Computational Molecular Biology 2014, Vol. 4, No. 9, 1-6
http://cmb.biopublisher.ca

Research Article **Open Access**

Comparative study of five Legume species based on De Novo Sequence Assembly and Annotation

Sagar S. Patel[1], Dipti B. Shah[1], Hetalkumar J. Panchal[2]

1. G. H. Patel Post Graduate Department of Computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat-388120, India
2. Gujarat Agricultural Biotechnology Institute, Navsari Agricultural University, Surat, Gujarat- 395007, India

Corresponding author email: sgr308@gmail.com

Computational Molecular Biology, 2014, Vol.4, No.9 doi: 10.5376/cmb.2014.04.0009
Received: 03 Sep., 2014
Accepted: 25 Sep., 2014
Published: 23 Oct., 2014

Preferred citation for this article:
Patel et al., 2014, Comparative study of five Legume species based on De Novo Sequence Assembly and Annotation, Computational Molecular Biology, Vol.4, No.9, 1-6 (doi: 10.5376/cmb.2014.04.0009)

Abstract Legume species are an important oilseed crop in tropical and subtropical regions of the world. Recently, next-generation sequencing technology, termed RNA-seq, has provided a powerful approach for analysing the Transcriptome. This study is focus on RNA-seq of five legume species which are *Arachis hypogaea* L. (The peanut) of *SRR1212866*, *Cicer arietinum* L. of *SRR627764*, *Phaseolus vulgaris* L. of *SRR1283084*, *Trigonella foenum-graecum* L. of *SRR066197* and *Vicia sativa* L. of *SRR403901* from NCBI database. Comparative study focuses on various important features like; reads were generated with N50, sequence assembly contigs which is further searched with known proteins and genes; among these, how many genes were annotated with gene ontology (GO) functional categories and sequences mapped to pathways by searching against the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG). These data will be useful for gene discovery and functional studies and the large number of transcripts reported in the current study will serve as a valuable genetic resource of these five legume species.

Keywords De Novo assembly; Bioinformatics; Legume species; Sequence Assembly and Annotation

Introduction

Next generation sequencing methods for high throughput RNA sequencing (transcriptome) is becoming increasingly utilized as the technology of choice to detect and quantify known and novel transcripts in plants. This Transcriptome analysis method is fast and simple because it does not require cloning of the cDNAs. Direct sequencing of these cDNAs can generate short reads at an extraordinary depth. After sequencing, the resulting reads can be assembled into a genome-scale transcription profile. It is a more comprehensive and efficient way to measure Transcriptome composition, obtain RNA expression patterns, and discovers new exons and genes (Mortazavi et al., 2008; Wang et al.,2009); sequencing data of Transcriptome was assembled using various assembly tools, functional annotation of genes and pathway analysis carried with various Bioinformatics tools. The large number of transcripts reported in the current study will serve as a valuable genetic resource for described five legume species.

High-throughput short-read sequencing is one of the latest sequencing technologies to be released to the genomics community. For example, on average a single run on the Illumina Genome Analyser can result in over 30 to 40 million single-end (~35 nt) sequences. However, the resulting output can easily overwhelm genomic analysis systems designed for the length of traditional Sanger sequencing, or even the smaller volumes of data resulting from 454 (Roche) sequencing technology. Typically, the initial use of short-read sequencing was confined to matching data from genomes that were nearly identical to the reference genome. Transcriptome analysis on a global gene expression level is an ideal application of short-read sequencing. Traditionally such analysis involved complementary DNA (cDNA) library construction, Sanger sequencing of ESTs, and microarray analysis. Next generation sequencing has become a feasible method for increasing sequencing depth and coverage while reducing time and cost compared to the traditional Sanger method (L J Collins et al.).

1 Methods

1.1 Sequence Retrieval

This study is focus on the de novo assembly and sequence annotation of five legume species which are *Arachis hypogaea* L. (The peanut) of *SRR1212866*, *Cicer arietinum* L. of *SRR627764*, *Phaseolus vulgaris* L. of *SRR1283084*, *Trigonella foenum-graecum* L. of *SRR066197* and *Vicia sativa* L. of *SRR403901* from NCBI database for de novo Transcriptome analysis. Raw data downloaded from NCBI SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/) which are from Illumina HiSeq 2000 platform and LS454 platform- 454 GS FLX. Raw sequence was converted into fastq file format for further annotation with the use of SRA TOOL KIT from NCBI (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software).

1.2 NGS QC Toolkit

NGS QC Toolkit, it is an application for quality check and filtering of high-quality data. This toolkit is a standalone and open source application freely available at http://www.nipgr.res.in/ngsqctoolkit.html. The toolkit is comprised of user-friendly tools for QC of sequencing data generated using Roche 454 and Illumina platforms, and additional tools to aid QC (sequence format converter and trimming tools) and analysis (statistics tools). A variety of options have been provided to facilitate the QC at user-defined parameters. The toolkit is expected to be very useful for the QC of NGS data to facilitate better downstream analysis (Patel RK, et al).

1.3 De novo sequence assembly by CLC GENOMICS WORKBENCH 7

A comprehensive and user-friendly analysis package for analyzing, comparing, and visualizing next generation sequencing data. This package was used for de novo sequence assembly of sequence with by default parameters of de novo assembly tool (http://www.clcbio.com/products/clc-genomics-workbench/).

1.4 BLASTX

The assembled file was further considered for annotation in which first step was to identify translated protein sequences from contigs. BLASTX at NCBI (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) performed with changing few parameters like non redundant protein database (nr) selected as Database; *Eudicots* selected in organism option and in Algorithm parameters Max target Sequences set to 10 and Expect threshold set to 6.

1.5 Blast2GO

Blast2GO is an ALL in ONE tool for functional annotation of (novel) sequences and the analysis of annotation data (http://www.blast2go.com/b2ghome). Based on the results of the protein database annotation, Blast2GO was employed to obtain the functional classification of the unigenes based on GO terms. The transcript contigs were classified under three GO terms such as molecular function, cellular process and biological process (Ness et al., 2011; Shi et al., 2011; Wang et al., 2010). WEGO (http://www.wego.genomics.org.cn) tool was used to perform the GO functional classification for all of the unigenes and to understand the distribution of the gene functions of this species at the macro level. The KEGG database (http://www.genome.jp/kegg/pathway.html) was used to annotate the pathway of these unigenes.

1.6 SSR mining

We employed MIcroSAtellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/) for microsatellite mining which gives various statistical outputs of transcripts with useful information.

1.7 Plant transcription factor

PlantTFcat: An Online Plant Transcription Factor and Transcriptional Regulator Categorization and Analysis Tool used for identifying plant transcription factor in sequences (http://plantgrn.noble.org/PlantTFcat/).

2 Result and Discussions:

2.1 Sequence Comparison

(Tbale 1).

2.2 NGS QC Toolkit

Sequence was filtered with this tool by removing adaptors and other contaminated materials then quality of sequence also checked with this tool and finally high quality filter sequence file considered for de novo sequence assembly (Table 2).

Table 1 Species comparison based on sequence

Species	SRR Number	Reads	%GC Content	Platform
Arachis hypogaea L.	SRR1212866	7.3 M spots	48.5	Illumina HiSeq 2000
Cicer arietinum L.	SRR627764	36 M spots	41.8	Illumina
Phaseolus vulgaris L.	SRR1283084	20.4 M spots	46.4	Illumina HiSeq 2000
Trigonella foenum-graecum L.	SRR066197	627,117 spots	45.2	454 GS FLX
Vicia sativa L.	SRR403901	12.4 M spots	42.4	Illumina HiSeq 2000

Table 2 NGS QC Toolkit Result

Species	Total number of reads (Original File)	Total number of reads (High Quality (HQ) Filter file)	Total number of bases (Original File)	Total number of bases (High Quality (HQ) Filter file)	Percentage of HQ reads
Arachis hypogaea L.	7300624	7216150	365031200	360807500	98.84%
Cicer arietinum L.	1942297463	1942030133	1904983	1904959	99.99%
Phaseolus vulgaris L.	20444892	13418027	1042689492	684319377	65.63%
Trigonella foenum-graecum L.	627117	609237	146335656	141577237	97.15%
Vicia sativa L.	12427455	12131939	608945295	594465011	97.62%

2.3 De novo Sequence Assembly

CLC GENOMICS WORKBENCH 7 considered for de novo sequence assembly with by default parameters like Mismatch Cost = 2, Insertion Cost = 3, Deletion Cost = 3, Length Fraction = 0.5, Similarity Fraction = 0.8, Word size = 21 and contigs generated with average values by this software and other details are shown in Table 3.

Table 3 Contig measurement in Length

Species	N50	Minimum	Maximum	Average	Count (Contigs)
Arachis hypogaea L.	448	199	6635	425	10824
Cicer arietinum L.	1239	179	8439	805	34678
Phaseolus vulgaris L.	293	187	5386	302	6999
Trigonella foenum-graecum L.	470	86	3231	445	7256
Vicia sativa L.	588	197	6080	503	22748

2.4 Functional annotation with BLASTX and blast2GO

2.4.1 BLASTX

BLASTX was performed to align the contigs against non-redundant sequences database using an E value threshold of 10-6. Various statistical information of BLAST result is given in Table 4.

2.4.2 Enzyme Code (EC) Classification

Enzyme classified with sequences which are further classified into six classes which are of Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases which is shown in Table 5.

Table 4 Blast Result comparison

Species	Without Blast Results	Without Blast Hits	With Blast Results	With Mapping Results	Annotated Sequences	Total Sequences
Arachis hypogaea L.	60	688	4789	568	4719	10824
Cicer arietinum L.	3492	3996	25459	786	945	34678
Phaseolus vulgaris L.	102	2601	1988	629	1679	6999
Trigonella foenum-graecum L.	167	2656	1983	192	2258	7256
Vicia sativa L.	0	1114	13482	500	7652	22748

Table 5 Enzyme Code (EC) Classification

Species	Oxidoreducta---ses	Transferases	Hydrola-ses	Lyases	Isomera--ses	Ligases	Total
Arachis hypogaea L.	301	614	431	78	51	71	1546
Cicer arietinum L.	51	92	76	20	5	4	248
Phaseolus vulgaris L.	110	232	147	23	20	31	563
Trigonella foenum-graecum L.	148	149	179	34	38	27	575
Vicia sativa L.	429	927	718	80	82	100	2336

2.4.3 Gene Ontology (GO) Classification

To functionally categorize various legume transcript contigs, Gene Ontology (GO) terms were assigned to each assembled transcript contigs. Transcript contigs were grouped into GO functional categories (http://www.geneontology.org), which are distributed under the three main categories of Molecular Function, Biological Process and Cellular Components (Table 6).

Table 6 Gene Ontology (GO) Classification

Species	Molecular Function	Biological Process	Cellular Components	Total
Arachis hypogaea L.	4512 (43%)	3352 (32%)	2607 (25%)	10471
Cicer arietinum L.	916 (40%)	734 (32%)	654 (28%)	2304
Phaseolus vulgaris L.	1727 (47%)	1168 (31%)	829 (22%)	3724
Trigonella foenum-graecum L.	2792 (28%)	4407 (43%)	2980 (29%)	10179
Vicia sativa L.	7026 (37%)	5815 (31%)	5920 (32%)	18761

Figure 1 which is output of WEGO tool; it shows that, Within the Molecular Function category, genes encoding binding proteins and proteins related to catalytic activity were the most enriched. Proteins related to metabolic processes and cellular processes were enriched in the Biological Process category. With regard to the Cellular Components category, the cell and cell part were the most highly represented categories. We found same in all other legume species so we have considered only this one figure for illustration of WEGO tool.

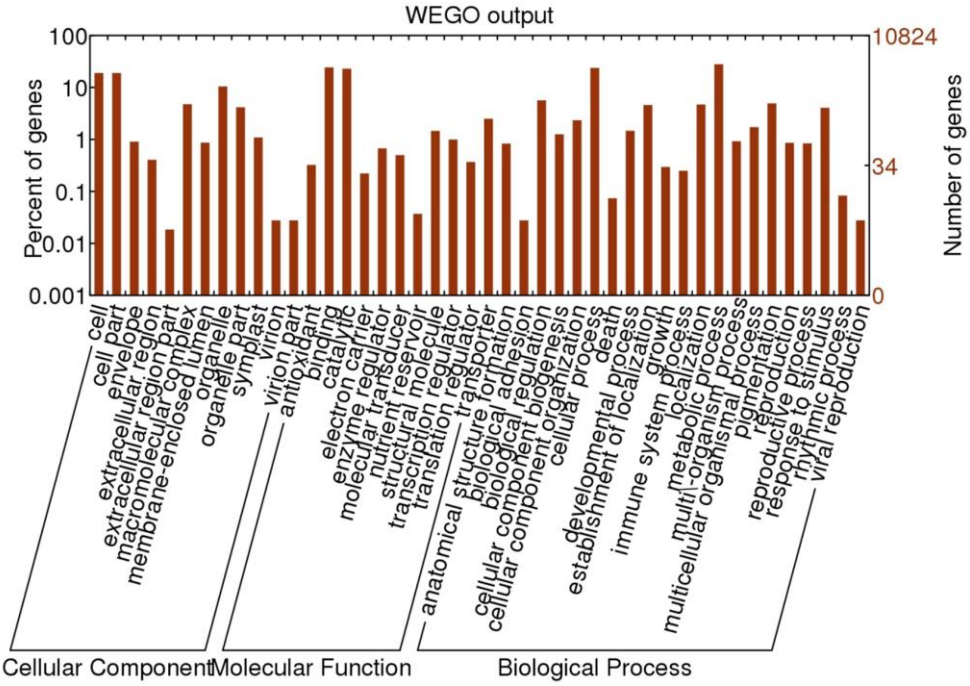

Figure 1 WEGO Tool Result of Arachis hypogaea L.

Many genes were annotated with different pathways in the KEGG database (http://www.genome.jp/kegg/pathway.html). Further comparative result is shown in Table 7. Many transcripts include various pathways like metabolic pathways, plant-pathogen interaction pathways, fatty acid metabolism pathway and fatty acid biosynthesis.

Table 7 KEGG Result

Species	Genes	KEGG Pathway
Arachis hypogaea L.	568	109
Cicer arietinum L.	786	78
Phaseolus vulgaris L.	629	89
Trigonella foenum-graecum L.	192	87
Vicia sativa L.	500	122

2.5 SSR mining

Microsatellite markers (SSR markers) are some of the most successful molecular markers in the construction of a peanut genetic map and in diversity analysis (Zhang et al). For identification of SSRs, all transcripts were searched with perl script MISA. SSR mining result is described in Table 8 which shows detailed information of each species' SSR result. The mono-nucleotide SSRs represented the largest fraction of SSRs identified followed by tri-nucleotide and di-nucleotide SSRs. Although only a small fraction of tetra-, penta- and hexa-nucleotide SSRs were identified in transcripts, the number is quite significant in most of species.

Table 8 Statistics of SSRs identified in transcripts

SSR Mining:	Species				
	Arachis hypogaea L.	Cicer arietinum L.	Phaseolus vulgaris L.	Trigonella foenum-graecum L.	Vicia sativa L.
Total number of sequences examined:	10824	34678	6999	7256	22748
Total size of examined sequences (bp):	4605095	27932177	2110290	3226271	11444673
Total number of identified SSRs:	742	5228	1405	3107	1150
Number of SSR containing sequences:	649	4391	1304	2191	1055
Number of sequences containing more than one SSR:	74	681	86	747	92
Number of SSRs present in compound formation:	48	337	64	747	48
Distribution to different repeat type classes:					
Mono-nucleotide	265	2019	1218	2589	362
Di-nucleotide	164	1271	87	235	243
Tri-nucleotide	299	1818	90	243	529
Tetra-nucleotide	10	78	7	28	10
Penta-nucleotide	2	17	2	10	3
Hexa-nucleotide	2	25	1	2	3

2.6 Plant Transcription Factor

Further, transcription factor encoding transcripts were identified by sequence comparison to known transcription factor gene families. Result in Table 9 shows that transcription factor genes distributed with families were identified and which is described in Table 9 and Figure 2 which is Plant Transcription Factor Result of *Trigonella foenum-graecum L.*.The overall distribution of transcription factor encoding transcripts among the various known protein families is very similar with that of other legumes as predicted earlier (Libault et al., 2009).

3 Conclusion

This study is focus on five different legume species from NCBI database for de novo sequence assembly and analysis by RNA-seq using next-generation Illumina and 454 sequencing. The transcriptome sequencing enables various functional genomics studies for an organism. Although several high throughput technologies have been developed for

Table 9 Plant Transcription Factor Result

Species	At least different families
Arachis hypogaea L.	70
Cicer arietinum L.	97
Phaseolus vulgaris L.	43
Trigonella foenum-graecum L.	45
Vicia sativa L.	82

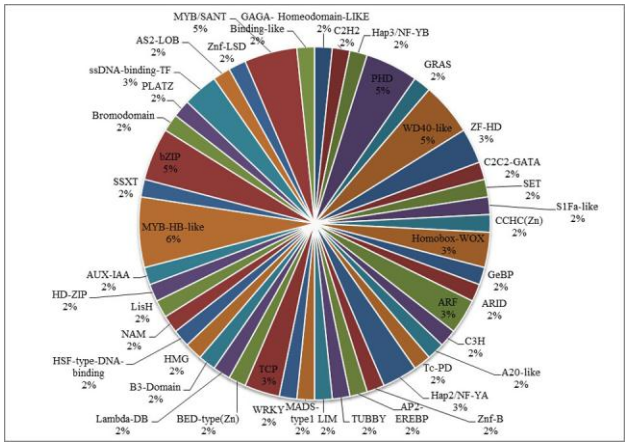

Figure 2 Plant Transcription Factor Result of *Trigonella foenum-graecum L.*

rapid sequencing and characterization of transcriptomes, expressed sequence data are still not available for many organisms, including many crop plants. In this study, we performed de novo functional annotation of five different legume species without considering any reference species with significant non-redundant set transcripts. The detailed analyses of the data set has provided several important features of five species such as GC content, conserved genes across legumes and other plant species, assignment of functional categories by GO terms and identification of SSRs by MISA tool. It is noted that this comparative study of five different legume species which are *Arachis hypogaea* L., *Cicer arietinum* L., *Phaseolus vulgaris* L., *Trigonella foenum-graecum* L. and *Vicia sativa* L. will be useful for further functional genomics studies as it includes useful information of each species with full annotation.

Acknowledgement

We are heartily thankful to Prof. (Dr.) P.V. Virparia, Director, GDCST, Sardar Patel University, Vallabh Vidyanagar, for providing us facilities for the research work.

References

Collins J. L., Biggs J. P., Voelckel C. and Joly S., 2008, An approach to transcriptome analysis of non-model organisms using short-read sequences, Genome Informatics 21:3-14
http://dx.doi.org/10.1142/9781848163324_0001

Jianan Zhang, Shan Liang, Jialei Duan, Jin Wang, Silong Chen, Zengshu Cheng, Qiang Zhang, Xuanqiang Liang and Yurong Li, 2012, De novo assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (Arachis hypogaea L.), BMC Genomics 2012 13:90
http://dx.doi.org/10.1186/1471-2164-13-90

Libault, M., Joshi, T., Benedito, V.A., Xu, D., Udvardi, M.K., and Stacey, G., 2009, Legume Transcription Factor Genes: What makes legumes so special?. Plant Physiology 151: 991-1001
http://dx.doi.org/10.1104/pp.109.144105

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5(7): 621-8
http://dx.doi.org/10.1038/nmeth.1226

Ness, R.W., Siol, M., and Barrett S.C.H., 2011, De novo sequence assembly and characterization of the floral transcriptome in cross and self-fertilizing plants, BMC Genomics 12: 298
http://dx.doi.org/10.1186/1471-2164-12-298

Patel RK, Jain M, 2012, NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data, PLoS ONE 7(2): e30619. doi:10.1371/journal.pone.0030619
http://dx.doi.org/10.1371/journal.pone.0030619

Rohini Garg, Ravi K. Patel, Akhilesh K. Tyagi, and Mukesh Jain., 2011, De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification, DNA RESEARCH 18, 53–63; doi:10.1093/dnares/dsq028
http://dx.doi.org/10.1093/dnares/dsq028

Shi, C.Y., Yang, H., and Wei, C.L., 2011, Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds, BMC Genomics 12: 131
http://dx.doi.org/10.1186/1471-2164-12-131

Vaidya K., Ghosh A., Kumar V, Chaudhary S, Srivastava N, Katudia K, Tiwari T and Chikara K., 2012, De novo transcriptome sequencing in Trigonella foenum-graecum to identify genes involved in the biosynthesis of diosgenin. The Plant Genome:doi: 10.3835/plantgenome 2012.08.0021
http://dx.doi.org/10.3835/plantgenome2012.08.0021

Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., and Liu, S.S., 2010, De novo characterization of a whitefly transcriptome and analysis of its gene expression during development, BMC Genomics 11: 400
http://dx.doi.org/10.1186/1471-2164-11-400

Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet. 10(1): 57-63
http://dx.doi.org/10.1038/nrg2484
http://www.blast.ncbi.nlm.nih.gov/Blast.cgi
http://www.blast2go.com/b2ghome
http://www.clcbio.com/products/clc-genomics-workbench/
http://www.genome.jp/kegg/pathway.html
http://www.ncbi.nlm.nih.gov/
http://www.nipgr.res.in/ngsqctoolkit.html
http://www.pgrc.ipk-gatersleben.de/misa/misa.html
http://www.plantgrn.noble.org/PlantTFcat/
http://www.trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
http://www.wego.genomics.org.cn