


Article

High Quality Unigenes and Microsatellite Markers from Tissue Specific Transcriptome and Development of a Database in Clusterbean (*Cyamopsis tetragonoloba* (L.) Taub.)

Hukam C. Rawal¹, Shrawan Kumar¹, Amitha Mithra S.V.¹, Amolkumar U. Solanke¹, Deepti Nigam¹ , Swati Saxena¹, Anshika Tyagi¹, Sureshkumar V.¹, Pritam Kalia², Narendra Pratap Singh³, Neelam R. Yadav⁴, Nagendra Kumar Singh¹, Tilak Raj Sharma^{1,†} and Kishor Gaikwad^{1,*}

¹ ICAR-National Research Centre on Plant Biotechnology, New Delhi 110012, India; hukam.rawal@gmail.com (H.C.R.); kumarshrawan12@gmail.com (S.K.); amithamithra.nrcpb@gmail.com (A.M.S.V.); amolsgene@gmail.com (A.U.S.); deep_t_mbi@yahoo.co.in (D.N.); swatisaxena605@gmail.com (S.S.); tyagi.anshika9@gmail.com (A.T.); sureshkumarv1996@gmail.com (S.V.); nksingh@nrcpb.org (N.K.S.); trsharma1965@gmail.com (T.R.S.)

² ICAR-Indian Agricultural Research Institute, New Delhi 110012, India; pritam.kalia@gmail.com

³ ICAR-Indian Institute of Pulse Research, Kanpur 208204, India; npsingh.iipr@gmail.com

⁴ Department of Biotechnology and Molecular Biology, CCS Haryana Agricultural University, Hisar 125004, India; nryadav@hau.ernet.in

* Correspondence: kish2012@nrcpb.org; Tel.: +91-011-25841789

† Current address: National Agri-Food Biotechnology Institute, Mohali 140306, India.

Received: 23 August 2017; Accepted: 6 November 2017; Published: 9 November 2017

Abstract: Clusterbean (*Cyamopsis tetragonoloba* L. Taub), is an important industrial, vegetable and forage crop. This crop owes its commercial importance to the presence of guar gum (galactomannans) in its endosperm which is used as a lubricant in a range of industries. Despite its relevance to agriculture and industry, genomic resources available in this crop are limited. Therefore, the present study was undertaken to generate RNA-Seq based transcriptome from leaf, shoot, and flower tissues. A total of 145 million high quality Illumina reads were assembled using Trinity into 127,706 transcripts and 48,007 non-redundant high quality (HQ) unigenes. We annotated 79% unigenes against Plant Genes from the National Center for Biotechnology Information (NCBI), Swiss-Prot, Pfam, gene ontology (GO) and KEGG databases. Among the annotated unigenes, 30,020 were assigned with 116,964 GO terms, 9984 with EC and 6111 with 137 KEGG pathways. At different fragments per kilobase of transcript per millions fragments sequenced (FPKM) levels, genes were found expressed higher in flower tissue followed by shoot and leaf. Additionally, we identified 8687 potential simple sequence repeats (SSRs) with an average frequency of one SSR per 8.75 kb. A total of 28 amplified SSRs in 21 clusterbean genotypes resulted in polymorphism in 13 markers with average polymorphic information content (PIC) of 0.21. We also constructed a database named 'ClustergeneDB' for easy retrieval of unigenes and the microsatellite markers. The tissue specific genes identified and the molecular marker resources developed in this study is expected to aid in genetic improvement of clusterbean for its end use.

Keywords: *Cyamopsis tetragonoloba*; clusterbean; transcriptome; RNA-Seq; tissue-specific; polymorphism; microsatellite markers; database

1. Introduction

Cyamopsis tetragonoloba L. Taub (Clusterbean, $2n = 14$) commonly known as guar, is a leguminous *kharif* crop grown for forage, industrial and vegetable purpose [1]. Clusterbean is a drought-tolerant legume which can survive in marginal lands including saline and low fertile soils [2,3]. In recent years, it has emerged as an industrial crop due to its gum content which serves as a raw material in a wide range of industrial applications from mining for oil and gas to food, cosmetic and textile industries [4–6]. Clusterbean also has medicinal value and is used in the treatment of high cholesterol and diabetics in indigenous medicine preparations [7]. The great advantage of guar gum, chemically galactomannans, compared to those of other species, such as carob (*Ceratonia siliqua* L.), is that it is extremely viscous even at low concentrations and is highly soluble in cold water [8]. Moreover, clusterbean meal, obtained as a byproduct after the extraction of gum from the endosperm is an excellent feed supplement for the livestock, broiler and fish, due to its high protein content (32–52%) [9]. India is the largest producer of clusterbean comprising of 80% of the world production [10]. India is also the major exporter of guar gum, especially to the USA, China, Germany, Russia, and Canada [9].

The major grain pulses of the sub-continent, chickpea and pigeonpea have huge genomic resources developed in the last decade [11–14]. However, clusterbean has limited genomic resources despite its huge economic importance. Due to lack of genomic resources, presently, conventional breeding is the only means of clusterbean improvement. In this regard, availability of genomic resources can serve as a good platform for clusterbean improvement. As of now, there are 78,686 sequences (62,146 unigenes, 16,476 expressed sequence tags (EST)s and 61 nucleotides sequences) available in the National Center for Biotechnology Information (NCBI) database generated from early and late developing clusterbean seeds, and leaf transcriptomes generated from two clusterbean varieties using Illumina sequencing platform [15,16]. Recently, we have published the chloroplast genome of clusterbean [17].

For precise breeding applications, availability of DNA markers is a prerequisite. Few large-scale molecular marker development efforts are reported in this crop. So far, only anonymous molecular marker systems such as inter simple sequence repeats (ISSR), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and ribosomal DNA (rDNA) have been deployed in clusterbean [18–22]. Development of locus specific codominant DNA markers would be more appropriate for constructing high density genetic maps, marker assisted selection (MAS), evolutionary and population genetics studies of this species. Thirty-nine polymorphic simple sequence repeats (SSRs) developed from 16,476 EST sequences are the only locus specific markers available in this legume crop [23].

For large-scale discovery and characterization of functional genes and genome assembly, global exploration of the transcriptome is a useful strategy. Sequencing of RNA remains the gold standard for annotation of both coding and non-coding genes [24,25]. RNA-Seq method offers a holistic view of the transcriptome, revealing many novel transcribed regions, splice isoforms, genic microsatellites and the precise location of transcription boundaries [26–29]. In the present study, transcriptome analysis of three different tissues namely, leaf, shoot, and flower from clusterbean are reported which will serve as a valuable resource for whole genome assembly of clusterbean, identification of tissue specific genes and promoters, and development of molecular markers.

2. Materials and Methods

2.1. Plant Material and RNA Isolation

RGC 936, a popular, short-duration, early maturing clusterbean variety was used for transcriptome analysis. It is a branched variety bearing white flowers and round and pink colored seeds containing 33% gum. Plants were grown in 10'' pots at the Net house of ICAR-NRCPB (Indian council of Agricultural Research-National Research Centre on Plant Biotechnology), New Delhi under natural conditions in the crop season. The plants were irrigated every alternate day with normal tap water. Leaf, shoot and flower tissues were harvested at the first flowering stage, i.e., 42 days after sowing,

frozen in liquid nitrogen and kept at -80°C till further use. Total RNA from three biological replicates for each tissue was isolated from the harvested samples using Spectrum Plant Total RNA Kit (Sigma, St. Louis, MO, USA) following the manufacturer's protocol.

2.2. Library Preparation and RNA Sequencing

Total RNA was quantified using Nanodrop spectrophotometer (Thermo Scientific, Waltham, MA, USA) and the quality assessment was performed with RNA 6000 Nano assay kit using Bioanalyser 2100 (Agilent, Santa Clara, CA, USA). RNA from a single tissue from all the three biological replicates type was pooled in equi-molar concentrations. The RNA sequencing library of leaf, shoot and flower tissues were separately constructed using Truseq RNA Sample prep kit (Illumina, Singapore) following the manufacturer's protocol. The average size of library was 260 bp. The libraries were then sequenced by Illumina paired end sequencing technology. The raw sequence data for three tissues has been deposited at NCBI Short Read Archive (SRA) with accession number SRR5428802, SRR5428803 and SRR5428804.

2.3. RNA-Seq Data Processing and De Novo Assembly

RNA-Seq raw reads were first processed for trimming (with Phred Score 33) using Trimmomatic-0.36 to remove low quality sequences and reads shorter than 36 bp [30]. The resultant high quality trimmed reads were de novo assembled with three different assemblers including Trinity2.2.0 [31], CLC Genomics workbench 7.0 (CLC Bio, Aarhus, Denmark) and SPAdes 3.9.0 [32] at different k -mers. The different assemblies, so obtained, were compared for different aspects including assembled transcriptome size, transcripts number, average length and N50 length of assembled transcripts. The assembly with Trinity2.2.0 at k -mer 25 with normalization 30 and the minimum k -mer coverage of 2 was found to be the best and used in downstream analysis (Supplementary Table S1).

Further, transcripts were cleaned using perl script SeqClean (<https://sourceforge.net/projects/seqclean/>) to remove contaminations like rRNA, low-complexity RNA, and polyA stretches. Cleaned transcripts were clustered by CD-HIT V4.6 to remove redundancies and unigenes were obtained with sequence identity and similarity cut-off set at 97% and 95%, respectively [33]. A commonly accepted estimate for the expression level of unigenes, FPKM (Fragments per kilobase of transcript per millions fragments sequenced) was calculated by running RSEM (RNA-Seq by Expectation Maximization) module from Trinity package [34]. Unigenes with FPKM < 1 and length ≤ 200 bp were removed to avoid any potential assembly errors and to ensure the quality of the resulted assembly. The high quality (HQ) unigenes (with FPKM ≥ 1 and length > 200 bp) were used in the downstream analysis.

2.4. Functional Annotation of High Quality Unigenes

HQ unigenes were BLAST-searched against different databases including Annotated Plant Genes (APG) database from NCBI (22,77,559 genes), Swiss-Prot (4,64,207 genes), Pfam, gene ontology (GO), KEGG and Enzyme Commission (EC) numbers using BLASTx program with a cut-off E -value of 1×10^{-10} [35]. Functional descriptions were assigned to HQ unigenes with BLAST results against APG database using Blast2GO 4.0.2 [36]. Blast2GO was also used to perform InterproScan (IPS) Search, assign Gene ontology terms (GO) and carry out pathway analyses (using KEGG).

2.5. Differential Gene Expression Analysis

To perform the tissue-specific transcriptome analysis, high-quality trimmed reads from each of the three samples were mapped against the assembled transcripts using RSEM package version 1.2.31 and the abundance estimates were obtained [34]. Transcripts with FPKM value ≥ 1 in all 3 tissue samples were considered as housekeeping genes [37]. In order to identify tissue-specific genes, the expression results were further parsed based on FPKM values with X -fold higher for one tissue as compared to the FPKM values for the remaining two tissue samples for a specific gene, where $X = 5, 8, 10$ and 50.

To estimate the gene expression level, we categorized all assembled genes into seven different categories, at a threshold of 5-fold FPKM. “Un-expressed” are those with <1 FPKM in all 3 tissues, while “tissue specific genes” are those with at least 5-fold higher FPKM in one tissue as compared to the other two tissues. The contigs with >5 FPKM value in all the three tissues were categorized as “Expressed in all” while those in only two tissues were grouped as “Mixed expressed”. Among the contigs which had <5 FPKM, those detected in only one tissue were “Low expressed but tissue specific” while the ones detected in two tissues were, “Mixed but low expressed” and those detected in all 3 tissues but at least one tissue <5 FPKM value were “Expressed in all low”.

Using the Bioconductor package edgeR (Extraction of Differential Gene Expression R Package 3.3.1) with dispersion value as 0.1 and other parameters using default settings, differentially expressed genes (DEGs) were extracted [38]. We used a *p*-value cut-off of ≤ 0.05 with at least two-fold change (*C* value ≥ 1) to identify significant DEGs. For graphical illustration of the expression profiles of the identified DEGs, heatmaps and plots were generated. A clustered heatmap was generated for pairwise comparison between three tissue samples representing the Pearson correlation matrix.

2.6. Simple Sequence Repeat Mining

The Perl script MISA [39] (<http://pgrc.ipkgatersleben.de/misa/>) was used to identify simple repeat sequences (SSRs) in the assembled transcripts as well as in the HQ unigenes. The minimum number of nucleotide repeats searched during the SSR analysis was set as six for di-nucleotide repeats and five for tri-, tetra-, penta- and hexa-nucleotide repeats with maximal number of bases interrupting 2 SSRs in a compound microsatellite as 100.

2.7. Marker Validation and Diversity Analysis

Validation of the SSR markers was carried out using 21 clusterbean accessions, obtained from Chaudhary Charan Singh Haryana Agricultural University, Hisar, India. Genomic DNA was extracted from young leaves of these accessions using cetyltrimethylammonium bromide (CTAB) extraction method [40]. DNA quality was evaluated by 0.8% agarose gel electrophoresis. The working concentration of DNA was adjusted to 50 ng/ μ L, for use in marker genotyping. Amplification was performed in 20 μ L volume reactions containing, PCR buffer (10 mM Tris pH 9.0, 50 mM KCl), 1.5 mM MgCl₂, 0.6 U Taq DNA polymerase (Bangalore Genei, Bangalore, India), 2 μ M of dNTP, 10 pM of primer, and 50 ng of genomic DNA. Microsatellite loci were amplified on a Thermal Cycler (Applied Biosystems Veriti, Foster City, CA, USA). PCR amplification was carried out with the following cycling conditions: one cycle of 4 min at 94 °C followed by 35 cycles at 94 °C for 30 s, 55–60 °C for 30 s and 72 °C for 30 s. The final extension was performed at 72 °C for 10 min. After completion of the amplification, 2.5 mL 6 \times blue dye was added to the samples, and the amplified DNA was analyzed on 3.5% MetaPhoreTM agarose (Lonza Rockland Inc., Rockland, ME, USA) gels in 1 \times TBE buffer at 120 V for 4–5 h. A 50 bp DNA ladder was also resolved to determine the approximate size of the fragments. The gel was documented in the gel documentation unit (Syngene, Cambridge, UK). Since we identified only a maximum of two alleles in polymorphic markers, Numerical Taxonomy System (NTSYS-pc) ver. 2.1 (Exeter Software, Setauket, NY, USA) was used for construction of similarity matrix using the SSR genotyping data scored as presence and absence. From the binary data matrix, Jaccard’s similarity coefficient between pairs of accessions was calculated in the SIMQUAL module. Further, using the un-weighted pair grouped method arithmetic average (UPGMA), a dendrogram depicting diversity and genetic relationship among the accessions was constructed.

2.8. Database Design

ClustergeneDB, a database for retrieving information on the unigenes of clusterbean was constructed by using XAMPP (Apache, MariaDB, PHP and Perl) server. Backend of the database was designed using MySQL while the front-end was designed using HTML5 and CSS3. JQuery and Javascript were used to create the user framework. PHP5 was used to connect users and server to access

queries. The database is hosted in the server environment, FUJITSU PrimeRGY-Rx600S6 and Windows operating system. Microsatellites present in the unigenes and their expression profiles were also added in the relational database. The database can be accessed at <http://14.139.229.201/clustergenedb>.

3. Results

3.1. Transcriptome Sequencing and De Novo Assembly

To comprehensively construct the complete transcriptome of *C. tetragonoloba*, RNA from three tissues representing different development stages, including leaves, shoots and flowers were sequenced. In total, 149 million paired-end raw reads with an average read length of 100 bp were generated from three tissue samples (Table 1). After quality check, trimming of adapter sequences and size selection, 145 million HQ reads (97.3%) remained for assembly. From 22.16 Gb of trimmed reads (109 million paired reads and 36 million un-paired reads), a total of 127,706 transcripts were reconstructed into 179 Mbp with N50 of 2263 bp and the largest transcript length of 16.94 kb (Tables 1 and 2). To reduce redundancy and potential assembly errors, we clustered assembled transcripts into 110,485 unigenes and removed transcripts with FPKM values <1 and sequence length < 200 bp, since these are more likely to be error prone. As a result, a final dataset of 48,007 HQ unigenes with an average length of 1583.43 bp and an N50 of 2179 bp was obtained (Table 2). The assembled size of these HQ unigenes accounted for 76.01 Mb. The average GC content was found to be less than 40%. The size and GC% distribution for these HQ unigenes is shown in Supplementary Figures S1 and S2, respectively. Their size distribution depicted that 63.07% of these unigenes were >1 kb long while 19.26% were 500–1000 bp long. Only 17.63% of total HQ unigenes were shorter than 500 bp. A high proportion of these unigenes (84.77%) had GC content in the range of 35 to 45%.

Table 1. Summary of the trimming results with Trimmomatic for each cDNA library sequenced.

Library/Sample	Number of Raw Reads (Paired)	Number of HQ Reads (Paired)	Number of HQ Reads (Un-Paired)	HQ Reads (Bases)	Average Length (HQ Paired Reads)
Flower	16,325,263	11,896,062	3,967,411	2,408,942,984	88.54
Leaf	41,575,280	30,676,700	9,701,552	6,192,249,152	88.91
Shoot	92,030,452	66,439,261	22,849,685	13,560,175,894	88.90
Total	149,930,995	109,012,023	36,518,648	22,161,368,030	

HQ: high quality.

Table 2. Transcriptome assembly and functional annotation of *Cyamopsis tetragonoloba*.

Assembly Statistics	Data
Transcripts	
Total Assembled	127,706 (179.50 Mb)
Average Length	1405.63 bp
GC%	39.22
≥1000 bp	64,606 (150.06 Mb)
≥5000 bp	2218 (138.75 Mb)
≥10,000 bp	53 (628.55 kb)
Largest Transcripts	16,940 bp
N50 Length	2263 bp
N75 Length	2931 bp
L50	26,460
L75	14,819

Table 2. Cont.

Assembly Statistics	Data
Unigenes	
Total Number	110,485 (152.13 Mb)
Average Length	1376.95 bp
Number of HQ Unigenes	48,007 (76.01 Mb)
Average Length (HQ Unigenes)	1583.43 bp
N50 Length (HQ Unigenes)	2179 bp
GC% (HQ Unigenes)	39.87
Annotation of HQ Unigenes	
Database Searched	Unigenes with significant hits
Against NCBI-Plant-Genes	37,382
Against SwissProt DB	28,905
Against Pfam	34,752
With Gene Ontology (GO) terms	30,020
With Enzyme Commission (EC) numbers	9984
All annotated transcript	37,442
With No Significant hit	10,565

3.2. Functional Annotation of HQ Unigenes

A total of 37,382 unigenes (77.87%) were found to show significant matches against downloaded plant proteins with known functions with top BLAST hits having E -value cut-off $\leq 1 \times 10^{-10}$ (Table 2). As the efficiency of BLAST hits can be judged best with the length of query transcripts sequences [41], we plotted length of unigenes against the percentage of unigenes with significant matches. We found a very high proportion of matching efficiency for longer assembled unigenes (98% for >3 kb long transcripts; and 94.69% for transcripts between 1 kb and 3 kb long), however only 26.47% of shorter transcripts (≤ 500 bp) could find a match (Figure 1a). Similarity distribution of the best BLAST hits revealed that 70.33% the unigenes with significant matches had sequence similarity in the range of 80% to 100% while 29.37% unigenes had 50% to 80% similarity (Figure 1b). E -value distribution showed that 62.64% of the HQ unigenes had higher homology with top BLAST hits with small E -values ($\leq 1 \times 10^{-100}$) including 39.24% of unigenes which had perfect E -value of 0 (Figure 1c). Species distribution of BLAST results showed that 31.70% of unigenes with significant matches had top BLAST hits against *Glycine max*, followed by *Vigna angularis* (24.45%), *Cajanus cajan* (13.92%), and other legumes with less than 5% matches (Figure 1d). Further, 28,905 unigenes (60.21%) showed significant (E -value $\leq 1 \times 10^{-10}$) BLAST matches with Swiss-prot compared to that of 77.87% with APG database (Table 2). On comparison, we found that there were 28,845 unigenes (60.08%) with hits in both of these two databases. Finally, we obtained 37,442 unigenes (77.99%) with significant hits against any of these databases, which were considered as the annotated genes of *C. tetragonoloba* transcriptome (Table 2).

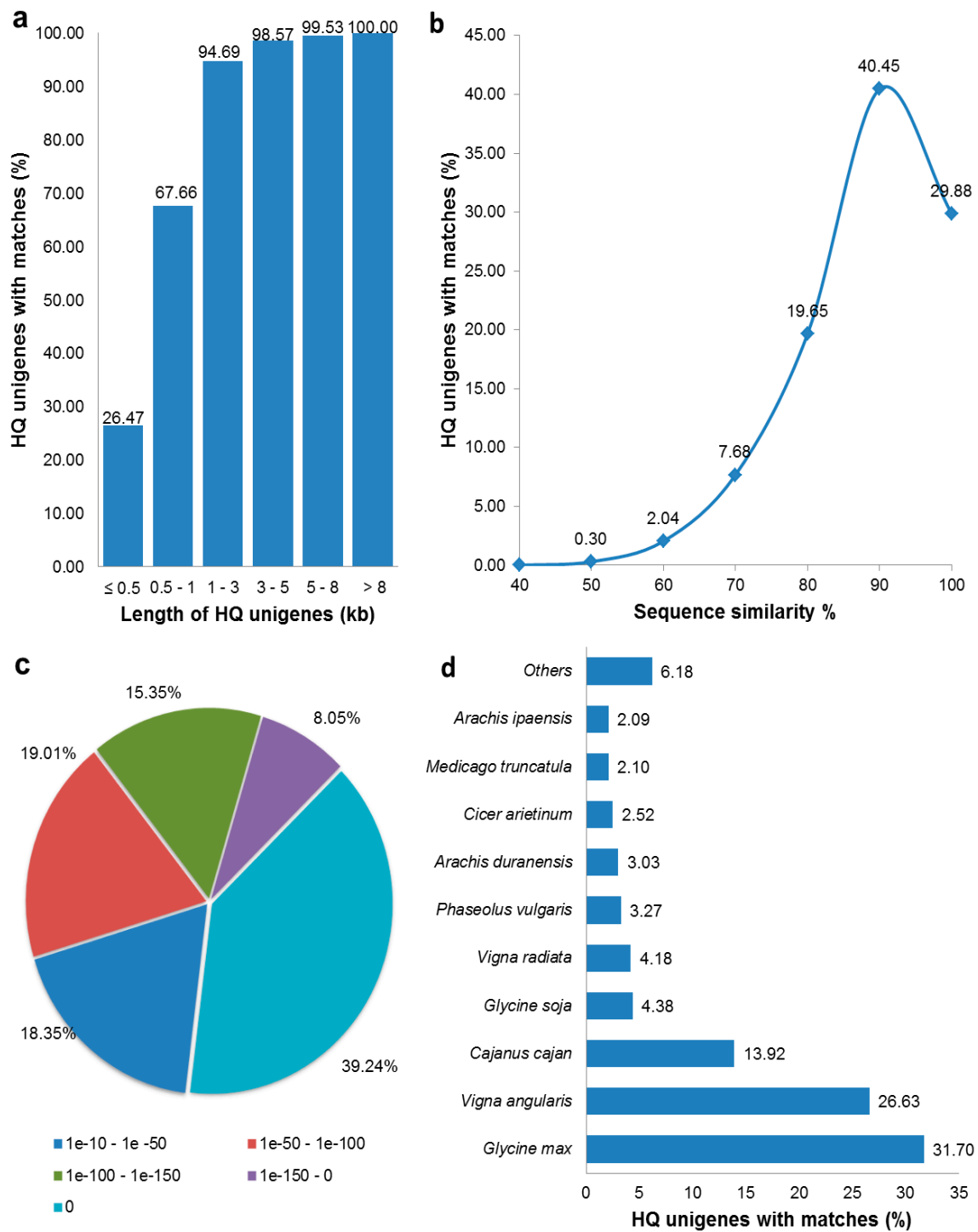


Figure 1. Statistics of BLAST search results of HQ Unigenes against Plant-genes database. (a) Length-wise distribution of HQ unigenes (query) sequence with significant matches ($E\text{-value} \leq 1 \times 10^{-10}$). A very high proportion (>98%) of large unigenes (>3 kb) showing significant matches; (b) similarity distribution of the best BLAST hits for each of the unigene with significant matches showing that 70.33% these having sequence similarity from 80 to 100%; (c) percent distribution of HQ unigenes on the basis of their E-values; (d) species distribution showing percentage of the HQ unigenes (query) sequence with significant matches against different species with maximum (31.70%) of these were having top BLAST hits against *Glycine max*.

3.3. Functional Classification with Gene Ontology Terms, Enzyme Commission Numbers and InterproScan Search

Out of the 37,382 unigenes, 30,020 unigenes were assigned GO terms while the remaining 9984 were assigned EC numbers (Table 2). A total of 116,964 GO terms were assigned to 30,020 unigenes and their distribution at GO level 2 is shown for molecular functions, biological processes and cellular components in Figure 2. Based on assigned GO terms, unigenes were found to be highly enriched in “cellular process” and “metabolic process” under biological process classification, “catalytic activity” and “binding” under the molecular function classification, and “cell” and “cell part” under the cellular component classification (Figure 2). Out of 9984 unigenes with EC numbers, maximum number of unigenes encoded hydrolases followed by transferases and oxidoreductases (Supplementary Figure S3). Investigation of the biological pathways identified a total of 6111 unigenes sequences mapped to 137 KEGG pathways.

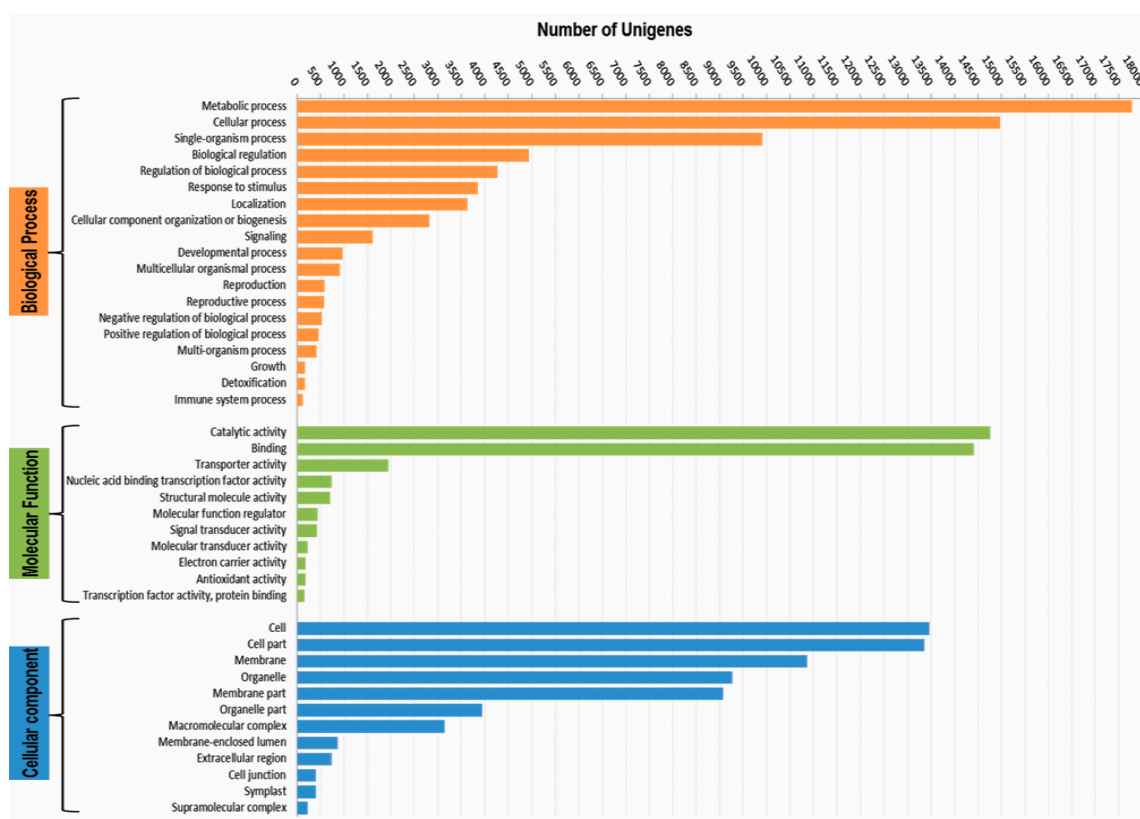


Figure 2. Functional classification of unigenes based on GO terms, showing GO category distribution of unigenes at GO level 2 into 3 categories: Biological Process, Molecular Function and Cellular Component.

InterproScan search revealed that maximum numbers of hits were found for Pentatricopeptide repeat (540 seq), Cytochrome P450 (241 seq) and Protein kinase-like domain (1601 seq) under the IPS repeat, family and domain category, respectively. Protein kinase domain [PF00069] showed maximum hits for IPS search against Pfam Database with 890 seq, followed by Serine-threonine/tyrosine protein kinase catalytic domain [PF07714], Pentatricopeptide repeat [PF13041], Pentatricopeptide repeat [PF001535] and RNA recognition motif domain [PF00076] (Figure 3).

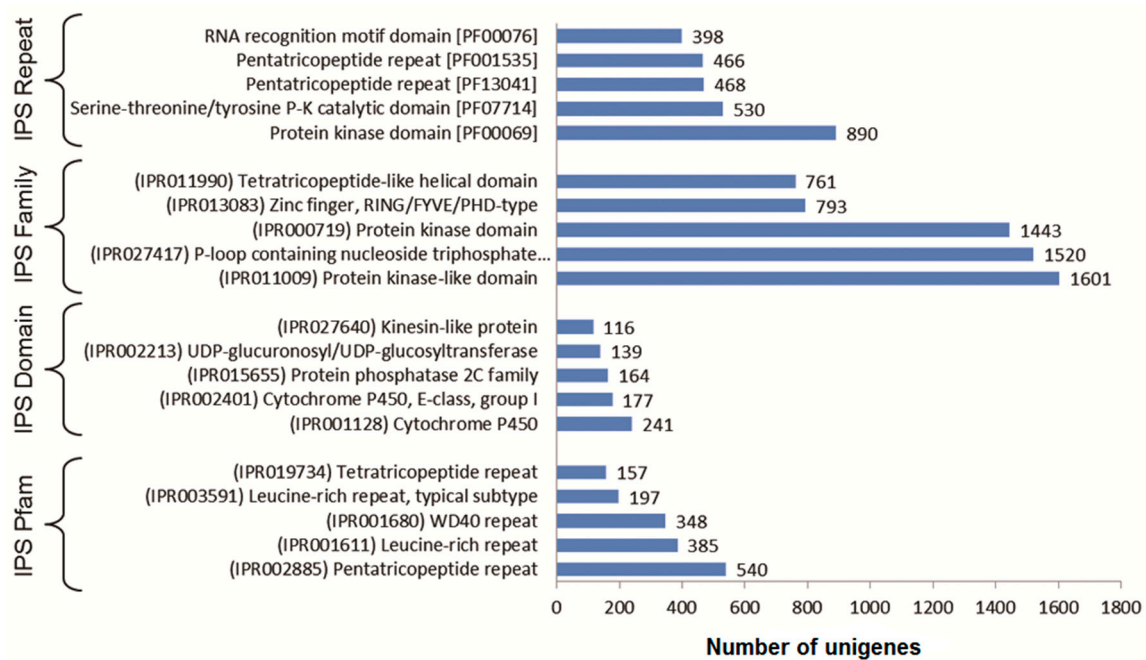


Figure 3. Result showing top 5 hits against interproscan repeat, family, domain and Pfam database.

3.4. Tissue-Specific Transcriptome Analysis

A total of 28,089 genes were identified as housekeeping genes with FPKM value ≥ 1 in each of the three tissue samples which included 12,754 (45.40%) genes with the highest FPKM value in shoot tissues (Figure 4a). At all fold level of comparisons namely 5, 8, 10, and 50, genes were found to be expressed higher in flower tissue sample (reproductive stage) followed by shoot and leaf (vegetative stage) (Figure 4b, Supplementary Table S2). A total of 790 genes were found to be highly tissue enriched with 50-fold higher FPKM value in one tissue compared to other tissues. Among these 790 highly tissue enriched genes, 58.48% were floral tissue specific followed by 22.15% and 19.37% in shoot and leaf respectively. Of the total 127,706 transcripts, a total of 50,394 were found “un-expressed” with < 1 FPKM in all 3 tissues. Among the expressed genes (at a threshold of 5-fold FPKM), 16,650 were found to be tissue specific genes while 10,678 were “genes with mixed-expression” (expressed in two tissues, whether low or high) and 49,984 as “Expressed in all” (expressed in all 3 tissues, whether low or high) which accounts for 21.54%, 13.81% and 64.65% of total expressed genes, respectively (Supplementary Table S3). Out of these, a small set of randomly selected genes from “Expressed in all” tissue-group was validated with quantitative real time PCR (qRT-PCR) and observed similar expression pattern (Supplementary Figure S4).

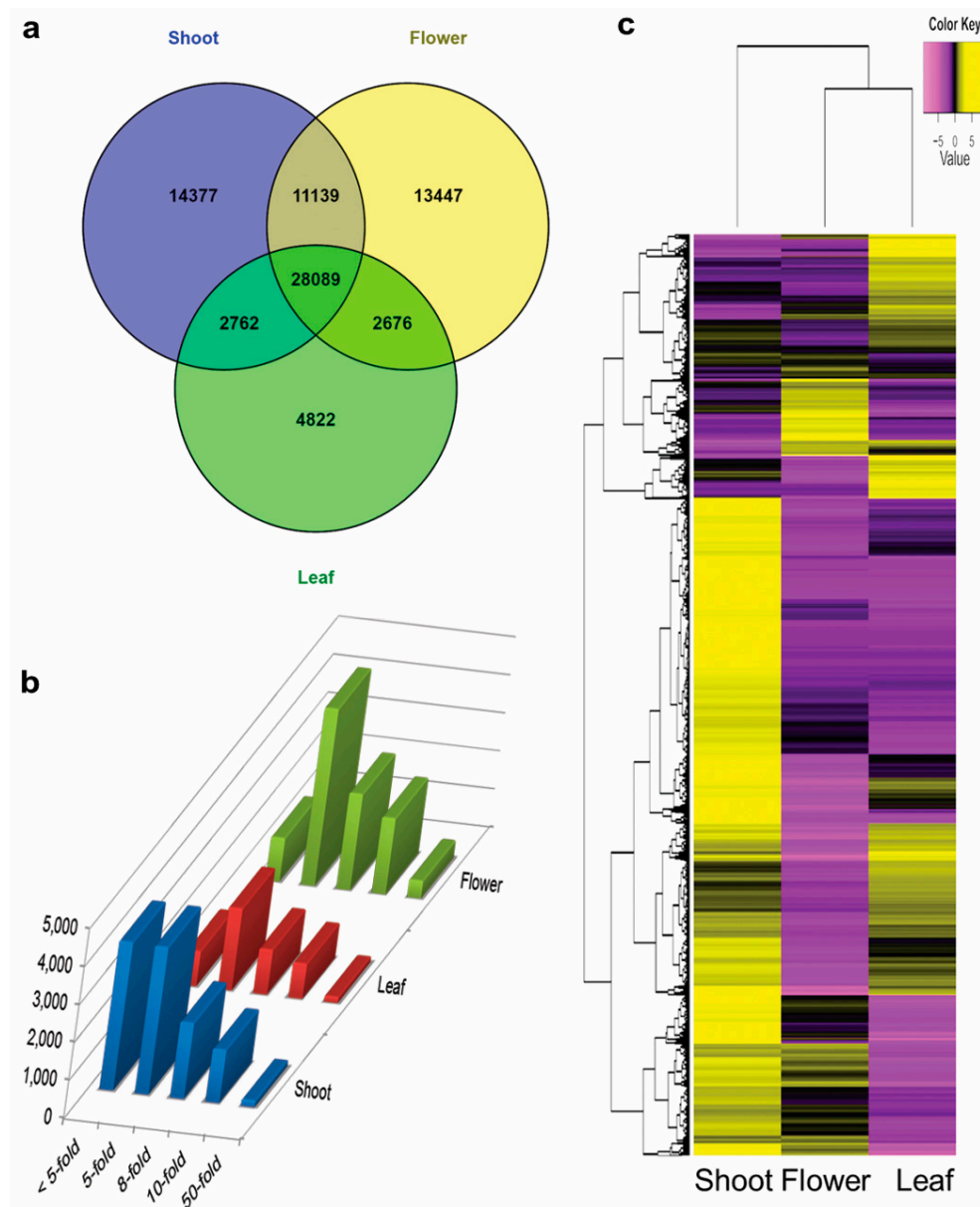


Figure 4. Tissue specific expression of *C. tetragonoloba*. (a) Expressed genes (FPKM ≥ 1) in 3 sample tissues; (b) expression of genes at different folds, with at each fold level of 5 or higher FPKM value, genes were found expressed higher in flower tissue sample (reproductive stage) as compared to tissue of vegetative stage (shoot and leaf); (c) differentially expressed genes (DEGs) vs. samples Heatmap showing cluster analysis of 38,423 differentially expressed genes for tissue-specific expression in all the 3 tissues. DEGs partitioned into 10 gene clusters with similar expression patterns with genes in each cluster ranging from 3755 to 3940. Color scale representing normalized expression values (left-top).

3.5. Differential Expression of Genes

A total of 38,423 Differentially Expressed Genes (DEGs) were identified at a p -value cut-off of ≤ 0.05 with at least two-fold change (C value ≥ 1). MA and Volcano plots of DEGs for Flower vs. Leaf, Flower vs. Shoot and Leaf vs. Shoot comparison are given in Supplementary Figure S5 wherein significant differential expression is represented as red dots while genes with “no significant expression” is represented by black ones. A significant global and relative relation between the floral and vegetative tissue as well as between the two vegetative tissues was found with positive correlation coefficient of 0.446, 0.435, and 0.401 (Supplementary Figure S6). Cluster analysis partitioned the DEGs

into 10 gene clusters, each cluster having similar expression and comprising of 3755 to 3940 genes. The clustered heatmap suggested that flowers followed a contrasting transcriptomic profile to leaves, while shoots showed totally different profile from these two tissues (Figure 4c).

3.6. Simple Sequence Repeat Prediction

From the assembled transcripts and HQ unigenes, we obtained a total of 17,593 and 8687 SSRs with an average frequency of one SSR per 10.20 and 8.75 kb in assembled transcripts and HQ unigenes, respectively (Table 3). Out of total 48,007 HQ unigenes, 7047 (14.68%) were found to contain SSR and 1297 of these unigenes had more than one SSR with 590 of these present in compound formation. The most abundant class of repeat motifs was found to be of those trinucleotide (51.11%) followed by dinucleotides (43.10%) SSRs. Other repeat motifs were just a fraction of these amounting to 4.43%, 0.71% and 0.64%, of tetra, penta and hexanucleotide repeats respectively (Supplementary Tables S4 and S5).

Table 3. Statistics of simple sequence repeats (SSRs) identified by MISA.

Features	Transcripts	HQ Unigenes
Total number of sequences examined	127,706	48,007
Total size of examined sequences (bp)	179,507,503	76,015,970
Total number of identified SSRs	17,593	8687
Number of SSR containing sequences	14,566	7047
Number of sequences containing more than 1 SSR	2430	1297
Number of SSRs present in compound formation	1137	590
Frequency of SSRs	1 SSR/10.20 kb	1 SSR/8.75 kb

3.7. Marker Validation

A total of 40 SSR markers were chosen randomly for validation in 21 clusterbean genotypes. The details of the SSR containing sequences and primers synthesized are provided in Supplementary Table S6. Out of 40 SSR primers, 28 SSR primers resulted in amplification in the target clusterbean varieties and were thus validated (Figure 5a). Two of the markers amplified multiple fragments (PCR products) while 26 showed single amplicons. Of the 28 SSRs, 13 were polymorphic with a significant polymorphism rate of 46.43%. The polymorphic information content (PIC) of the polymorphic markers ranged from 0.091 to 0.363 with an average of 0.21. The 21 genotypes were clustered into two distinct groups based on these markers (Figure 5b). To test the cross transferability of the markers, PCR was done with three genotypes each of pigeonpea and chickpea. However, no amplification could be found in any of the 28 primers in these two legume species. The FASTA sequences of predicted SSRs in HQ unigenes were BLAST searched against the EST sequences of *Cajanus cajan* (25,576 ESTs) and *Cicer arietinum* (52,788 ESTs), but no single hit was found with any EST sequence of these two legumes indicating the presence of uncharacterized SSRs unique to the clusterbean genome.

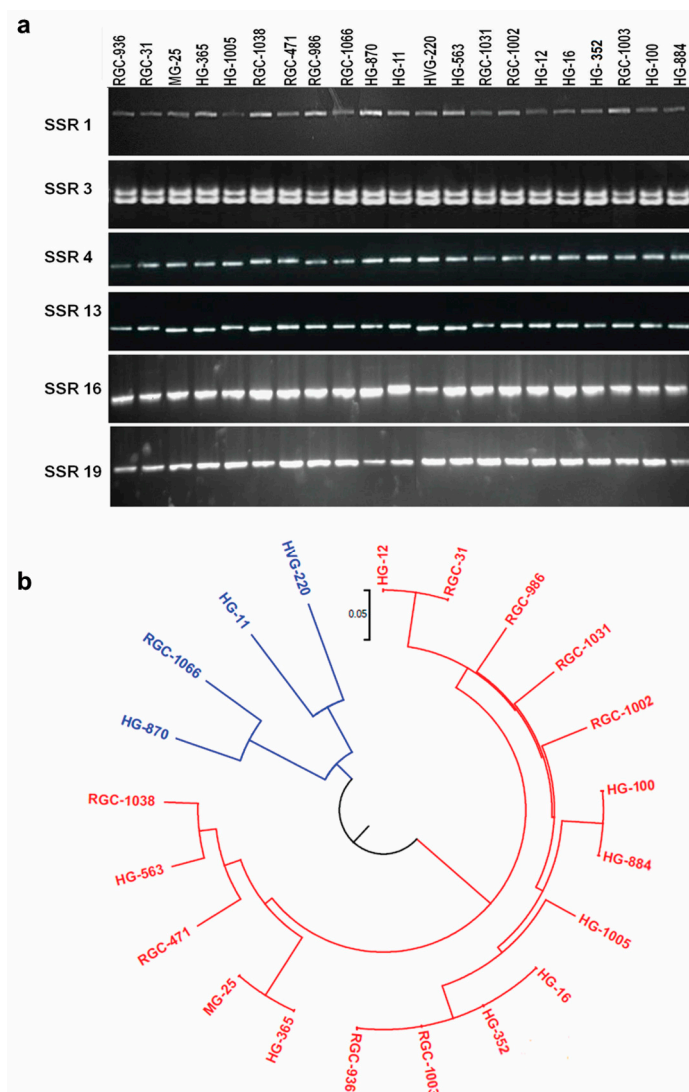


Figure 5. Validation of randomly selected simple sequence repeats from HQ unigenes. (a) Banding pattern of SSR primers' amplification on genomic DNA of 21 varieties of *C. tetragonoloba*; (b) genetic relationship among the 21 clusterbean accessions as revealed by the UPGMA method in the Numerical Taxonomy System (NTSYS-pc) ver. 2.1

3.8. Database for Clusterbean Unigenes and Microsatellite Markers

The clusterbean unigenes database has all the 48,007 unigenes identified in the present study along with their details on transcript length, protein description, Blast2Go annotation, *E*-value, expression status in leaf, stem and flower tissues and the SSR motifs present in them. Searching unigenes by GO IDs, GO names, enzyme IDs, enzyme names, key words such as MYB, WRKY, Zinc ion binding, chloroplast (targeted) and the IDs assigned by us has been enabled in the database (Supplementary Figure S7). Separate tabs for retrieving information on the expression status of selected genes in individual tissues, in batch of 10 genes is given. Since SSR markers remain breeders' choice, owing to the easy to genotype and locus specific nature of the markers, another tab for searching SSRs in unigenes has also been enabled. This information is also provided in the Supplementary Table S7.

4. Discussion

C. tetragonoloba is an annual legume crop and is a source of gum, food, fodder, and medicines [5,6,42]. It is a recently evolved and the only cultivated crop species among the 3 species of fabaceous genus

Cyamopsis [43]. To augment the genomic resources and to facilitate whole genome assembly and marker development in this industrially and nutritionally important but orphan crop, we carried out transcriptome sequencing from leaf, shoot and flower tissues of clusterbean variety RG 936 in the present study.

From the ~150 million paired-end raw reads generated, we could retain ~145 million high-quality trimmed reads (22.16 Gb) despite stringent standards (quality score of 33). With 127,706 (179.50 Mb) transcripts and 48,007 (76.01 Mb) high-quality unigenes and 22.16 Gb raw reads, the depth of sequencing of the present study was in the range of 150–300x. Average length of contigs and N50 are the two important aspects to judge the quality of an assembly and on these both accounts, our assembly was better than the recently reported de novo transcriptomes across plant species such as *Onobrychis viciifolia*, *Nicotiana benthamiana* and leaf transcriptome of clusterbean where the N50 values were in the range of 1000–1400 bp and average lengths were around 700–780 bp [16,44,45]. In our study, N50 value was >2 kb with the average length of assembled transcript reaching 1376 bp which further increased to 1583 bp for unigenes. The average length and N50 value of our assembly was also found better than other closely related legumes with published transcriptomes, including *Cicer arietinum* [46,47], *Arachis hypogaea* [48], *Vigna radiata* [49,50] and *Trifolium pratense* [51]. Among the 48,007 high-quality unigenes, 63% (30,276) were more than 1 kb long and only 17.63% (8466) were less than 500 bp, which is again very high compared to the legume transcriptomes available in the public domain. The genome completeness of this transcriptome assembly was checked using CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline [52] and the results showed that the transcriptome was complete to the tune of 98.79% with core eukaryotic genes (CEGs).

Of the 48,007 HQ unigenes of *C. tetragonoloba* transcriptome, 37,382 unigenes (77.87%) could be annotated even with a higher and more significant *E*-value cut-off of $\leq 1 \times 10^{-10}$. Only 22.13% unigenes had no significant matches, which could be either because of the high cut-off or shorter transcript length. Moreover, these genes may represent the novel or clusterbean lineage specific genes or non-coding RNAs. From the cDNA libraries of clusterbean, 27% of the genes could not be annotated and could be non-coding RNAs [15]. BLAST results showed that clusterbean unigenes had the highest similarity with *Glycine max* (31.70%) which is consistent with the study by Tanwar et al. [16]. However, in the latter study the other species that showed higher similarity (6–15%) were *Phaseolus vulgaris*, *Cicer arietinum*, *Sphingomonas melonis* and *Medicago truncatula*. Though *Phaseolus vulgaris* (3.27%), *Cicer arietinum* (2.52%) showed similarity with clusterbean HQ genes in our study, the degree of similarity was much lower, while *Sphingomonas melonis* and *Medicago truncatula* did not ever feature in the top similarity species list. Rather, all the three species of Phaseoleae tribe of legumes, namely *Glycine max* followed by *Vigna angularis* (26.63%) *Cajanus cajan* (13.92%) showed the highest matches in our study. Moreover, the presence of only legumes in the high level of similarity list with clusterbean indicated good coverage of the homologous legume sequences in the assembly [51].

Coding capacity of the genomes is well captured when multiple tissues and multiple growth conditions are used for transcriptome sequencing. Since ESTs/transcripts from seed and leaf tissues are already reported, our transcriptome study with leaves, shoots and flowers is an improvement over the available resources [15,16]. After meta-analysis, it would be possible to identify tissue specific genes for leaves, flowers, seeds and shoots.

Molecular markers identified from transcriptome-based studies are genic in nature and hence are expected to be more useful in molecular breeding applications. Transcriptome-based markers are helpful in contrast to the markers in non-transcribed regions owing to their high amplification rates and cross-species transferability [53]. Though SNPs are the markers of choice for understanding the trait architecture [54], for breeding application, SSRs and other gene/length polymorphism based markers are preferred. In the present work, one SSR per 8.75 kb in the HQ unigenes was identified which is in line with the earlier reports of one SSR per 7.31 kb [16] and one SSR per 7.9 kb [23] in clusterbean. Compared to other legumes, chickpea with one SSR per every 5.80 kb and common bean with one SSR per 4.70 kb, the frequency of SSRs in clusterbean seems to be lower [46,55].

Of the 40 randomly chosen SSRs, 28 could be validated by amplification. Such rates of successful amplification (70%) show that the majority of unigenes were precisely assembled. The failure of the remaining primer pairs to generate amplicons might be attributed to either long intervening introns or the location of primers across splice sites. This has also provided a novel set of genic SSRs to clusterbean research community. The cross-transferability studies with *Cajanus cajan* and *Cicer arietinum* failed to amplify SSRs even for a single primer pair. Following this, we subjected the complete set of FASTA sequences containing microsatellite markers and their flanking sequences to BLAST analysis with the two species and found no hits. This suggested that the clusterbean genome is unique and is quite diverse from the rest of the legumes including those belonging to the tribe Phaseoleae. Thus, further efforts would be required to generate multi-tissue transcriptomes, whole genome sequencing and useful molecular markers in this legume for breeding applications, evolutionary studies and understanding its genetic architecture.

5. Conclusions

The current study is the first report on multi-tissue developmental transcriptome from clusterbean and is third in a row after seed specific ESTs and leaf specific transcriptome. We have identified 48,007 high quality unigenes of which more than 98% have complete ORFs and 10,565 are clusterbean specific. Further, unigene sequence information and SSR markers have been provided in a database for easy access to researchers. Since clusterbean is an important crop in terms of gum and cosmetic industry, genomic and genetic information generated in the present study will serve as a platform for precise breeding applications.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4425/8/11/313/s1. Figure S1: Length distribution of HQ unigenes; Figure S2: GC content distribution of HQ unigenes; Figure S3: EC numbers, to categorize unigenes into 6 EC Classes; Figure S4: Validation of expression patterns of 7 randomly selected differentially expressed unigenes (based on FPKM values) using qRT-PCR to show the similar patterns in both FPKM and qRT-PCR expression values, in blue and red color bar, respectively. The FPKM expression value of leaf tissue has been normalized with qRT-PCR. The respective unigene IDs are shown at the top and Y-axis represents relative expression values; Figure S5: Volcano and MA plots of DEGs for all 3 possible pair of tissue samples: Flower vs. Leaf, Flower vs. Shoot and Leaf vs. Shoot, respectively, with red dots as significant expression and black ones representing 'no significant expression' [FDR = False Discovery Rate; FC = Fold Change]; Figure S6: A clustered heatmap showing the Pearson correlation matrix for pairwise comparison between three tissue samples by comparing the complete transcriptome; Figure S7: Screenshot of ClustergeneDB, a database for retrieving information on the unigenes of clusterbean; Table S1: Assembly with different assembler; Table S2: Number of expressed genes for expression level of genes at different fold FPKM value; Table S3: Categorization of expressed genes; Table S4: Distribution of SSRs in different repeat classes in Transcripts; Table S5: Distribution of SSRs in different repeat classes in HQ unigenes; Table S6: The details of the SSR containing sequences which were used for validation; Table S7: Primer sequences of SSR markers identified from HQ unigenes.

Acknowledgments: We acknowledge the financial support from Indian Council of Agricultural Research-Consortium Research Project on Genomics (ICAR-CRP Genomics) coordinated by National Bureau of Fish Genetic Resources (NBFGR), Lucknow, India. We also acknowledge the logistic support provided by PD, ICAR-NRCPB, New Delhi.

Author Contributions: K.G. conceived the study, designed the experiments, finalized the manuscript and coordinated the work. S.S. and A.T. collected the tissue, isolated RNA, prepared the genomic library, sequenced the library and filtered high quality reads. D.N. submitted the raw data and performed the preliminary analysis of data. H.C.R. analyzed the data, and wrote the first draft; S.K. carried out the SSR markers discovery and validation. P.K., and N.R.Y. provided the seeds of different varieties and contributed to manuscript. S.V.A.M. and A.U.S. interpreted data, wrote the manuscript and finalized the figures and tables. S.V. performed the database related work. K.G., N.R.Y., N.P.S., N.K.S. and T.R.S. contributed in data analysis and manuscript finalisation. All the authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krishnan, G.S.; Dwivedi, N.K.; Singh, J.P. Primitive weedy forms of guar, adak guar: Possible missing link in the domestication of guar *Cyamopsis tetragonoloba* (L.). *Genet. Resour. Crop Evol.* **2011**, *58*, 961–996. [[CrossRef](#)]
2. Francois, L.E.; Donovan, T.J.; Maas, E.V. Salinity effects on emergence, vegetative growth, and seed yield of guar. *Agron. J.* **1990**, *82*, 587–592. [[CrossRef](#)]

3. Ashraf, M.Y.; Akhtar, K.; Sarwar, G.; Ashraf, M. Evaluation of arid and semi-arid ecotypes of guar (*Cyamopsis tetragonoloba* L.) for salinity (NaCl) tolerance. *J. Arid Environ.* **2002**, *52*, 473–482. [[CrossRef](#)]
4. Dhugga, K.S.; Barreiro, R.; Whitten, B.; Stecca, K.; Hazebroek, J.; Randhawa, G.S.; Dolan, M.; Kinney, A.J.; Tomes, D.; Nichols, S.; et al. Guar seed β -mannan synthase is a member of the cellulose synthase super gene family. *Science* **2004**, *303*, 363–366. [[CrossRef](#)] [[PubMed](#)]
5. Vaughan, S.F.; Berhow, M.A.; Winkler-Moser, J.K.; Lee, E. Formulation of a biodegradable, odor-reducing cat litter from solvent-extracted corn dried distillers grains. *Ind. Crops Prod.* **2011**, *34*, 999–1002. [[CrossRef](#)]
6. Lubbe, A.; Verpoorte, R. Cultivation of Medicinal and Aromatic Plants for Specialty Industrial Materials. *Ind. Crop Prod.* **2011**, *34*, 785–801. [[CrossRef](#)]
7. Butt, M.S.; Shahzadi, N.; Sharif, M.K.; Shahzadi, N.; Sharif, M.K.; Nasir, M. Guar gum: A miracle therapy for hypercholesterolemia, hyperglycemia and obesity. *Crit. Rev. Food Sci. Nutr.* **2007**, *47*, 389–396. [[CrossRef](#)] [[PubMed](#)]
8. Mathur, N.K. *Industrial Galactomannan Polysaccharides*; CRC Press Taylor & Francis Group: Boca Raton, FL, USA, 2012; p. 187.
9. Sharma, P.; Gummagolmath, K.C. Reforming Guar Industry in India: Issues and Strategies. *Agric. Econ. Res. Rev.* **2012**, *25*, 37–48. [[CrossRef](#)]
10. Rai, D.K. *Trends and Economic Dynamics of Guar in India*; Indian Council for Research on International Economic Relations: New Delhi, India, 2015.
11. Kumawat, G.; Raje, R.S.; Bhutani, S.; Pal, J.K.; Mithra, S.V.; Gaikwad, K.; Sharma, T.R.; Singh, N.K. Molecular mapping of QTLs for plant type and earliness traits in pigeonpea (*Cajanus cajan* L. Millsp.). *BMC Genet.* **2012**, *13*, 84. [[CrossRef](#)] [[PubMed](#)]
12. Singh, N.K.; Gupta, D.K.; Jayaswal, P.K.; Mahato, A.K.; Dutta, S.; Singh, S.; Bhutani, S.; Dogra, V.; Singh, B.P.; Kumawat, G.; et al. The first draft of the pigeonpea genome sequence. *J. Plant Biochem. Biotechnol.* **2012**, *21*, 98–112. [[CrossRef](#)] [[PubMed](#)]
13. Varshney, R.K.; Chen, W.; Li, Y.; Bharti, A.K.; Saxena, R.K.; Schlueter, J.A.; Donoghue, M.T.A.; Azam, S.; Fan, G.; Whaley, A.M.; et al. Draft genome sequence of pigeonpea (*Cajanus cajan* L.), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **2010**, *30*, 83–89. [[CrossRef](#)] [[PubMed](#)]
14. Varshney, R.K.; Song, C.; Saxena, R.K.; Azam, S.; Yu, S.; Sharpe, A.G.; Cannon, S.; Baek, J.; Rosen, B.D.; Tar'an, B.; et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **2013**, *31*, 240–246. [[CrossRef](#)] [[PubMed](#)]
15. Naoumkina, M.; Torres-Jerez, I.; Allen, S.; He, J.; Zhao, P.X.; Dixon, R.A.; May, G.D. Analysis of cDNA libraries from developing seeds of guar (*Cyamopsis tetragonoloba* (L.) Taub). *BMC Plant Biol.* **2007**, *7*, 62. [[CrossRef](#)] [[PubMed](#)]
16. Tanwar, U.K.; Pruthi, V.; Randhawa, G.S. RNA-Seq of Guar (*Cyamopsis tetragonoloba*, L. Taub.) Leaves: De novo Transcriptome Assembly, Functional Annotation and Development of Genomic Resources. *Front. Plant Sci.* **2017**, *8*, 91. [[CrossRef](#)] [[PubMed](#)]
17. Kaila, T.; Chaduvla, P.K.; Rawal, H.C.; Saxena, S.; Tyagi, A.; Mithra, S.V.A.; Solanke, A.U.; Kalia, P.; Sharma, T.R.; Singh, N.K.; et al. Chloroplast Genome Sequence of Clusterbean (*Cyamopsis tetragonoloba* L.): Genome Structure and Comparative Analysis. *Genes* **2017**, *8*, 212. [[CrossRef](#)] [[PubMed](#)]
18. Punia, A.; Yadav, R.; Arora, P.; Chaudhury, A. Molecular and morphophysiological characterization of superior cluster bean (*Cyamopsis tetragonoloba*) varieties. *J. Crop Sci. Biotechnol.* **2009**, *12*, 143–148. [[CrossRef](#)]
19. Pathak, R.; Singh, S.K.; Singh, M. Assessment of genetic diversity in clusterbean using nuclear rDNA and RAPD markers. *J. Food Legum.* **2011**, *24*, 180–183.
20. Pathak, R.; Singh, S.K.; Singh, M.; Henry, A. Molecular assessment of genetic diversity in cluster bean (*Cyamopsis tetragonoloba*) genotypes. *J. Genet.* **2010**, *89*, 243–246. [[CrossRef](#)] [[PubMed](#)]
21. Sharma, P.; Kumar, V.; Raman, K.V.; Tiwari, K. A set of SCAR markers in cluster bean (*Cyamopsis tetragonoloba* L. Taub) genotypes. *Adv. Biosci. Biotechnol.* **2014**, *5*, 131–141. [[CrossRef](#)]
22. Gresta, F.; Mercati, F.; Santonoceto, C.; Abenavoli, M.R.; Ceravolo, G.; Araniti, F.; Anastasi, U.; Sunseri, F. Morpho-agronomic and AFLP characterization to explore guar (*Cyamopsis tetragonoloba* L.) genotypes for the 1059 Mediterranean environment. *Ind. Crop Prod.* **2016**, *86*, 23–30. [[CrossRef](#)]
23. Kumar, S.; Parekh, M.J.; Patel, C.B.; Zala, H.N.; Sharma, R.; Kulkarni, K.S.; Fougat, R.S.; Bhatt, R.K.; Sakure, A.A. Development and validation of EST-derived SSR markers and diversity analysis in cluster bean (*Cyamopsis tetragonoloba*). *J. Plant Biochem. Biotechnol.* **2016**, *25*, 263–269. [[CrossRef](#)]

24. Adams, M.D.; Kelley, J.M.; Gocayne, J.D.; Dubnick, M.; Polymeropoulos, M.H.; Xiao, H.; Merril, C.R.; Wu, A.; Olde, B.; Moreno, R.F.; et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **1991**, *252*, 1651–1656. [[CrossRef](#)] [[PubMed](#)]
25. Haas, B.J.; Volfovsky, N.; Town, C.D.; Troukhan, M.; Alexandrov, N.; Feldmann, K.A.; Flavell, R.B.; White, O.; Salzberg, S.L. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **2002**, *3*, research0029:1–research0029:12. [[CrossRef](#)]
26. Cloonan, N.; Forrest, A.R.R.; Kolle, G.; Gardiner, B.B.A.; Faulkner, G.J.; Brown, M.K.; Taylor, D.F.; Steptoe, A.L.; Wani, S.; Bethel, G.; et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **2008**, *5*, 613–619. [[CrossRef](#)] [[PubMed](#)]
27. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [[CrossRef](#)] [[PubMed](#)]
28. Li, B.; Ruotti, V.; Stewart, R.M.; Thomson, J.A.; Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **2010**, *26*, 493–500. [[CrossRef](#)] [[PubMed](#)]
29. Wilhelm, B.T.; Marguerat, S.; Goodhead, I.; Bahler, J. Defining transcribed regions using RNA-Seq. *Nat. Protoc.* **2010**, *5*, 255–266. [[CrossRef](#)] [[PubMed](#)]
30. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
31. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [[CrossRef](#)] [[PubMed](#)]
32. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Pribelski, A.D.; et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [[CrossRef](#)] [[PubMed](#)]
33. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [[CrossRef](#)] [[PubMed](#)]
34. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [[CrossRef](#)] [[PubMed](#)]
35. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
36. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [[CrossRef](#)] [[PubMed](#)]
37. Salem, M.; Paneru, B.; Al-Tobasei, R.; Abdouni, F.; Thorgaard, G.H.; Rexroad, C.E.; Yao, J. Transcriptome Assembly, Gene Annotation and Tissue Gene Expression Atlas of the Rainbow Trout. *PLoS ONE* **2015**, *10*, e0121778. [[CrossRef](#)] [[PubMed](#)]
38. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [[CrossRef](#)] [[PubMed](#)]
39. Thiel, T.; Michalek, W.; Varshney, R.K.; Graner, A. Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422. [[CrossRef](#)] [[PubMed](#)]
40. Doyle, J.J. Isolation of plant DNA from fresh tissue. *Focus* **1990**, *12*, 13–15.
41. Shi, C.Y.; Yang, H.; Wei, C.L.; Yu, O.; Zhang, Z.Z.; Jiang, C.J.; Sun, J.; Li, Y.Y.; Chen, Q.; Xia, T.; et al. Deep sequencing of the *C. sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genom.* **2011**, *12*, 131. [[CrossRef](#)] [[PubMed](#)]
42. Dwivedi, N.K.; Bhandari, D.C.; Dubas, B.S.; Agrawal, R.C.; Mandal, S.; Rana, R.S. *Catalogue on Cluster Bean (Cyamopsis tetragonoloba (L.) Taub) Germplasm Part III*; NBPGR: New Delhi, India, 1995.
43. Patil, C.G. Nuclear DNA Amount Variation in *Cyamopsis* D.C. (Fabaceae). *Cytologia* **2004**, *69*, 59–62. [[CrossRef](#)]
44. Nakasugi, K.; Crowhurst, R.N.; Bally, J.; Wood, C.C.; Hellens, R.P.; Waterhouse, P.M. *De novo* transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. *PLoS ONE* **2013**, *8*, e59534. [[CrossRef](#)] [[PubMed](#)]
45. Mora-Ortiz, M.; Swain, M.T.; Vickers, M.J.; Hegarty, M.J.; Kelly, R.; Smith, L.M.J.; Skøt, L. *De novo* transcriptome assembly for gene identification, analysis, annotation, and molecular marker discovery in *Onobrychis viciifolia*. *BMC Genom.* **2016**, *17*, 756. [[CrossRef](#)] [[PubMed](#)]

46. Garg, R.; Patel, R.; Tyagi, A.; Jain, M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* **2011**, *18*, 53–63. [[CrossRef](#)] [[PubMed](#)]
47. Kudapa, H.; Azam, S.; Sharpe, A.G.; Taran, B.; Li, R.; Deonovic, B.; Cameron, C.; Farmer, A.D.; Cannon, S.B.; Varshney, R.K. Comprehensive transcriptome assembly of Chickpea (*Cicer arietinum* L.) using Sanger and Next Generation Sequencing platforms: Development and applications. *PLoS ONE* **2014**, *9*, e86039. [[CrossRef](#)] [[PubMed](#)]
48. Zhang, J.; Liang, S.; Duan, J.; Wang, J.; Chen, S.; Cheng, Z.; Zhang, Q.; Liang, X.; Li, Y. De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). *BMC Genom.* **2012**, *13*, 90. [[CrossRef](#)] [[PubMed](#)]
49. Chen, H.; Wang, L.; Wang, S.; Liu, C.; Blair, M.W.; Cheng, X. Transcriptome sequencing of mung bean (*Vigna radiata* L.) genes and the identification of EST-SSR markers. *PLoS ONE* **2015**, *10*, e0120273. [[CrossRef](#)]
50. Liu, C.; Fan, B.; Cao, Z.; Su, Q.; Wang, Y.; Zhang, Z.; Wu, J.; Tian, J. A deep sequencing analysis of transcriptomes and the development of EST-SSR markers in mungbean (*Vigna radiata*). *J. Genet.* **2016**, *95*, 527–535. [[CrossRef](#)] [[PubMed](#)]
51. Chakrabarti, M.; Dinkins, R.D.; Hunt, A.G. De novo transcriptome assembly and dynamic spatial gene expression analysis in red clover. *Plant Genome* **2016**, *9*. [[CrossRef](#)] [[PubMed](#)]
52. Parra, G.; Bradnam, K.; Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **2007**, *23*, 1061–1067. [[CrossRef](#)] [[PubMed](#)]
53. Barbara, T.; Palma-Silva, C.; Paggi, G.M.; Bered, F.; Fay, M.F.; Lexer, C. Cross-species transfer of nuclear microsatellite markers: Potential and limitations. *Mol. Ecol.* **2007**, *16*, 3759–3767. [[CrossRef](#)] [[PubMed](#)]
54. Nepolean, T.; Singh, I.; Hossain, F.; Pandey, N.; Gupta, H.S. Molecular characterization and assessment of genetic diversity of inbred lines showing variability for drought tolerance in maize. *J. Plant Biochem. Biotechnol.* **2013**, *22*, 71. [[CrossRef](#)]
55. Wu, J.; Wang, L.; Li, L.; Wang, S. De novo assembly of the common bean transcriptome using short reads for the discovery of drought-responsive genes. *PLoS ONE* **2014**, *9*, e109262. [[CrossRef](#)] [[PubMed](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).