj, A. Singh Bhatnagar, P. Kumar, P. Chandrashekar, Coconut oil: chemistry, production and its applications - a review, Indian Coconut J. 53 (2010) 15–27.

[6] N. Nejat, D.M. Cahill, G. Vadamalai, M. Ziemann, J. Rookes, N. Naderali, Transcriptomics-based analysis using RNA-Seq of the coconut (Cocos nucifera) leaf in response to yellow decline phytoplasma infection, Mol. Genet. Genomics 290 (2015) 1899–1910, http://dx.doi.org/10.1007/s00438-015-1046-2.

[7] A.W. Meerow, R.R. Krueger, R. Singh, E.-T.L. Low, M. Ithnin, L.C.-L. Ooi, Coconut, date, and oil palm genomics, in: Genomics Tree Crop, Springer New York, New York, NY, 2012, pp. 299–351, http://dx.doi.org/10.1007/978-1-4614-0920-5_10.

[8] D. Chakrabarty, P.S. Chauhan, A.S. Chauhan, Y. Indoliya, U.C. Lavania, C.S. Nautiyal, *De novo* assembly and characterization of root transcriptome in two distinct morphotypes of vetiver, *Chrysopogon zizaniodes (L.) Roberty*, Sci. Rep. 5 (2015) 18630, http://dx.doi.org/10.1038/srep18630.

[9] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, http://dx.doi.org/10.1093/bioinformatics/btu170.

[10] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, http://dx.doi.org/10.1038/nbt.1883.

[11] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512, http://dx.doi.org/10.1038/nprot.2013.084.

[12] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877, http://dx.doi.org/10.1101/GR.9.9.868.

[13] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25, http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[14] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinforma. 12 (2011) 323, http://dx.doi.org/10.1186/1471-2105-12-323.

[15] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140, http://dx.doi.org/10.1093/bioinformatics/btp616.

[16] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, BMC Bioinforma. 10 (2009) 421, http://dx.doi.org/10.1186/1471-2105-10-421.

[17] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676, http://dx.doi.org/10.1093/bioinformatics/bti610.

[18] J. Jin, F. Tian, D.-C. Yang, Y.-Q. Meng, L. Kong, J. Luo, G. Gao, PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants, gkw982, Nucleic Acids Res. (2016), http://dx.doi.org/10.1093/nar/gkw982.

[19] T. Thiel, W. Michalek, R.K. Varshney, A. Graner, Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.), Theor. Appl. Genet. 106 (2003) 411–422, http://dx.doi.org/10.1007/s00122-002-1031-0.

[20] M.-Y. Li, F. Wang, Q. Jiang, J. Ma, A.-S. Xiong, Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens L.*) by deep transcriptome sequencing, Hortic. Res. 1 (2014) 10, http://dx.doi.org/10.1038/hortres.2014.10.

[21] N. Fu, Q. Wang, H.-L. Shen, De Novo, Assembly, gene annotation and marker development using Illumina paired-end transcriptome sequences in celery (Apium graveolens L.), PLoS One 8 (2) (2013) e57686, http://dx.doi.org/10.1371/journal.pone.0057686.

[22] E.J. Oliveira, J.G. Pádua, M.I. Zucchi, R. Vencovsky, M.L.C. Vieira, Origin, evolution and genome distribution of microsatellites, Genet. Mol. Biol. 29 (2) (2006).

[23] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm, S.G. Rozen, Primer3—new capabilities and interfaces, Nucleic Acids Res. 40 (2012) e115, http://dx.doi.org/10.1093/nar/gks596.

[24] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760, http://dx.doi.org/10.1093/bioinformatics/btp324.

[25] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the

sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, http://dx.doi.org/10.1093/bioinformatics/btp352.
[26] R.K. Varshney, A. Graner, M.E. Sorrells, Genic microsatellite markers in plants: features and applications, Trends Biotechnol. 23 (1) (2005) 48–55.
[27] J.K. Yu, H. Paik, J.P. Choi, J.-H. Han, J.K. Choe, C.-G. Hur, Functional domain marker (FDM): an in silico demonstration in solanaceae using simple sequence repeats (SSRs), Plant Mol. Biol. Rep. 28 (2) (2009) 352–356.
[28] S.K. Parida, A. Pandit, K. Gaikwad, T.R. Sharma, P.S. Srivastava, N.K. Singh, T. Mohapatra, Functionally relevant microsatellites in sugarcane unigenes, BMC Plant Biol. 10 (2010) 251, http://dx.doi.org/10.1186/1471-2229-10-251.
[29] S. Gupta, R. Shukla, S. Roy, N. Sen, A. Sharma, In silico SSR and FDM analysis through EST sequences in, Ocimum Basilicum. POJ 3 (4) (2010) 121–128.
[30] U. Bhattacharyya, S.K. Pandey, T. Dasgupta, Identification of EST-SSRs and FDM in sesame (*Sesamum indicum L.*) through data mining, Sch. J. Agric. Sci. 4 (2) (2014) 60–69.
[31] T.J. Tranbarger, W. Kluabmongkol, D. Sangsrakru, F. Morcillo, W.J. Tregear, S. Tragoonrung, N. Billotte, SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*, BMC Plant Biol. 12 (2012) 1, http://dx.doi.org/10.1186/1471-2229-12-1.
[32] J. Sahu, R. Sarmah, B. Dehury, K. Sarma, S. Sahoo, M. Sahu, M. Barooah, M.K. Modi, P. Sen, Mining for SSRs and FDMs from expressed sequence tags of *Camellia sinensis*, Bioinformation 8 (6) (2012) 260–266.
[33] M.K. Rajesh, K.E. Rachana, S.A. Naganeeswaran, R. Shafeeq, R.J. Thomas, M. Shareefa, B. Merin, K. Anitha, Identification of expressed resistance gene analog sequences in coconut leaf transcriptome and their evolutionary analysis, TURKISH J. Agric. 39 (2015) 489–502.
[34] Z. Liu, Y. Wang, J. Xiao, J. Zhao, M. Liu, Identification of genes associated with phytoplasma resistance through suppressive subtraction hybridization in Chinese jujube, Physiol. Mol. Plant Pathol. 86 (2014) 43–48, http://dx.doi.org/10.1016/j.pmpp.2014.03.001.
[35] H. Kanzaki, H. Saitoh, A. Ito, S. Fujisawa, S. Kamoun, S. Katou, H. Yoshioka, R. Terauchi, Cytosolic HSP90 and HSP70 are essential components of INF1-mediated hypersensitive response and non-host resistance to Pseudomonas cichorii in Nicotiana benthamiana, Mol. Plant Pathol. 4 (2003) 383–391, http://dx.doi.org/10.1046/j.1364-3703.2003.00186.x.
[36] Y.-Q. Gu, M.C. Wildermuth, S. Chakravarthy, Y.-T. Loh, C. Yang, X. He, Y. Han, G.B. Martin, Tomato transcription factors pti4, pti5, and pti6 activate defense responses when expressed in arabidopsis, Plant Cell 14 (2002) 817–831.
[37] W. Wu, Y. Ding, W. Wei, R.E. Davis, I.-M. Lee, R.W. Hammond, Y. Zhao, Salicylic acid-mediated elicitation of tomato defence against infection by potato purple top phytoplasma, Ann. Appl. Biol. 161 (2012) 36–45, http://dx.doi.org/10.1111/j.1744-7348.2012.00550.x.
[38] Y. Nyathi, A. Baker, Plant peroxisomes as a source of signalling molecules, Biochim. Biophys. Acta 1763 (2006) 1478–1495, http://dx.doi.org/10.1016/j.bbamcr.2006.08.031.
[39] X. Meng, S. Zhang, MAPK cascades in plant disease resistance signaling, Annu. Rev. Phytopathol. 51 (2013) 245–266, http://dx.doi.org/10.1146/annurev-phyto-082712-102314.
[40] R. Liu, Y. Dong, G. Fan, Z. Zhao, M. Deng, X. Cao, S. Niu, Discovery of genes related to witches broom disease in Paulownia tomentosa × Paulownia fortunei by a De Novo assembled transcriptome, PLoS One 8 (2013) e80238, http://dx.doi.org/10.1371/journal.pone.0080238.
[41] A.R. War, M.G. Paulraj, M.Y. War, S. Ignacimuthu, Role of salicylic acid in induction of plant defense system in chickpea (Cicer arietinum L.), Plant Signal. Behav. 6 (2011) 1787–1792, http://dx.doi.org/10.4161/psb.6.11.17685.
[42] F.G. Malinovsky, J.U. Fangel, W.G.T. Willats, The role of the cell wall in plant immunity, Front. Plant Sci. 5 (2014) 178, http://dx.doi.org/10.3389/fpls.2014.00178.
[43] S. Wang, X. Yang, M. Xu, X. Lin, T. Lin, J. Qi, G. Shao, N. Tian, Q. Yang, Z. Zhang, S. Huang, A rare SNP identified a TCP transcription factor essential for tendril development in cucumber, Mol. Plant 8 (2015) 1795–1808.
[44] S. Takeda, M. Matsuoka, Genetic approaches to crop improvement: responding to environmental and population changes, Nat. Rev. Genet. 9 (2008) 444–457, http://dx.doi.org/10.1038/nrg2342.
[45] R.C. Rabara, P. Tripathi, P.J. Rushton, The potential of transcription factor-based genetic engineering in improving crop tolerance to drought, OMICS 18 (10) (2014) 601–614.
[46] M. Wang, A. Vannozzi, G. Wang, Y.-H. Liang, G.B. Tornielli, S. Zenoni, E. Cavallini, M. Pezzotti, Z.-M.M. Cheng, Genome and transcriptome analysis of the grapevine (Vitis vinifera L.) WRKY gene family, Hortic. Res. 1 (2014) 14016, http://dx.doi.org/10.1038/hortres.2014.16.
[47] M.S. Alves, S.P. Dadalto, A.B. Gonçalves, G.B. De Souza, V.A. Barros, L.G. Fietto, Plant bZIP transcription factors responsive to pathogens: a review, Int. J. Mol. Sci. 14 (2013) 7815–7828, http://dx.doi.org/10.3390/ijms14047815.
[48] F. Ehya, A. Monavarfeshani, E. Mohseni Fard, L. Karimi Farsad, M. Khayam Nekouei, M. Mardi, G.H. Salekdeh, Phytoplasma-Responsive microRNAs modulate hormonal, nutritional, and stress signalling pathways in mexican lime trees, PLoS One 8 (2013) e66372, http://dx.doi.org/10.1371/journal.pone.0066372.
[49] N. Lannoo, E.J.M. Van Damme, Lectin domains at the frontiers of plant defense, Front. Plant Sci. 5 (2014) 397, http://dx.doi.org/10.3389/fpls.2014.00397.
[50] S. Li, The Arabidopsis thaliana TCP transcription factors: a broadening horizon beyond development, Plant Signal. Behav. 10 (2015) e1044192, http://dx.doi.org/10.1080/15592324.2015.1044192.
[51] Y. Xiao, Y. Luo, Y. Yang, H. Fan, W. Xia, A.S. Mason, S. Zhao, R. Sager, F. Qiao, Development of microsatellite markers in Cocos nucifera and their application in evaluating the level of genetic diversity of Cocos nucifera, POJ 6 (2013) 193–200.
[52] F.E. Ribeiro, L. Baudouin, P. Lebrun, L.J. Chaves, C. Brondani, M.I. Zucchi, R. Vencovsky, Population structures of Brazilian tall coconut (Cocos nucifera L.) by microsatellite markers, Genet. Mol. Biol. 33 (2010) 696–702, http://dx.doi.org/10.1590/S1415-47572010005000077.
[53] D.V. Rasam, N.B. Gokhale, S.V. Sawardekar, D.M. Patil, Molecular characterisation of coconut (Cocos nucifera L.) varieties using ISSR and SSR markers, J. Hortic. Sci. Biotechnol. 91 (2016) 347–352, http://dx.doi.org/10.1080/14620316.2016.1160544.