International Journal of Agriculture, Environment and Biotechnology
Citation: IJAEB: 8(2): 295-302 June 2015
DOI Number: 10.5958/2230-732X.2015.00036.4

GENETICS

Gene Expression Study of *Arachis Hypogaea* L.

Sagar Patel[1*], Dipti Shah[1] and Hetalkumar 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: sgr308@gmail.com

Paper No. 320 Received: 25 August 2014 Accepted: 28 April 2015 Published: 29 June 2015

Abstract

Arachis hypogaea L. (The peanut) is an important oilseed crop in tropical and subtropical regions of the world. This species belongs to the subfamily Fabaceae and family Leguminosae. Different parts of the plant such as leaves and seeds are used for many purposes in India. Next-generation sequencing technology (NGS) such as RNA-seq has provided a powerful approach for analyzing the Transcriptome accurately and cheaply. This study is focus on gene expression study of RNA-seq of *Arachis hypogaea* L. (The peanut); Three SRA files of BioProject ID 243319 downloaded from NCBI database and genome of *Arabidopsis thaliana* was considered as reference genome for gene expression study. Data analysis carried out with many Bioinformatics tools such as TopHat2, Cufflinks, Cuffmerge, Cuffcompare and Cuffdiff. CummeRbund tool was used to manage, visualize and integrate all of the data produced by a Cuffdiff tool for gene expression analysis. These data reported in the current study will serve as a valuable genetic resource of the *Arachis hypogaea* L..

Highlights

- Gene expression study of Arachis hypogaea L. (The peanut) analyzed with three SRA files of BioProject ID 243319.
- Data analysis done with many Bioinformatics tools to get useful results.
- Next-generation sequence data analysis of Arachis hypogaea L. (The peanut) based on Gene expression study will be very useful for further study or analysis purpose.

Keywords: Transcriptome, gene expression, bioinformatics, *Arachis hypogaea* L.

The development of the peanut seed has been studied intensely to understand the physiological, biochemical, and molecular characteristics that determine the oil quality and their beneficial nutritional contributions. However, the development of the peanut seed is a complex process involving a cascade of biochemical changes, which involve the transcriptional modulation of many genes, yet little is known about these transcriptional changes and their regulation (Zhang *et al.* 2012).

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. This method is cheap as compare to other methods. 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).

High-throughput short-read sequencing is one of the latest sequencing technologies to be released to the genomics community.

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. 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.* 2008).

This study shows Gene Expression study of three different conditions of *Arachis hypogaea* L. which were treated with different method and analysis was done by using various Bioinformatics tools to get detail information of Gene Expression which is reported in current study.

Materials and Methods

Sequence Retrieval

As a part of research, three SRA sequences SRR1212866, SRR1212867 and SRR1212868 are downloaded from BioProject ID 243319 from NCBI database for Gene expression study. These data files submitted on 2-April-2014. These SRA files are converted into .fastq files by SRA TOOLKIT of NCBI.

NGS QC Toolkit

NGS QC Toolkit, it is an application for quality check and filtering of high-quality data. 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) (Patel RK, *et al.* 2012). NGS TOOLKIT package is used for sequence filtering and filtered sequences then uploaded to Galaxy server for *FASTQ GROOMER* process.

TopHat2

TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. TopHat can find splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. In this step, we have considered genome of *Arabidopsis thaliana* as the reference genome for Tophat2 analysis. This step was done for three times for each RNA-seq sample and all results of TopHat2 in .bam file format are considered for further Cufflinks analysis.

Cufflinks

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks assembles individual transcripts from RNA-seq reads that have been aligned to the genome (Cole Trapnell *et al.* 2012). In this step, we have used genes of *Arabidopsis thaliana* as reference (.gtf file) annotation in Cufflinks. This step was done for three times for each RNA-seq sample in .bam files which are output of TopHat2 and all results of Cufflinks (.gtf file format) which are considered for further Cuffmerge analysis.

Cuffmerge

Cuffmerge used to merge together several Cufflinks assemblies. In this step, three .gtf files of each three Cufflinks results and reference genes file of *Arabidopsis thaliana* considered as reference annotation and result was one merged transcript file of three transcripts.

Cuffdiff

Cuffdiff reports numerous output files containing the results of its differential analysis of the samples. Gene and transcript expression level changes are reported in simple tabular output files. Cuffdiff also reports additional differential analysis results beyond simple changes in gene expression. The program can identify genes that are differentially spliced or differentially regulated via promoter switching (Cole Trapnell *et al.* 2012). In this step, one merged file of

all three transcripts and individual mapped reads in .gtf file format are considered and result of Cuffdiff reported gene expression of three samples which is further analyzed with CummeRbund package in R language.

CummeRbund

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output. CummeRbund handles the transformation of Cuffdiff data into the R statistical computing environment, making RNA-seq expression analysis with Cuffdiff more compatible with many other advanced statistical analysis and plotting packages (Cole Trapnell *et al.* 2012).

Results and Discussion

Sequence Comparison

For gene expression study of *Arachis Hypogaea* L.; three RNA-Seq downloaded from NCBI database of SRA BioProject ID 243319 and details of each RNA-Seq is given in Table 1.

Table 1. Comparison of three RNA-Seq.

SRR Number	Spots	Bases	Size	%GC Content
SRR1212866	7.3 M	365.0 Mbp	254.2 M	48.5%
SRR1212867	7.7 M	383.8 Mbp	268.9 M	49.2%
SRR1212868	7.1 M	355.7 Mbp	240.1 M	44.1%

NGS QC Toolkit

Sequences filtered with this tool by removing adaptors and other contaminated materials then quality of sequence also checked and finally high-quality filter sequence file considered for further analysis (Table 2).

Gene Expression results

1. Overall result

Results shown in Table 3 are from CummeRbund tool, in which files from Cuffdiff tool inserted and after implementing few commands results are shown in Table 3, which shows numbers of differentially expressed genes in this study.

Table 3. Overview of Gene expression result.

Content	Result
samples	3
genes	27126
isoforms	35267
TSS	28643
CDS	32744
splicing	81378
relCDS	81048

2. Dispersion plot

Figure 2 shows the Dispersion plot of three RNA-Seq samples. All three samples are in different colors, and comparative view can easily see in Figure 2.

Table 2. NGS QC Toolkit Result.

SRR Number	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
SRR1212866	7300624	7216150	365031200	356155344	98.84%
SRR1212867	7676569	7591722	383828450	374677397	98.89%
SRR1212868	7113198	7016388	355659900	346595531	98.64%

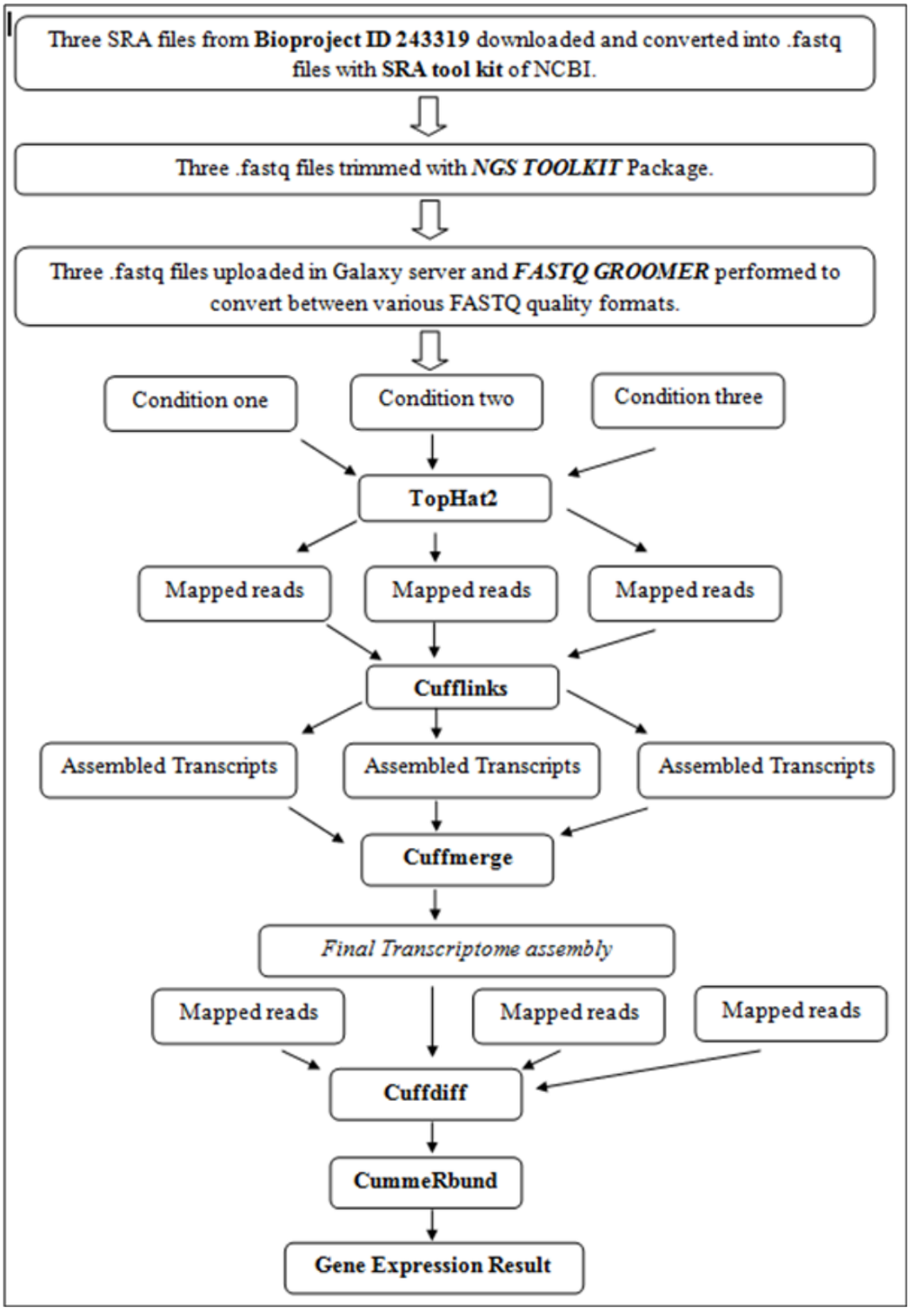

Figure 1. Pipeline for gene expression study.

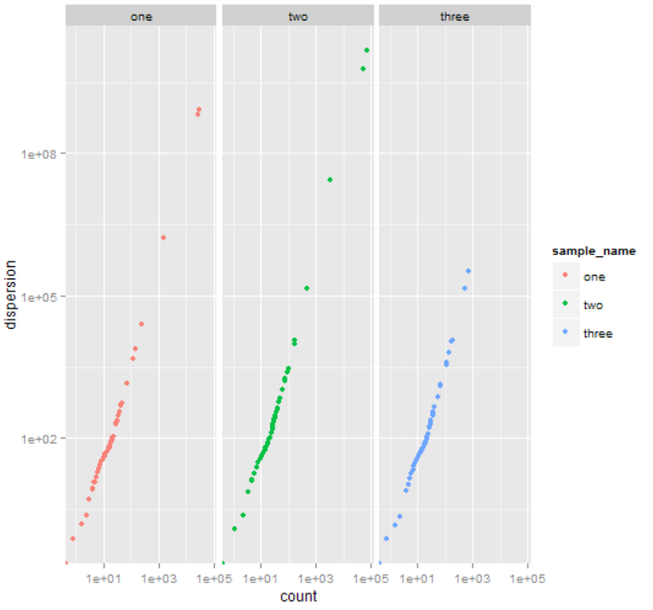

Figure 2. Dispersion plot of three RNA-Seq samples.

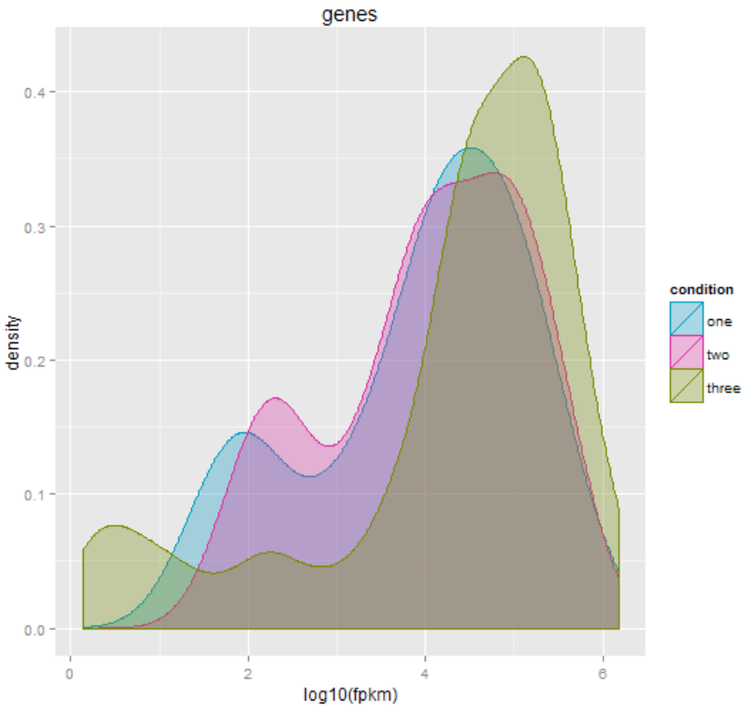

Figure 3. Gene density plot of three RNA-Seq samples.

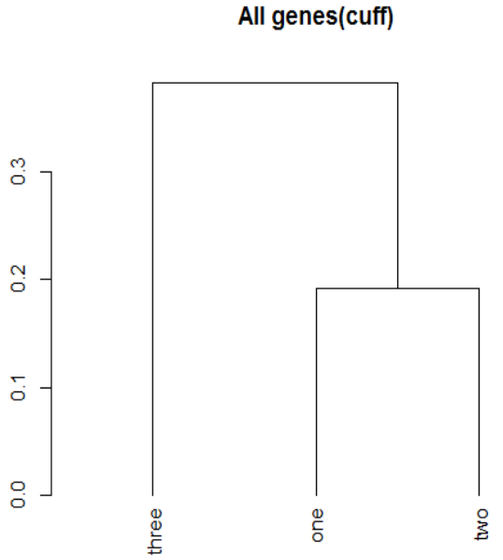

Figure 4. Dendrogram of three RNA-Seq samples.

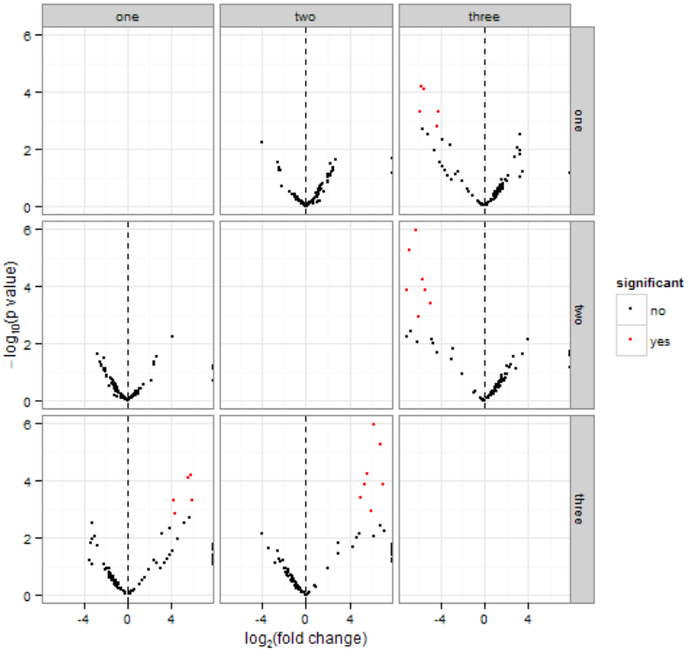

Figure 5. Volcano plots of three RNA-Seq samples.

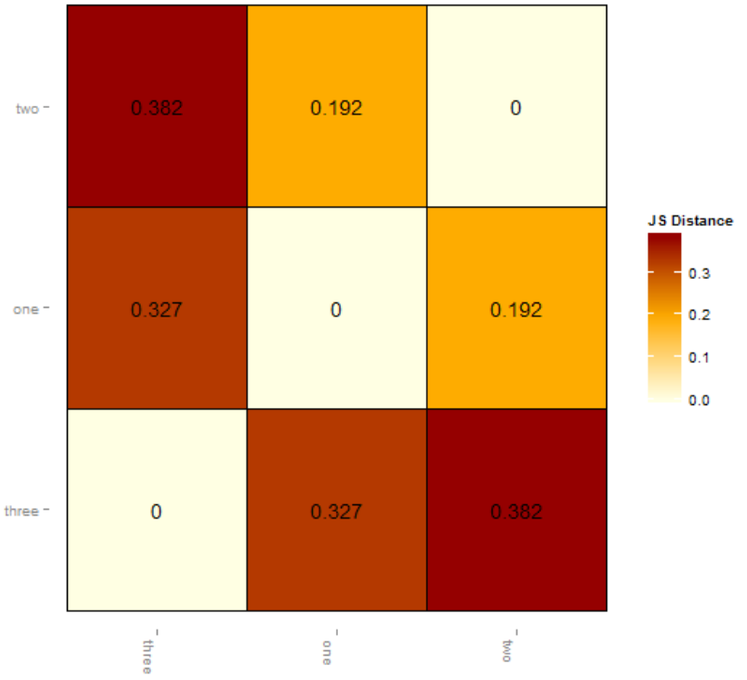

Figure 6. Pairwise similarities of three RNA-Seq samples.

3. Gene density plot

Figure 3 shows gene density plot of three RNA-Seq samples, which shows the distributions of FPKM scores across three samples with different scores.

4. Dendrogram

Figure 4 shows the dendrogram of three RNA-Seq samples for three different conditions. Sample three is distantly related with respect to samples one and two. While sample one and two share common node form a clade.

5. Volcano plots

Figure 5 shows the volcano plots of three RNA-Seq data in which, red dots shows significant genes while black dots shows non-significant genes.

6. Similarities between conditions

Similarities between conditions can provide useful insight into the relationship between various groupings of conditions and Figure 6 provides the csDistHeat() method to visualize the pair wise similarities between three conditions.

Conclusion

Next-generation sequencing (NGS) 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.* 2008). There are several software packages exists for short read alignment, and recently specialized algorithms for transcriptome alignment have been developed, e.g. TopHat2 for aligning reads to a reference genome to discover splice sites, Cufflinks to assemble the transcripts and compare/merge them with others and CummeRbund is designed to help simplify the analysis and exploration portion of RNA-Seq data derived from the output of a differential expression analysis using Cuffdiff with the goal of providing fast and intuitive access of results. (L J Collins *et al.* 2008).

This study is of Gene expression of *Arachis hypogaea* L. and as a part of research, three SRA sequences SRR1212866, SRR1212867 and SRR1212868 downloaded from BioProject ID 243319 from NCBI database. These SRA files are converted to .fastq files by SRA TOOLKIT of NCBI. For Gene expression study, TUXEDO Protocol considered for analysis, in which various Bioinformatics tools used for Gene expression study and CummeRbund tool used for visualization of data.

The results of this study shows that three RNA-Seq samples with total 27126 genes present in three different conditions and various interesting results obtained from CummeRbund tool. There are various statistical results also obtained from this study which are discussed in Result section. Next-generation sequence data analysis (NGS) of *Arachis hypogaea* L. (The peanut) based on Gene expression study will be very useful for further study or analysis purpose.

Acknowledgments

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

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**(7): 621-8. http://dx.doi.org/10.1038/nmeth.1226

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn and Lior Pachter 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578. http://dx.doi.org/10.1038/nprot.2012.016

Cheng-Ying Shi, Hua Yang, Chao-Ling Wei, Oliver Yu, Zheng-Zhu Zhang, Chang-Jun Jiang, Jun Sun, Ye-Yun Li, Qi Chen, Tao Xia and Xiao-Chun Wan 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

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* **13**:90. http://dx.doi.org/10.1186/1471-2164-13-90

Lesley J Collins, Patrick J Biggs, Claudia Voelckel and Simon Joly 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

Marc Libault, Trupti Joshi, Vagner A Benedito, Dong Xu, Michael K Udvardi and Gary Stacey 2009. Legume Transcription Factor Genes: What makes legumes so special?. *Plant Physiology* **151**: 991-1001. http://dx.doi.org/10.1104/pp.109.144105

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

Rob W Ness, Mathieu Siol and Spencer CH Barrett 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

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. http://dx.doi.org/10.1093/dnares/dsq028

Patel, Sagar and Panchal, Hetal Kumar 2014. Bioinformatics Information of Leguminosae Family in Gujarat State. *International Journal of Agriculture, Environment and Biotechnology* **7**(1): 11-15. http://dx.doi.org/10.5958/j.2230-732X.7.1.002

Thamodharan, G and Pillai, Arumugam M 2014.. Role of Antioxidative Enzymes Activity in Salt Stress and Salinity Screening in Rice Grown Under in vitro Condition. *International Journal of Agriculture, Environment and Biotechnology* **7**(2): 261-268. http://dx.doi.org/10.5958/2230-732X.2014.00243.5

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*: Published ahead of print 12 Dec. 2012; http://dx.doi.org/10.3835/plantgenome2012.08.0021

Xiao-Wei Wang, Jun-Bo Luan, Jun-Min Li, Yan-Yuan Bao, Chuan-Xi Zhang and Shu-Sheng Liu 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

Zhong Wang, Mark Gerstein and Michael Snyder 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1): 57-63. http://dx.doi.org/10.1038/nrg2484