Research Article **Open Access**

De Novo RNA Seq Assembly and Annotation of *Vicia sativa* L. (SRR403901)

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

Genomics and Applied Biology, 2015, Vol.6, No.2 doi: 10.5376/gab.2015.06.0002
Received: 19 Nov., 2014
Accepted: 15 Jan., 2015
Published: 29 Jan., 2015

Preferred citation for this article:
Sagar S. Patel et al., 2015, De Novo RNA Seq Assembly and Annotation of *Vicia sativa* L. (SRR403901), Genomics and Applied Biology, Vol.6, No.2, 1-7 (doi: 10.5376/gab.2015.06.0002)

Abstract *Vicia sativa* L. which is also known as common vetch is nitrogen fixing leguminous plant in the family Fabaceae. Recently, next-generation sequencing technology, termed RNA-seq, has provided a powerful approach for analyzing the Transcriptome. This study is focus on RNA-seq of *Vicia sativa* L. of SRR403901 from NCBI database for de novo Transcriptome analysis. A total of 12.4 million single reads were generated with N50 of 588 bp. Sequence assembly contained total 22748 contigs which is further search with known proteins, a total of 7652 genes were identified. Among these, only 500 unigenes were annotated with 18761 gene ontology (GO) functional categories and sequences mapped to 122 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 the *Vicia sativa* L..

Keywords Transcriptome; Bioinformatics; *Vicia sativa* L..

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 *Vicia sativa* L.

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.).

Methods

1. Sequence Retrieval:

This study is focus on the de novo assembly and sequence annotation of *Vicia sativa* L. of *SRR403901* from NCBI database. Raw data downloaded from NCBI SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/?run= SRR403901) which is from Illumina HiSeq 2000 platform and the sample is single ended with 12.4 M spots and 42.4% GC content. Raw sequence was converted in to 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)

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).

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/).

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.

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.

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.

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/).

Results and Discussions

1. 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 1).

Table 1. NGS QC Toolkit Result

File Details	Original File	High Quality (HQ) Filter file
Total number of reads	12427455	12131939
Total number of bases	608945295	594465011
Percentage of HQ reads	--	97.62%

2. 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 finally 22748 contigs generated with average value of 503 by this software and other details are shown in Table 2.

Table 2. Contig measurement

Description	Length
N75	348
N50	588
N25	1056
Minimum	197
Maximum	6080
Average	503
Count (Contigs)	22748

3. Functional annotation with BLASTX and blast2GO

3.1 BLASTX

BLASTX was performed to align the contigs against non-redundant sequences database using an E value threshold of 10-6. Out of 22748 transcript contigs, 13482 were having BLAST hits to known proteins with high significant similarity and 1114 had no BLAST hits (Table 3). Out of total transcripts contigs, Table 4 and Figure 1 shows that species distribution in which 9819 sequences showed significant similarity with *Medicago truncatula* and least similarity was found with *Prunus mume* (24).

Table 3. Blast Result

Without Blast Results	0
Without Blast Hits	1114
With Blast Results	13482
With Mapping Results	500
Annotated Sequences	7652
Total Sequences	22748

Table 4. Blast Result of Species Distribution

Species	Blast Hit
Medicago truncatula	9819
Cicer arietinum	7942
Glycine max	1050
Pisum sativum	553
Phaseolus vulgaris	513
Lotus japonicus	168
Vicia faba	131
Vitis vinifera	118
Medicago sativa	81
Citrus sinensis	80
Cucumis sativus	74
Populus trichocarpa	73
Theobroma cacao	66
Trifolium pratense	65
Morus notabilis	60
Eucalyptus grandis	52
Prunus persica	46
Arabidopsis thaliana	46
Ricinus communis	41
Erythranthe guttata	38
Fragaria vesca	38
Jatropha curcas	38

UDP-forming	34
Solanum tuberosum	34
Eutrema salsugineum	33
Citrus clementina	28
Genlisea aurea	26
Capsella rubella	26
Prunus mume	24
others	337

3.2 Enzyme Code (EC) Classification

Enzyme classified with total of 2336 sequences which is further classified into six classes which are of Oxidoreductases (429), Transferases (927), Hydrolases (718), Lyases (80), Isomerases (82) and Ligases (100) which is shown in Figure 2.

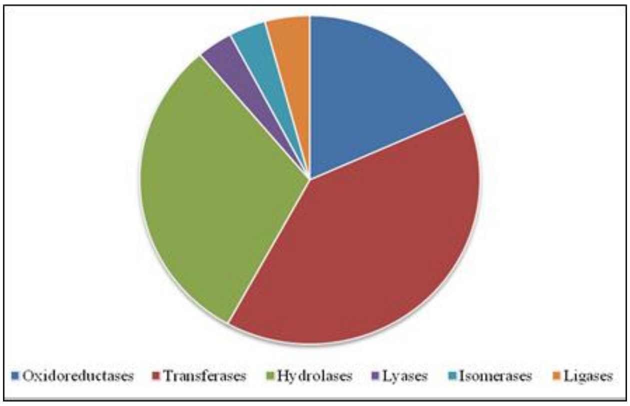

Figure 2 Enzyme Code (EC) Classification

3.3 Gene Ontology (GO) Classification

To functionally categorize *Vicia sativa* L. transcript contigs, Gene Ontology (GO) terms were assigned to each assembled transcript contigs. Out of 22748 transcript contigs, 18761 unigenes were grouped into GO functional categories (http://www.geneontology.org), which are distributed under the three main categories of Molecular Function (7026), Biological Process (5815) and Cellular Components (5920) (Figure 3). Figure 4 which is output of WEGO tool;

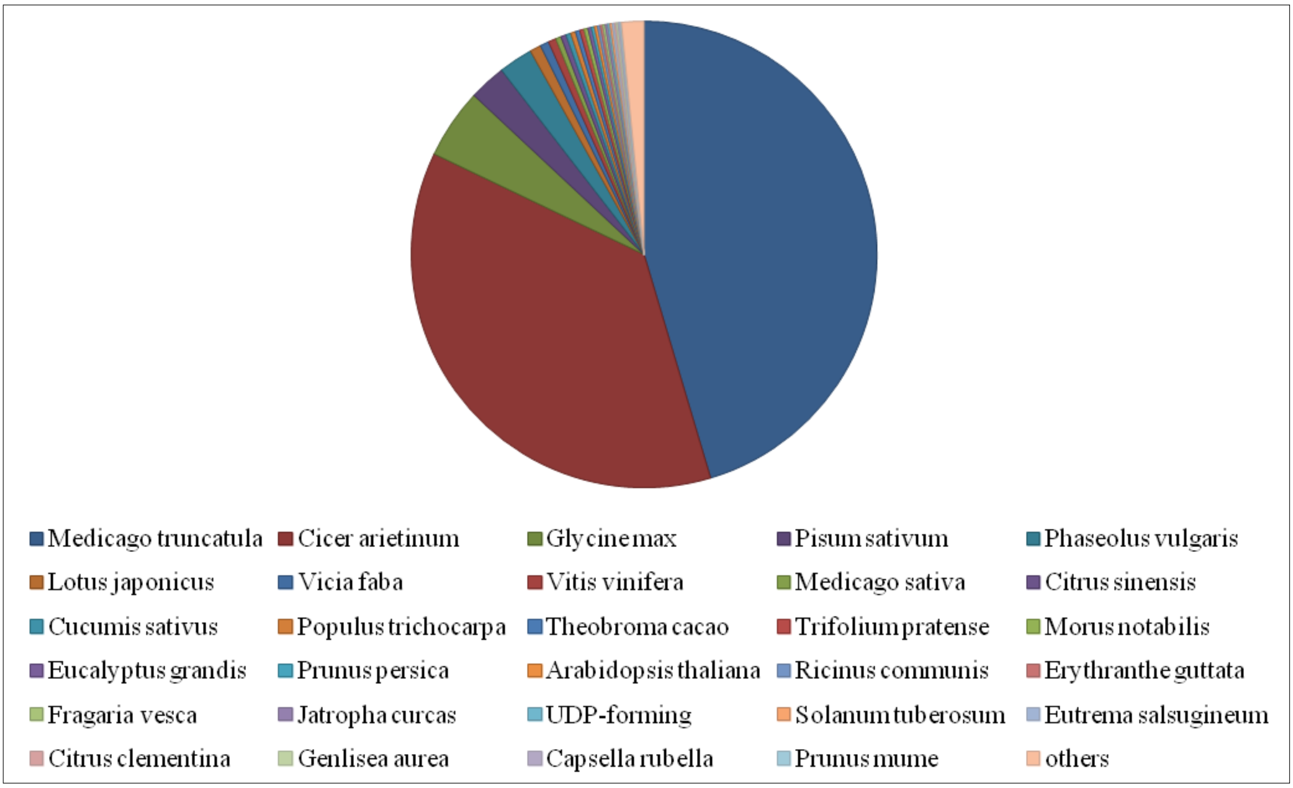

Figure 1. Blast Result of Species Distribution

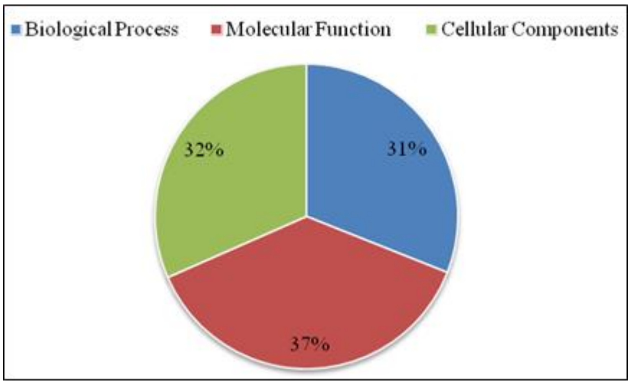

Figure 3 Gene Ontology Result

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.

A total of 500 unigenes were annotated with 122 pathways in the KEGG database (http://www.genome.jp/kegg/pathway.html). Many transcripts include various pathways like metabolic pathways, plant-pathogen interaction pathways, fatty acid metabolism pathway and fatty acid biosynthesis.

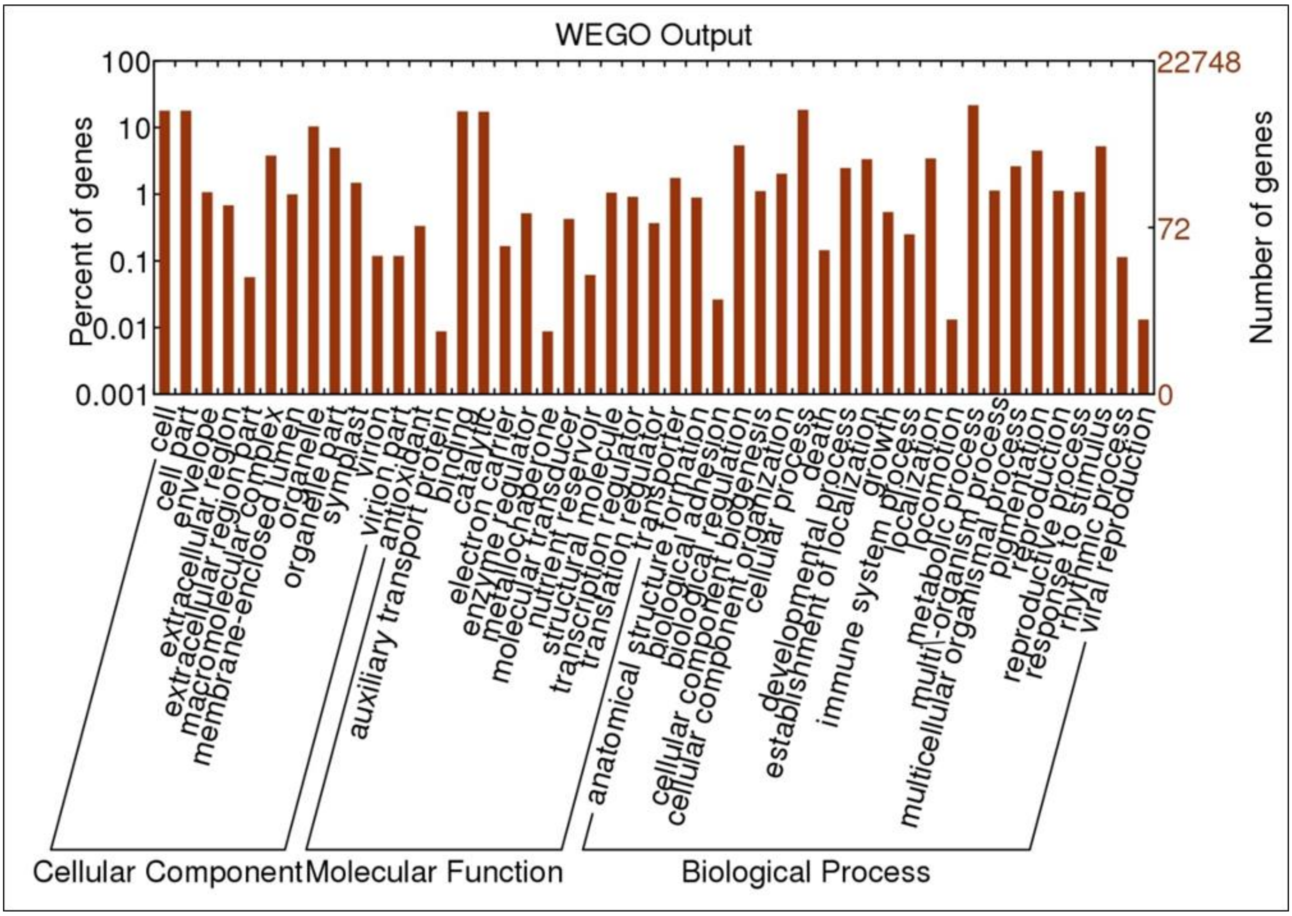

Figure 4 WEGO Tool Result

4. SSR mining

Microsatellite markers (SSR markers) are some of the most successful molecular markers in the construction of a *Vicia sativa* L. genetic map and in diversity analysis (Zhang et al). For identification of SSRs, all transcripts were searched with perl script MISA. We identified a total of 1150 SSRs in 1055 transcripts (Table 5). The mono-nucleotide SSRs represented the largest fraction of SSRs identified followed by tri-nucleotide and di-nucleotide SSRs. Although only

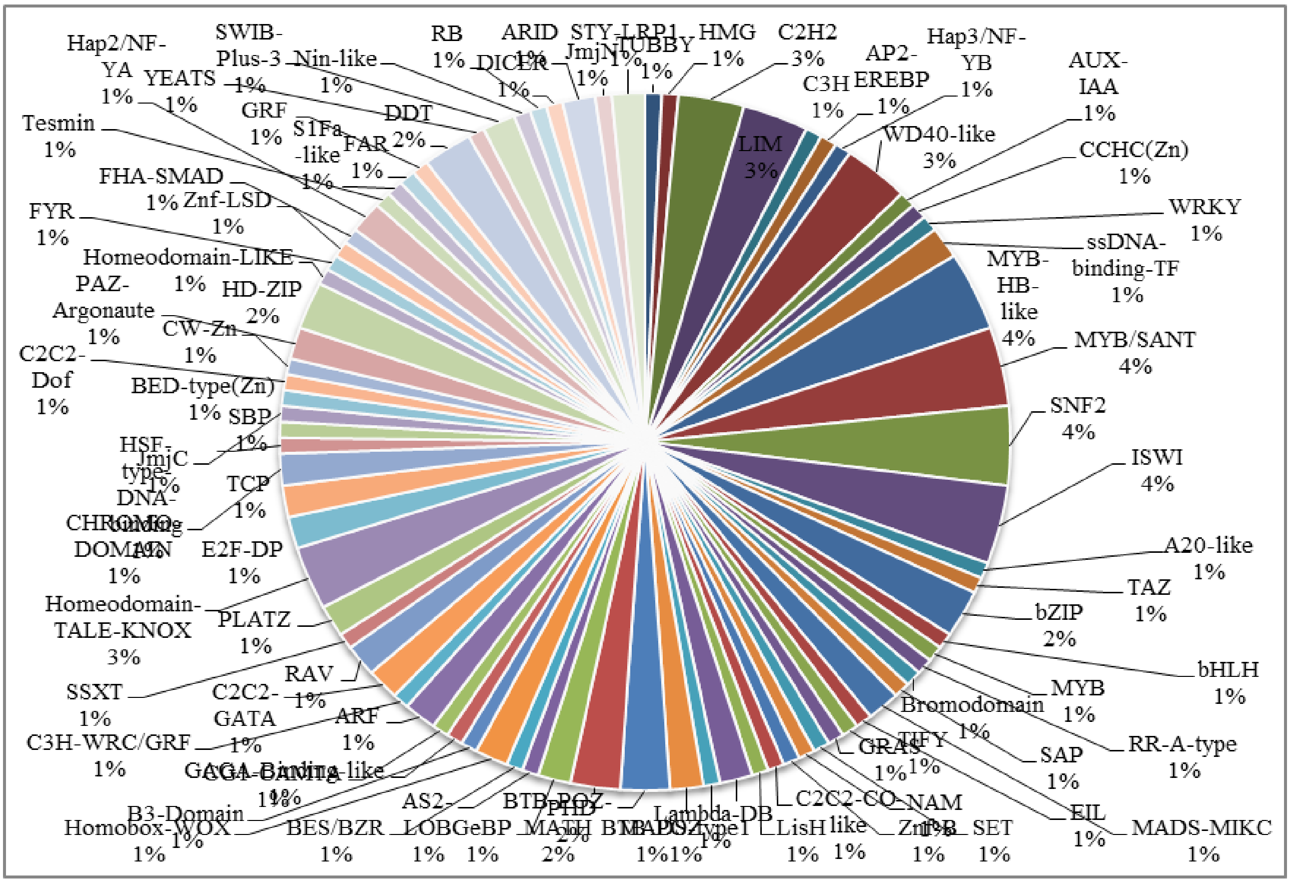

Figure 5 Plant Transcription Factor Result

a small fraction of tetra-, penta- and hexa-nucleotide SSRs were identified in transcripts, the number is quite significant.

5. Plant Transcription Factor

Further, transcription factor encoding transcripts were identified by sequence comparison to known transcription factor gene families. Result shows that transcription factor genes distributed with at least 82 families were identified (Figure 5). 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).

Conclusions

This study is focus on *Vicia sativa* L. species *(SRR403901)* from NCBI database for de novo Transcriptome analysis by RNA-seq using next-generation Illumina sequencing. The transcriptome sequencing enables various functional genomics

Table 5. Statistics of SSRs identified in transcripts

SSR Mining:	
Total number of sequences examined:	22748
Total size of examined sequences (bp):	11444673
Total number of identified SSRs:	1150
Number of SSR containing sequences:	1055
Number of sequences containing more than one SSR:	92
Number of SSRs present in compound formation:	48
Distribution to different repeat type classes:	
Mono-nucleotide	362
Di-nucleotide	243
Tri-nucleotide	529
Tetra-nucleotide	10
Penta-nucleotide	3
Hexa-nucleotide	3

studies for an organism. Although several high throughput technologies have been developed for 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 the *Vicia sativa* L. transcriptome without considering any reference species with significant non-redundant set of 34678 transcripts. The detailed analyses of the data set has provided several important features of *Vicia sativa* L. transcriptome 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 study of *Vicia sativa* L. will be useful for further functional genomics studies as it includes useful information of each transcript.

Acknowledgment

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 et al., 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

Garg et al., 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

Libault et al., 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 et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5(7): 621-8. http://dx.doi.org/10.1038/nmeth.1226

Ness et al., 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 et al., 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

Shi et al., 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 et al., 2012, De novo transcriptome sequencing in Trigonella foenum-graecum to identify genes involved in the biosynthesis of diosgenin. The Plant Genome:doi: 10.3835/plantgenome2012.08.0021. http://dx.doi.org/10.3835/plantgenome2012.08.0021

Wang et al., 2009. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet. 10(1): 57-63. http://dx.doi.org/10.1038/nrg2484

Wang et al., 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

Zhang et al., 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

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