आर.एन.ए-सैक डेटा का उपयोग करते हुए खेसारी में नमी तनाव के प्रति संवेदनशील जीनों का ट्रांस्क्रिप्टोम विश्लेषण

Transcriptome analysis of moisture stress responsive genes in *Lathyrus sativus* using RNA-Seq data

By

SNEHA MURMU

Master of Science

In

Bioinformatics

ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE

ICAR-INDIAN AGRICULTURAL RESEARCH INSTITUTE
NEW DELHI – 110012

2017

आर.एन.ए-सैक डेटा का उपयोग करते हुए खेसारी में नमी तनाव के प्रति संवेदनशील जीनों का ट्रांस्क्रिप्टोम विश्लेषण

Transcriptome analysis of moisture stress responsive genes in *Lathyrus sativus* using RNA-Seq data

by

SNEHA MURMU

A thesis submitted to the Faculty of the Post-Graduate School,
Indian Agricultural Research Institute, New Delhi,
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
IN
BIOINFORMATICS

Approved by:

Chairperson: _____
(Dr. SUNIL ARCHAK)

Co-Chairperson: _____
(Dr. K.V BHAT)

Members: _____
(Dr. RANJEET KUMAR PAUL)

जैव सूचना विज्ञान विभाग
भारतीय कृषि अनुसंधान संस्थान
पूसा कैंपस, नई दिल्ली- 110 012
INDIAN AGRICULTURAL RESEARCH INSTITUTE
रा पा आ सं ब्यूरो NBPGR

Sunil Archak
ICAR National Fellow
ICAR - National Bureau of Plant Genetic Resources
Pusa Campus, New Delhi 110 012

CERTIFICATE

This is to certify that the work incorporated in the thesis titled **"Transcriptome analysis of moisture stress responsive genes in *Lathyrus sativus* using RNA-Seq data"** submitted in partial fulfilment of the requirement for the degree of **MASTER OF SCIENCE** in **BIOINFORMATICS** of the P. G. School, ICAR-Indian Agricultural Research Institute, New Delhi, is a record of ***bona fide*** research work carried out by **Ms. Sneha Murmu** under my guidance and supervision and no part of this dissertation has been submitted for any other degree or diploma.

All assistance and help received during the course of this investigation has been duly acknowledged by her.

Date:

Place: New Delhi

(**Sunil Archak**)
Chairperson
Advisory Committee

ACKNOWLEDGEMENTS

This thesis is the fruitful outcome of the knowledge gained over the entire period of my M.Sc. study at ICAR-Indian Agricultural Statistics Research Institute (I.A.S.R.I), New Delhi during which I have been in touch with a great number of people whose contribution in varied yet myriad ways has led to the research and making of the thesis which deserves special mention. It is a pleasure to convey my gratitude to all of them by way of my humble acknowledgements.

First and foremost with reverence, I want to express deepest sense of gratitude to **Dr. Sunil Archak**, National Fellow, Division of Genomic Resources, National Bureau of Plant Genetic Resources (NBPGR), Indian Council of Agricultural Research (ICAR), New Delhi and Chairman of my Advisory Committee for his initiative, benevolence, endurance, constructive criticism and constant monitoring during the period of my investigation and also during preparation of this thesis. Above all and most needed, he provided me constant support and encouragement in various ways. I consider myself blessed having the privilege of being guided by him. I am really indebted to him.

I am also equally indebted to **Dr. K.V Bhat**, Principal Scientist, Division of Genomic Resources, ICAR-NBPGR and Co-Chairman of my advisory committee for his moral support during the course of my research work. He gave me constant encouragement from time to time for completion of my research work.

It is great privilege for me to express my esteem and profound sense of gratitude to **Dr. R.K Paul**, Scientist, Discipline of Agricultural Statistics, ICAR- I.A.S.R.I, New Delhi and Minor Member of my Advisory Committee for his valuable suggestions and help.

I would like to convey my deep sense of gratitude and appreciation to **Dr. A.R Rao** , Professor and Principal Scientist, Discipline of Bioinformatics, ICAR- I.A.S.R.I, New Delhi, for his help and support during my entire course work. I am indebted to him for his valuable advice and constructive criticism during preparation of this thesis.

I am deeply indebted to Dr. U. C. Sud, Director, ICAR-I.A.S.R.I for the help and infrastructures provided by him.

I want to express my respect and gratitude to my family, especially Grand- parents and parents for their sacrifices they did for me. They were the inspirations and moral boosting which gave me sufficient energy to complete this thesis in time.

I am thankful to my all batch mates, specially, Himanshu who has always helped me unconditionally, Shweta, Ritwika, Aamir, Asit, Dilip, Ramesh, Arpan, Yeasin, Akhilesh, Ronit, Samir, Ashish and Dipankar, for their friendly approach and moral support throughout my entire M.Sc. tenure.

I do not have any words to express the help and support given by my seniors, specially, Soumya Ma’am, Supriya Ma’am, Arfa Ma’am, Priyanka Ma’am, Sayanti Ma’am, Sonica Ma’am, Sapna Ma'am, Shaba Ma’am, Rajeev Sir, Asif Sir, Amit Sir, Anubhav Sir, Bulbul Sir, Nalini Sir, Sandeep Sir and my juniors, specially, Ankita, Garima,

Vivek, Lovkush, Lakshmi, Sayantani and Debdali for all their affectionate support and help.

I take this opportunity to appreciate the help rendered by the staff of TAC, ICARI.A.S.R.I. Special thanks to Sanjeev Sir and Gagan Sir.

I am indebted to Director, I.A.R.I, New Delhi and Dr. R. K. Jain, Dean and Joint Director (Education), I.A.R.I, for providing facilities to carry out this research work.

I like to thank all the staff of PG School for their helpful attitude and cooperation, throughout the period of study.

Last but not the least I am thankful to ICAR-I.A.S.R.I for the financial assistance provided to me in the form of Scholarship during the tenure of my study.

Finally, I would like to thank everybody who was imperative to the successful realization of this thesis, as well as articulate my apology that I could not mention personally one by one.

Date:

Place: New Delhi-110012 (Sneha Murmu)

CONTENTS

LIST OF ABBREVIATION

ABA	Abscisic Acid
ABI	Applied Biosystem
BLAST	Basic Local Alignment Search Tool
BP	Biological Process
BWT	Burrow-Wheeler Transformation
bZIP	Basic leucine-ZIPper
CAGE	Cap Analysis of Gene Expression
CC	Cellular Component
DEA	Differential Expression Analysis
DEGs	Differential Expressed Genes
DNA	DeoxyriboNucleic Acid
EST	Deoxyribonucleic acid
FDR	Expressed Sequence Tags
GC	Guanine Cytosine
GFF	General Feature Format
GO	Gene Ontology
GTP	Guanosine-5'-triphosphate
HPC	High Performance Computer
HT	High Throughput
KEGG	Kyoto Encyclopedia of Genes and Genomes
MF	Molecular Function
MPSS	Massively Parallel Signature Sequencing
MYB	Myoglobin
NCBI	National Centre for Biotechnology Information
NGS	Next-Generation Sequencing
NR	Non-Redundant
PE	Paired-Ends
PCR	Polymerase Chain Reaction
Q	Phred score or Quality score
RHEL	Red Hat Enterprise Limited
RNA	RiboNucleic Acid
RPKM	Reads per Kilobase of exon model per Million Mapped reads

SAGE	Serial Analysis of Gene Expression
SE	Single-Ends
SOAP	Short Oligonucleotide Analysis Package
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SRA	Sequence Read Archive
TF	Transcription Factor
TFDB	Transcription Factor Database
WD	Water Deficit

LIST OF TABLES

LIST OF FIGURES

Introduction

CHAPTER I

INTRODUCTION

Drought plays a very important role in plant growth and crop yields (Boyer, 1982). Drought is the most significant environmental stress on world agricultural production. It is the period during which the amount of moisture in the soil no longer meets the needs of the particular crop. The struggle to grow food crops during such scarcity of moisture is therefore a big issue. Drought tolerance has also been well documented to result from cooperative interactions among multiple morphological, physiological, and biochemical characters. Different genotypes may have diverse responses to drought stress (Shinozaki and Yamaguchi-Shinozaki, 2007). The evaluation or investigation for improved drought resistance is hence a major concern in breeding programs in many agriculturally important crops (Zivcak *et al.*, 2008). Lots of research has been dedicated for the improvement of plant responses to moisture or water deficit (WD) but majority of such efforts have been made upon cereal crops as compared to the crops belonging to the legume family (Jeuffroy and Ney, 1997). There is thus a need to increase the performance of pulse crops, particularly in developing countries, where most grain legume production is for human consumption and demand is increasing due to uncontrolled growth in population. Drought appears to be one of the major agronomic problems which limits the plant growth and yields. Therefore, efficient improvement requires an in-depth understanding of the gene expression regulation mechanisms in response to drought stress.

In legumes, drought resistance traits have been already identified (Nunes *et al.*, 2009) in *Medicago truncatula* which is a model legume, and deeper understanding regarding the molecular mechanisms that modulate the physiological response have also been achieved (Trindade *et al.*, 2010) in this plant. *Lathyrus* is a large genus with 187 species and subspecies among which *Lathyrus sativus* (2n=14) is the only cultivated species (Allkin *et al*., 1986) and it belongs to family Fabaceae and sub-family Papilionoideae. It is popularly known as grass pea or white pea and has great economic potential as food and feed crops. This plant shows high resilience to moisture stress conditions i.e., both drought and flood. It has also very high nutritive value. Grass pea (*Lathyrus sativus* L.) is a crop of immense economic and agronomic

importance which has multiple uses as food, feed and fodder and hence it is used for both human and livestock consumption. It is mostly cultivated in developing nations including India, Bangladesh, Pakistan, Nepal, and Ethiopia (Kumar *et al.*, 2011; Tripp and Heide, 1996) and also in China and in many countries of Europe, the Middle East, and Northern Africa. *Lathyrus sativus* offers an attractive choice for sustainable food production, owing to its intrinsic properties including limited water requirement and drought tolerance. Studies like the one conducted by Talukdar (2013) has proven that when compared with other legume crops like lentil, plant growth traits and seed yields components reduced significantly in both the crops but the effect was more severe in lentil compared with grass pea. Proline level increased significantly in both crops, but it decreased markedly in nodules of lentil whereas remained unchanged in grass pea. Excessive moisture stress may affect the grain yield up to certain extent but the crop manages to maintain its seed size (Gusmao *et al.*, 2012). Grass pea has also been effectively found to exploit the residual moisture left after the rice harvest when broadcasted into standing rice crops (Joshi *et al.*, 1997; Bharati, 1986). It is well adapted to arid conditions and is one of the hardiest pulses known till date. Its ability to thrive in adverse condition when majority of the crops fail to survive has enabled it to hold the tag of insurance crop. *Lathyrus* species present high genetic variability. The above mentioned characteristics of the crop make it one of the most promising crops for the arid areas and potential targets for further germplasm improvement. But in spite of possessing such agronomical valuable traits it has remained underused and neglected species and very limited research has been devoted to this crop. Physiological studies that could aid our understanding the mechanisms and traits resulting in drought resistance are scarce and also not well understood. Hence this study was undertaken to unravel the genes and its expression along with their metabolic pathway that assist the crop to withstand moisture stress condition. Also because of the complexity of the genome, biotechnological investments remain limited. Genes responsible for the plant's remarkable environmental tolerance are unknown (Yan *et al.,* 2006). The biotechnological potential of grass pea as a source of stress tolerance genes for general crop improvement remains to be exploited. Understanding the molecular mechanisms in the drought response in grass pea is therefore important for improvement of drought tolerance in other crops using molecular techniques. Hence

effort has been made in this present study, to provide an insight into moisture stress related gene activity which may accelerate knowledge based breeding.

Transcriptome analysis is expected to provide an insight into the gene expression of an organism. The transcriptome can be defined as the complete set of transcripts present in a cell, and their quantity, for a particular developmental stage or physiological condition. Various technologies have been developed to deduce and quantify the transcriptome, including hybridization-or sequence-based approaches. However the recent development of new high-throughput DNA sequencing methods has provided a novel method for quantifying transcriptomes that would give us an image of the gene expression level. This method is termed as RNA-Sequencing (RNA-Seq) and imparts clear advantages over the existing approaches. RNA-Seq uses deep sequencing technologies. The reads generated may vary typically from 30-400 bp, depending on the kind of sequencing technology used. Any DNA sequencing technology like Illumina IG, Applied Biosystems SOLiD and Roche 454 Life Science systems, can be used for RNA-Seq. We performed trancriptome analysis of gene expression between cultivars grown under two different conditions, controlled and drought stressed, using SOLiD next-generation sequencing. Since we are dealing with the data generated from SOLiD technology, hence it is important to first discuss and understand its advantages and disadvantages with respect to other popular sequencing technology before proceeding for its analysis.

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a commercial second generation next-generation sequencing (NGS) platform that is based on the principle on sequencing by ligation and di-base encoding. Initially SOLiD could produce more sequencing data than Illumina. But the advancement made in Illumina sequencing technology enabled it to produce more sequencing data and has now superseded SOLiD. Short read lengths and difficulty in sequencing the palindromic sequence are other limitations associated with SOLiD. In spite of these shortcomings the accuracy of this technology could be as high as 99.99%.

Transcriptome analysis can be carried out by two approaches: reference-genome based method and *de novo* method. When a high-quality reference genome sequence is available, for example of model organisms like mouse, fruit fly, *Arabidopsis*

thaliana, the assembly methods in which reads are first mapped against reference genome and then use the aligned reads to infer transcript structure are the most accurate approach. If the well annotated genome of the organism is available then choosing the reference based approach would be regarded best. If the genome of a particular organism is not present then the genome of closest related species is preferred as a reference. Unfortunately, the use of a reference genome is not always possible.

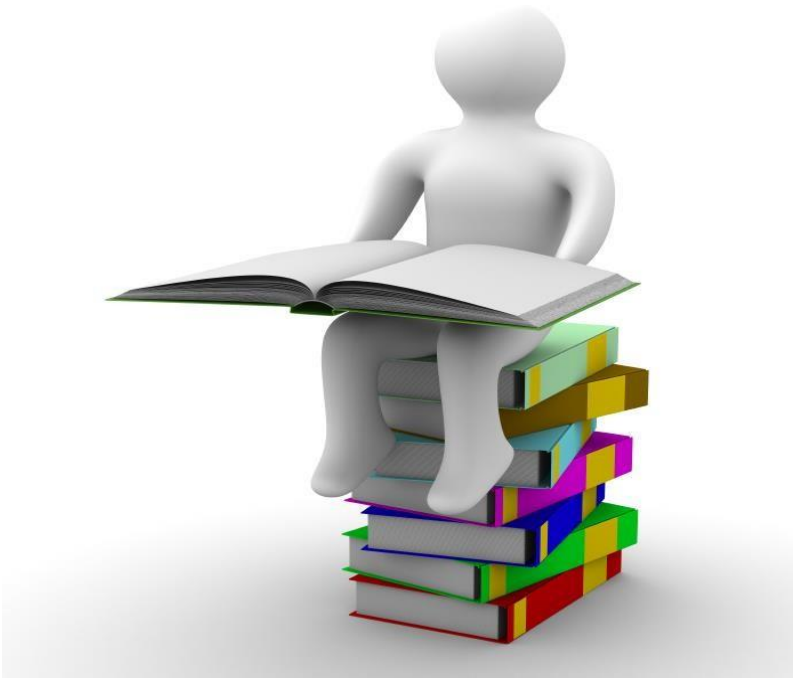
De novo assembly can be described as connecting reads into transcripts that involves identifying overlaps between the reads based on local similarities, establishing their correct order, and orientation, and then connecting all reads such that it satisfies these relationships. This is the most accepted approach in case of absence of reference-genome which is a well annotated sequenced genome. Transcript assemblies can be created by *de novo* genome assemblers, from the RNA-Seq reads in the absence of a reference genome (Birol *et al.*, 2009; Collins *et al.*, 2008; Jackson *et al.*, 2009). The number of *de novo* transcriptome programs developed for assembly of short sequence reads has increased within the past few years. Trans-Abyss (Robertson *et al.*, 2010), Trinity (Grabherr *et al.*, 2011), Oases (Schulz *et al.*, 2012) and SOAPdenovo-Trans (Xie *et al.*, 2014) are the popular examples of de novo assembler.

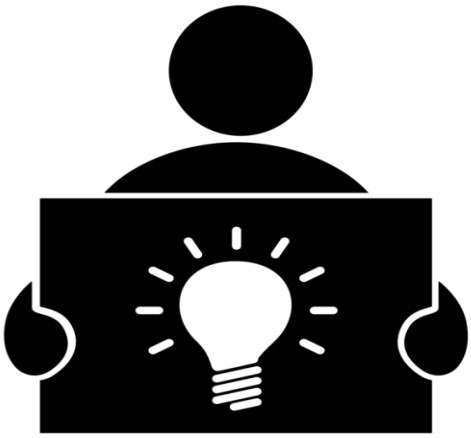
Furthermore, unlike whole-genome sequencing data which has almost constant read coverage along the genome are well explained by a Poisson distribution from statistical point of view, whereas in RNA-Seq data each gene may have a different expression level, which means some genes may be represented by thousands of reads and some by just a few reads. Thus, every gene essentially poses a different transcript assembly problem, where the goal is to assemble all expressed isoforms, and then count the reads deriving from each isoform. Transcript assembly programs must also be computationally efficient to process the vast amounts of data in an RNA-Seq experiment. Several genome-guided transcript assembly algorithms have emerged over the past few years that address all of these challenges, though in different ways.

Together with the growing popularity of RNA-Seq, a number of data analysis methods and pipelines have already been developed for this task. There are common assumptions that substantial gains occur in the quality of the results as read length

increases and when paired-ends (PE) are used. The current read length that is standard for many experiments is PE 100 bp reads (Chhangawala *et al.,* 2015). Currently, however, there is no clear consensus about the best practices for SOLiD short read single end data, which makes the choice of an appropriate method a daunting task especially for a basic user. Hence a comparative study of different RNA-Seq analysis software will be made with the aim to understand and assist the choice of selection of such methods for SOLiD transcriptome data. The study presented here compares and evaluates different tools for both, de novo as well as reference genome based method using *Glycine max* SOLiD transcriptome data. To date, this is the first study to compare the performances of the commonly used *de Bruijn* graph-based *de novo* assemblers and also reference-genome guided assembler and alignment tool.

With this aim we carried out the present study with two major objectives. The first objective was to compare different tools which are suitable for SOLiD trasncriptome data analysis in order to evaluate their performance in respect of *de novo* and reference genome based transcriptome analysis by using *Glycine max* RNA-Seq data. This objective would serve as a reference to carry out further analysis for the identification of differentially expressed genes (DEGs) in our plant of interest i.e., *Lathyrus sativus*. This will frame our second objective in this study. And to the best of our knowledge; this would be the first report to provide molecular insights into the drought tolerance of *Lathyrus sativus.*





Review of Literature

CHAPTER II

REVIEW OF LITERATURE

2.1 Evolution of sequencing technology

Concept of sequencing DNA was first given by Sanger in the year 1975 who later published a method for determining DNA sequences which is known as chain-termination method. Two years later in 1977, another two scientists, Allan Maxam and Walter Gilbert, developed a technique based on chemical-degradation method to sequence DNA in which the labeled DNA was cleaved at specific base and then the fragmented DNA was separated with the help of gel electrophoresis. GS 20 was the first commercially available next-generation sequencing (NGS) technology developed by Roche 454 Life Sciences in 2000 which was founded by Jonathan Rothberg but it was not introduced until 2005. This led to the birth of high throughput-next generation technology (HT-NGS). The principle behind HT-NGS is that the large numbers of DNA molecules can be sequenced simultaneously in a flow cell (Mardis, 2008). Second generation HT-NGS platforms were advance than the first HT-NGS in the sense that it could produce about several hundred millions bases of raw reads like in Roche to billions of bases in a single run as in Illumina and SOLiD (Sequencing by Oligonucleotide Ligation and Detection). SOLiD technology was developed by George Church in the year 2005. The later work of its advancement was carried out by Applied Biosystems in 2007 (Voelkerding *et al.,* 2009). Presently, the above mentioned three technologies are leading second generation HT-NGS platforms. This was based on the principle of emulsion Polymerase Chain Reaction (PCR) of DNA fragments. In spite of its several advantages, PCR amplification may result in changing the relative abundance of many DNA pieces that was present before the amplification. So the use of single DNA molecule without amplification to determine the DNA base sequence would help to overcome the above mentioned limitation. This solution formed the basis of third generation HT-NGS which follows the concept of sequencing-by-synthesis. Braslavsky and his co-workers introduced first technique based on the principle which was licensed by Helicos biosciences. But these technologies do not comprise the ultimate sequencers. The future sequencing technology would be based on single DNA molecule without the

need of amplification, that could generate reads longer than mega bases with no GC bias and high read accuracy. It should be cheap and its operation should be easy. The Oxford nano-pore sequencing technology is an emerging sequencing platform and it is expected to be comprised of above characteristics of ultimate sequencer. But the field of sequencing technology is still under development and continuous improvement.

2.2 Transcriptome analysis

Transcriptome can be defined as a complete set of all RNA or transcript molecules including coding RNA like mRNA and non-coding RNA like tRNA, rRNA, etc present in a cell at a particular time and a particular developmental stage. Functional genomics deal with the understanding of the function of the genes and there alterations at certain conditions like stress, disease, etc. Hybridization and sequence based approaches were the early methods that were used to analyze the transcriptome of an organism. The former method was dependant on the use of fluorescently labeled cDNA that was hybridized on a microarray chip. But its reliance on existing genome added a limitation to its use (Okoniewski and Miller, 2006). On the other hand, sequence based method was free from any such cross-hybridization and could determine the cDNA directly. Serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995), cap analysis of gene expression (CAGE) (Kodzius *et al.*, 2006) or massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000) are the example of sequence-based approach. All these methods depend upon the traditional Sanger sequencing to determine the cDNA sequence making it slower than the recent NGS technology.

2.3 RNA-Seq

HT-NGS also has application in the study of RNAs and in analysis of the transcriptome. RNA-Seq make use of NGS technology to precisely determine the expression level of several genes in particular time and in particular cell by sequencing RNA. It is also famous as whole transcriptome shotgun sequencing (Lister *et al.*, 2009). Till now RNA-Seq technology has seeked application in lots of research studies like the one dealing with detection of alternative splicing (Gan *et al.*, 2010), detection of gene fusion(Maher *et al.*, 2009), identification of splice junctions, identification of novel transcripts

(Robertson *et al.,* 2010) and gene expression quantification (Mortazavi *et al.,* 2008). Due to its vast applicability it has been very truly referred to as a "revolutionary tool for transcriptomics". This technology can be applied with various NGS platforms (Wang *et al.,* 2009).The three most commonly used NGS platforms for RNA sequencing are SOLiD and Ion Torrent both of which are distributed by Life Technologies, and Illumina's HiSeq. They sequence several millions cDNA fragments in each run. Sequencing by synthesis (SBS) forms the underlying principle of Illumina and Ion Torrent in which the addition of nucleotide is detected simultaneously at various fixed positions on a flow cell. Hiremath *et al.* (2011) used NGS technologies such as Roche/454 and Illumina to determine the sequence of most gene transcripts and to identify drought-responsive genes in chickpea. 1,03,215 tentative unique sequences and 21,491 ESTs were produced and about 3000 gene-based markers were developed.

Depending upon the sequencing platform the number and length of the reads may vary. RNA-Seq starts with the sequencing of fragmented RNA samples, which is achieved by first converting it into cDNA and then it is sequenced by a NGS platform as reads. A typical workflow for the identification of differential expression involves producing several millions of short to long reads, assembling, mapping, quantifying and applying test for differential expression and then deducing conclusion based on the output. The reads produced are then *de novo* assembled into transcripts in the absence of reference genome or if the there is availability of reference genome then the reads are mapped on the genome and then the mapped reads are assembled into transcripts. Almeida *et al*. (2014) generated comprehensive transcriptome assemblies from control and *Uromyces pisi* inoculated leafs of a susceptible and rust-resistant grass pea genotype by RNA-Seq and 134,914 contigs were used to analyze their differential expression in response to rust infection. They found considerable differences in regulation of major phytohormone signalling pathways. Salicylic and Abscisic Acid pathways were up-regulated in the resistant genotype where as Jasmonate and Ethylene pathways were down-regulated in the susceptible one.

2.4 Mapping

There are several read aligners which are publicly available in their respective websites. The choice of aligner not only depends upon the sequencing platform but also on the purpose of study. Two widely used methods for aligning reads are based upon either hash look-up table (hash-table) algorithms or burrow-wheeler transformation (BWT). Hash-table aligners are designed to inspect complex differences between the read sequences and the reference, for example GSNAP (Wu and Nacu, 2010) and SOAP (Li *et al.*, 2008), whereas BWT-based aligners for example Bowtie (Langmead *et al.*, 2009), BWA (Li and Durbin, 2009), and SOAP2 (Li *et al.*, 2009) performs efficiently when reads are mapped to the closely related sequences (Oshlack *et al.*, 2010). In case there is no closely related genome sequence then reads are assembled into longer transcripts and this is achieved by the use of de novo assemblers for example ABySS (Birol *et al.*, 2009), SOAPdenovo (Li and Durbin, 2009), Trinity (Grabherr *et al.*, 2011), Velvet (Zerbino and Birney, 2008) and Oases (Schulz *et al.*, 2012). These assemblers are based on the de Bruijn graphs.

2.5 Differential expression analysis (DEA)

One of the key aims of transcriptome analysis is to analyze the change in expression level of the genes between two or more sample collected from two different conditions. Expression levels can be quantified based on the number of reads aligned or mapped to the consensus transcriptome sequence which was obtained after the reads were assembled or to the reference genome. Tools like DESeq(), EdgeR() are used for calculating differential expression of genes which are based upon model which follows known probability distributions, such as Binomial, Poisson, and Negative Binomial distribution. Seyednasrollah *et al.* (2013) performed a systematic comparison of eight widely used software packages and pipelines for detecting the differential expression between sample groups and provided general guidelines for choosing a robust pipeline.

2.6 Annotation and pathway analysis

The ultimate purpose of RNA-Seq analysis is to gain a biological insight into the molecular mechanism of the concerned trait. The transcripts formed after the assembly or genes which are found to be differentially expressed can be annotated by using BLASTprogram to get their significant matches, the biological function of which is already known and to associate the particular query sequence with that related function of the genes annotated previously. Blast2GO is a high-quality functional annotation tool published in 2005 which was developed to serve this purpose (Conesa *et al.*, 2005). It provides information related to biological process, cellular components, and molecular functions of the uncharacterized gene products.The functional information are represented through three gene ontological (GO) terms namely, biological process, cellular component and molecular function.The biological process ontology puts several molecular functions in biological contexts, cellular component ontology describes the location where the gene product is functional, and the molecular function ontology explains the activities of the gene products (Ashburner *et al.*, 2000).Other information related to metabolic pathways can also be associated with this annotation. This can be achieved using KEGG (Kyoto Encyclopedia of Genes and Genomes) which is in-built in Blast2GO that enables the visualization of the metabolic pathways within the transcriptome (Gotz *et al.,* 2008).

2.7 Moisture stress and drought mechanism in plants

Tyagi *et al*. (1998) showed that the ABA responsive genes such as PLE 25, TAS 14 and RAB 17 are synthesized as response to water stress in *Lathyrus sativus,* the level of which decline with the increase in water stress. They also showed that the accumulation of proline was highest in leaves followed by stem and root during moisture stress condition. Talukdar (2013) conducted a study to ascertain the response of two legume crops, lentil and grass pea under different water stress regimes and found that plant growth traits and seed yields components reduced significantly in both the crops but the effect was more severe in lentil compared with grass pea. Proline level increased significantly in both crops, but it decreased markedly in nodules of lentil whereas remained unchanged in grass pea.

Materials and Methods

CHAPTER III

MATERIALS AND METHODS

3.1 Comparison of RNA-Seq analysis tools using *Glycine max* transcriptome data

The performance of the RNA-Seq analysistool has been reported to be driven by the type of data set which differs from one organism to another from which the data is derived. Hence it is difficult to say which tool or software is best to use for such bioinformatics analysis. And till now, no practice or tool has beenreported to be the best. So in order to carry out RNA-Seq analysis with the *Lathyrus sativus* (Grass pea), we first attempted to select an optimum performing tool for our given plant of interest. The genome of *Lathyrus sativus* has not been sequenced yet and hence its genome is unavailable. Therefore, the better alternative is to use similar data of the closely related species which also have a very well annotated genome. *Glycine max* shares a close relationship with *Lathyrus sativus* and also has a well-annotated genome. Besides, similar transcriptome data as that of our i.e., the one sequenced from ABI SOLiD System, it is also available online. The read length (50 bp) of both the transcriptome datais same. Hence, *Glycine max* SOLiD transcriptome data was used to evaluate and compare between commercialized and open source RNA-Seq analysis tool for both the *de novo* and reference genome based approach of RNA-Seq analysis. The sample was extracted from inflorescences pre-meiotic stage, 45-day soil-grown plants. RNA-Seq data of *Glycine max* was downloaded from the SRA database of NCBI. Sample description is given below in Table 3.1.

Table 3.1 Sample description of *Glycine max*

Parameters	Sample
Accession number	SRX487294
Run	SRR1190184
Link to Biosample	SAMN02688415
Project URLs	https://www.ncbi.nlm.nih.gov/sra/SRX487294[accn]
Layout	Single
Read length	50
File type	Conventional base call

Platform/encoding	ABI SOLiD System 4
Selection	PolyA
Strategy	RNA-Seq
Submitted by	Fudan University
Description	soybean IBM

There are several tools available for transcriptome assembly but till now none study that suggests which software package does the best assembly. Hence, the same data of Glycine max as mentioned above was used to compare two tools, one which is commercialized and the other which is open source software. CLC Genomics Workbench is an integrated software which was developed to perform analysis and visualization of NGS data. In this study, we made an attempt to compare CLC Genomics Workbench with Velvet/Oases for *de novo* assembly and with TopHat/Cufflinks for reference genome based approach.

3.1.1 Hardware and software requirement and installation

Computational analyses were carried out with a RHEL 7.0 server with dual CPU and 512 GB memory with 2 TB high-speed storage (SSD) which meets the entire hardware requirement.

CLCGenomics Workbench

CLC being a commercial integrated software tool, its license was first borrowed from NABG Linux cluster server (HPC) on the weekly basis in client server model. Prior to installation, we need to make sure that the system is updated in order to maintain the compatibility with the latest version of the tool. In this study, CLC Genomics Workbench software (Linux version 9.5, CLC Bio, Denmark) was used.

Velvet/Oases

Velvet (Zerbino *et al*., 2008) is a de Bruijn graph-based sequence assembly tool for short reads and it typically works in two steps: hashing and graph building. These steps require two Velvet executables, velveth and velvetg respectively. Velveth reads sequence files and forms all possible combination of words of length k, where k is ak-mer size and this parameter is provided by the userwhich defines exact local

alignments between the reads. Velvetg then reads these alignments, sbuilds a *de Bruijn* graph based on velveth command, removes errors and finally simplify the graph and resolve repeats on the basis of the parameters defined by the user, performs the job of assembly to yield contig sequences as an output along with various statistics which is discussed later in this thesis.

The archive of Velvet assembler was downloaded from the website https://github.com/dzerbino/velvet/archive/master.zip. The downloaded compressed folder was unzipped. In the terminal, the path was set to the velvet directory and compilation was done with the make command. Compilation results in the creation of two executables velveth and velvetg which was then followed by make install which is sufficient for the basic installation.

Oases is a *de novo* transcriptome assembler thatneed to be used in continuation with thevelvet assembler if the data to be assembled consists of transcriptome. It achieves this by using an array of hash lengths and the efficient merging of multiple assemblies to remove the redundancy (Schulz *et al.,* 2012).Therefore it is essentially required that Velvet is installed prior to Oases when a transcriptome assembly is to be performed. The input for Oases is the final output directory resulted after the Velvet run. The archive of Oases software was downloaded from the site https://github.com/dzerbino/oases. Care must be taken that the compilation and installation parameters for Oases should be same as that of Velvet, violation of which may result intocompilation error.

Tophat/Cufflink

TopHat, available at http://tophat.cbcb.umd.edu/ can be employed only if the reference genome is available. It aligns or maps reads to the genome and discovers transcript splice sites. These alignments can be used in several ways during down stream analysis. *Cufflinks,* available at http://cufflinks.cbcb.umd.edu/ uses this map against the genome to assemble the reads into transcripts. This assembly differs from the *de novo* assembly. These assemblies are then merged together using the *Cuffmerge* utility, which is included with the Cufflinks package. *Cufflinks* is the common name used for both the package and the program as well. This merged assembly serves as consensus sequencessimilar to the *de novo* assembly and forms the basis for calculating gene and transcript expression in each condition. The reads

and the merged assembly are provided as an input to *Cuffdiff*, which calculates expression levels and tests the statistical significance of observed variations. *Cuffdiff* is also a part of the *Cufflinks* package, that takes the aligned reads from two or more conditions and reports genes and transcripts that are differentially expressed using a rigorous statistical analysis. Till now this tool has been used in a number of recent high-resolution transcriptome studies. *CummeRbund*, available at http://compbio.mit.edu/cummeRbund/ is a recently developed powerful tool for plotting which provides functions for creating commonly used expression plots such as volcano, scatter and box plots. *CummeRbund* transforms Cufflinks output files into R objects suitable for analysis with a wide variety of other packages available within the R environment. *CummeRbund* can be accessed through the Bioconductor website http://www.bioconductor.org/. TopHat/Cufflinks is the most popular RNA-Seq analysis tool when the reference genome is available and together it is popularly known as *Tuxedo* protocol.

The linux commands for installing the software are given in the Appendix. These commands were used for the installation of the required software.

3.1.2 Data quality assessment

FastQC (version 0.11.5) tool, available at http://www.bioinformatics.babraham.ac .uk/projects/fastqc/ is the most popular bioinformatics tool for sequence quality visualization, which provide us with the quality score of data based on various parameters like basic statistics, per base sequence quality, sequence length distribution, per sequence quality scores, per sequence GC content, per base sequence content, per base N content, over represented sequences, sequence duplication levels and K-mer content. Quality of the data is then checked based on the value of the parameters whether the score is below or above the quality threshold value. If it passes then the data is approved for further analysis and if it is below the threshold value then we require cleaning the data by removing poor quality sequences from the dataset. FastQC was employed in this study for visualizing the quality of the sequence. After visualizing, the low quality reads whose phred score was less than 20 was removed. Since FastQC strictly require the input file to be in fastq format we converted the file downloaded in SRA format to fastq format with the help of SRA tookit, available at www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft. Fastq

format store sequence base information along with its corresponding quality score. Quality score or phred score (Q) was calculated by the given formula:

$$\mathbf{Q = -10 \log_{10} P}$$

Where, 'Q' denotes phred score and 'P' is the probability that the base is called wrong (error probability).

3.1.3 Transcriptome assembly

Sequence assembly involved alignment of short reads into longer stretch of DNA sequences in order to reconstruct the original sequence. The genome assembly algorithm employed considered all the small pieces of sequence-reads, aligned them, and detected overlaps. The process was iterated for overlapping reads to further merge. Assembly is a computationally intensive job. Two approaches were used for genome assembly:

i) *De novo* assembly (for un-sequenced genomes)
ii) Reference-guided assembly (for sequenced genomes)

3.1.3.1 *De novo* assembly

CLC bio

The reads were imported in fastq format and assembly was executed by following the default parameters. With the help of read information, contigs were created based on the *de Bruijn* algorithm. Reads werethen mapped onto thecontigs.

Velvet/Oases

Assembly of the clean reads was carried out by *Velvet*(an efficient genomic *de novo* assembler based on *de Bruijn* graphs) and *Oases* (to heuristically assemble RNA-Seq reads in the absence of a reference genome) assembler. Velvet was used to assemble sample reads at four different *k-mers* (25, 27, 29 and 31). Contigsof four assemblies were merged into a single non-redundant assembly using Oases, which processes the contigs into loci and associated transcripts. K-mer length of 25 was used to merge the contigs.

Command for merging the assembly

```
velveth MergedAssembly/ 25 -long directory*/transcripts.fa
velvetg MergedAssembly/ -read_trkg yes -conserveLong yes
oases MergedAssembly/ -merge
```

3.1.3.2 Reference genome based approach

The most critical step in reference-genome based approach of RNA-Seq analysis is the percentage of high-quality reads that actually mapped to the reference genome. Rest of the downstream analysis like assembly of mapped reads and testing for DEGs crucially depends upon the mapping percentage. Hence, the percentage of mapping was selected as criteria to judge the performance of the alignment tool used in RNA-Seq analysis.

CLC bio

High-quality reads of soybean was imported in fastq format. Then, reference genome of *Glycine max* was imported using Roche 454, fasta importer and was then converted into track. Once the reference track was created, all the genes were extracted from the reference genome using a gene track along with all transcripts using mRNA track.

We had additional information about the chromosome of other organelles like plastid and mitochondria and scaffolds along with the standard 20 chromosome of soybean. We chose to retain all these information in order to maximize the level of annotations. The reads of each sample were mapped to the reference genome annotated with the genes and transcripts. Mapping was performed on both the genic as well as intergenic regions with the default parameters. And the expression values were determined on the basis of RPKM (reads per kilobase of exon model per million mapped reads).

TopHat

Reference-genome based approach of transcriptome analysis was performed using *TopHat*. Genome of Glycine max was downloaded from NCBI in fasta format along with its general feature format (gff) file that was essentially required to create an index file by *Bowtie* that was later usedby the *TopHat* for aligning reads to the

genome. Reads were aligned to the reference genome. Bowtie, available at http://bowtie-bio.sourceforge.net/index.shtml till date is the most efficient alignment program and it serve as an alignment engine for *TopHat*. The aligned reads are then assembled into individual transcripts by *Cufflinks* program.

```
bowtie2-build -f glycine_max.fa glycine_max_index

tophat -p 6 -G glycine_max.gff -o /mapping /indexes/
      glycine_max_index glycine_max.fastq
```

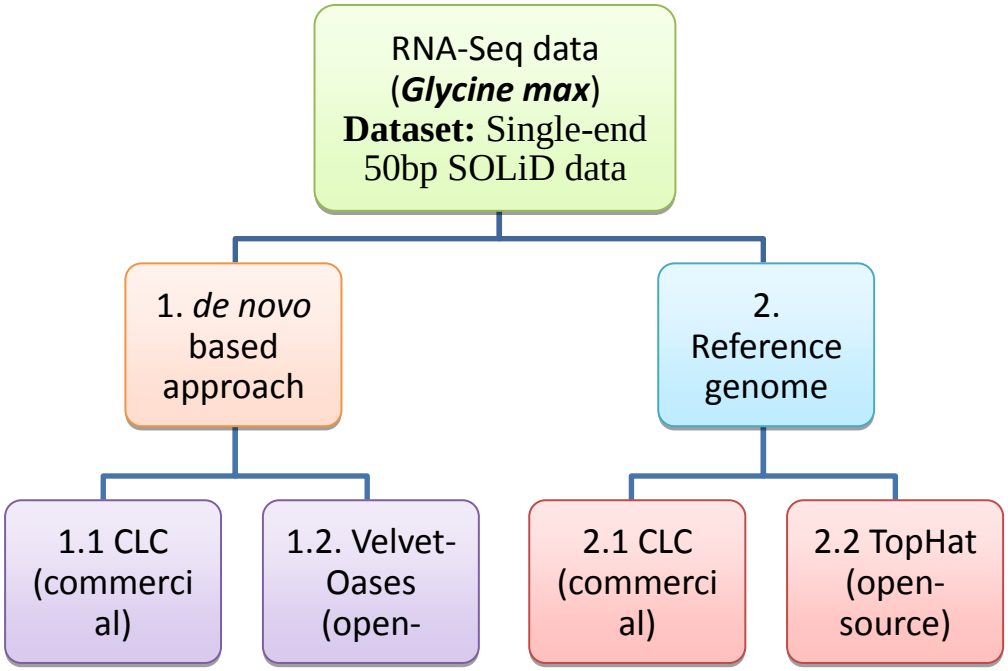


Figure 3.1: Tools used for *de novo* and reference-genome based approach for RNA-Seq analysis

3.2 Transcriptome analysis of *Lathyrus sativus* SOLiD transcriptome data.

3.2.1 Raw data summarization

Transcriptome analysis was performed to analyse moisture stress responsive genes using RNA-seq data of tissues sample from *Lathyrus sativus* under two conditions i.e., control and moisture stress. The quality of the raw reads (50bp) data generated from SOLiD sequencing was assessed using FastQC. The reads were pre-processed to remove the low quality reads Q<20 and low complex regions (reads with

ambiguous base ‘N’). Reads of size more than 25 bp were retained for the assembly of *Lathyrus* transcriptome.

RNA-Seq data of tissues sample from *Lathyrus sativus* under two conditions i.e., control and moisture stress was used for the transcriptome analysis of moisture stress responsive genes. Single end reads were generated from ABI SOLiD 4 System and output was in the form of csfasta + qual format, each of which was 50bp long. Both the sample data is also available online at SRA at National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov//) with project IDs SRR1005702 (control) and SRR1005703 (drought) were used in the study (Table 3.1).

Table 3.2 Sample description of *Lathyrus sativus*

Description	Identification of genes related to water stress tolerance in grasspeas (*Lathyrus sativus*)	
Submitted by	NBPGR, New Delhi	
Strategy	RNA-Seq	
Selection	Random	
Platform/encoding	AB SOLiD System 4	
File type	Conventional base call	
Layout	Single	
Read length	50	
Sample	**Control**	**Drought**
Accession number	SRX362148	SRX404301
Run	SRR1005702	SRR1005703
Link to Biosample	SAMN02364559	SAMN02364559
Project URLs	https://www.ncbi.nlm.nih.gov/sra/SRX362148[accn]	https://www.ncbi.nlm.nih.gov/sra/SRX404301[accn]

3.2.2 Quality assessment and pre-processing of data

Same procedure of quality assessment and pre-processing of data as mentioned in sections 3.1.2 and 3.1.3 was followed in *Lathyrus sativus*.

3.2.3 Transcriptome assembly

Same procedure of assembly as described in section 3.1.4 was followed in *Lathyrus sativus* as in case of *Glycine max* for both the approaches, de novo as well as reference genome based approach.

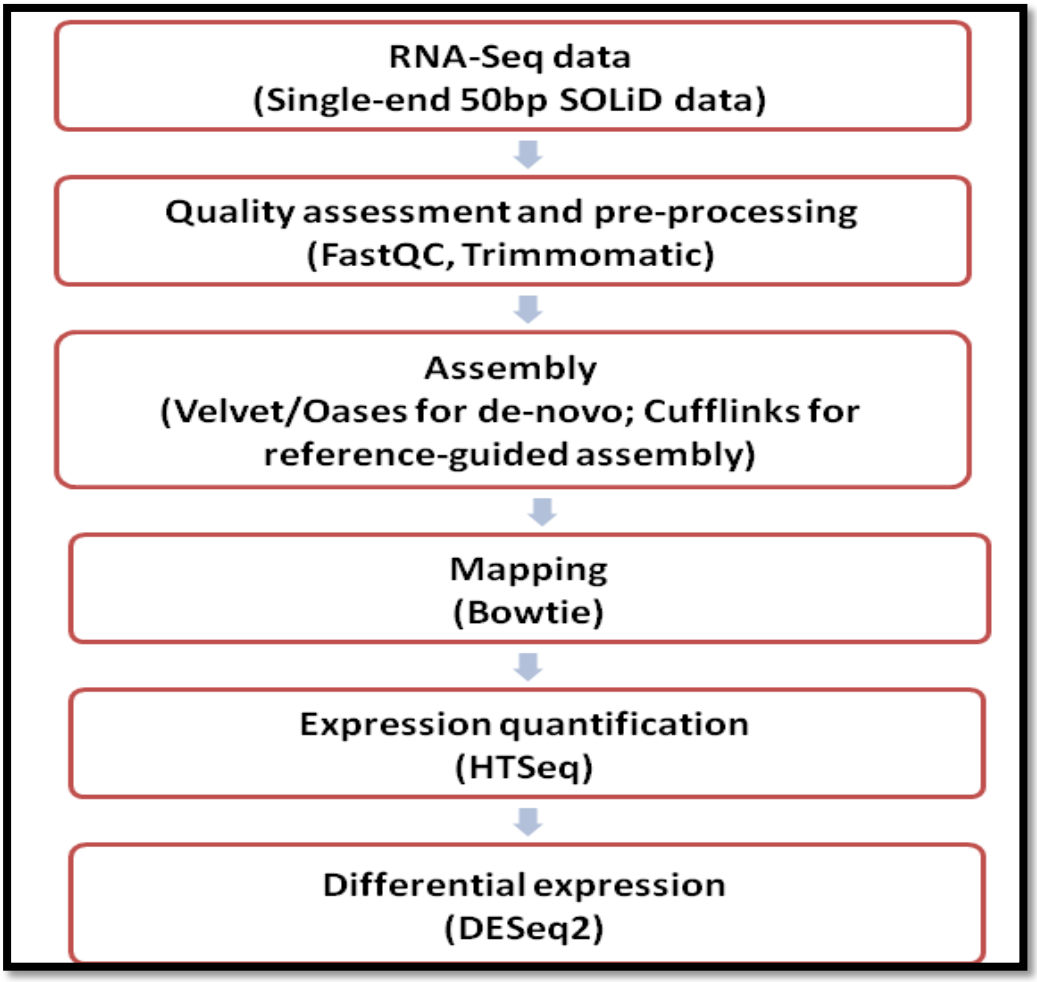


Figure 3.2: Flow diagram for identification of differentially expressed genes

3.2.4 Functional Annotation and classification of assembled transcripts

Annotation is done in order to facilitate the extraction of biological significance of sequence data that helps in better understanding of the biological processes and identification of genes in the given sequence data whose function is known in the related genome. For the functional annotation of all the transcripts obtained after assembly, the workflow depicted in Figure 3.2 was employed. The assembled

trasncripts were subjected to BLASTx algorithm (Altschul *et al.,* 1990) against NCBI non-redundant (nr)-protein sequence database version nr.38, available at ftp://ftp.ncbi.nlm.nih.giov/blast/db in virdiplantae (plant) kingdom which was locally installed. Annotation of the results were done using Blast2Go Pro version 3.1 software, available at https://www.blast2go.com (Conesa *et al.,* 2005) program in order to obtain GO annotation of the transcripts. KEGG pathway analysis was also performed to analyse the gene product during the metabolism process and related gene function in cellular process. The genes are grouped into molecular function depending upon the molecular activities of the gene or gene products, cellular component describing the location of the gene activity and biological process in which the gene or gene products participate. PlantTFDB v4.0, available at http://planttfdb.cbi.pku.edu.cn/ (Jin *et al.,* 2017) was used for the identification of transcriptional factors.

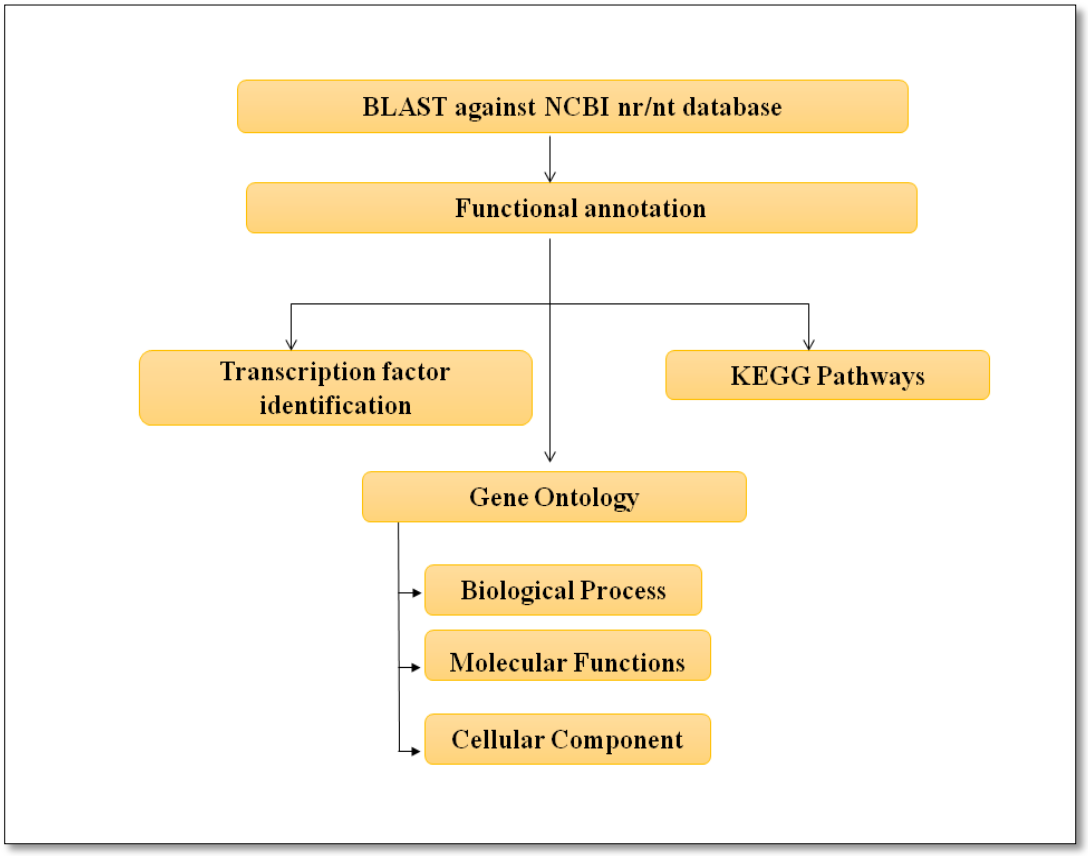


Figure 3.3 Flow chart of functional annotation

The annotation process completes in two steps: mapping and annotation. In mapping, B2G plugin fetches all Gene Ontology (GO) terms associated to hits obtained after Blast search. Once the mapping is completed results can be visualized in the form the

mapping statistics. Mapping results can be summarized on the basis of three evaluation charts: db resources of mapping that shows which databases was used to obtain annotation, evidence code distribution for blast hits and sequences. Mapping results in the generation of pool and in the process of annotation; those GO terms are selected and assigned to the query sequence. Blastx and Blast2GO parameter used are:

e-value <= 10-e6
Similarity >=35%
Annotation cutoff >= 55
GO weight cutoff >= 5

3.2.5 Mapping

Bowtie2 was used to map high quality filtered reads to the assembled transcripts for reads quantification. Input file for bowtie can be in fasta or fastq format. We took fastq file as an input which contains base sequence information of each reads and its quality score at each position. These reads were then aligned to the reference genome of *Glycine max.* Output of alignment is stored in BAM (binary alignment map) file format.

3.2.6Analysis of differential expression of genes (DEGs)

Differential expression analysis is one of theaims of RNA-Seq data analysis. DESeq2 (Love *et al.*, 2014) is one such package that is developed for the identification of differentially expressed genes (DEGs) between control and stress samples. DESeq2 is a R software package used to test differential expression based on the model using negative binomial distribution. It expects raw read count as an input which we obtained using HTSeq. Heat map, MA plot, and volcano plot was also obtained. Transcripts which showed absolute fold-change >2 at FDR corrected p-value <0.05 were identified as differentially expressed.

Command used for DESeq2 in R-3.4.0 is given in Appendix.

Results

CHAPTER IV

RESULT

4.1. Optimization of protocol using *Glycine max* RNA-Seq data

4.1.1 Quality assessment and pre-processing of RNA-Seq data

To evaluate the performance of RNA-Seq analysis tool we used a SOLiD transcriptome data of *Glycine max.* The sample consist of 112,537,370 raw reads, each of length 50bp sequenced from ABI-SOLiD 4 System. After quality assessment by using FastQC, poor quality reads were removed and if ambiguous bases are present more than two, they were also trimmed. Trimming resulted in variable length of the reads. Reads which were 25 bp long or more were retained and rest were discarded.

Table 4.1 Data cleaning statistics

Total raw reads	Total clean reads	Q20 %	GC%
112,537,370	78,326,009	69.6	47

Q20% is the proportion of the base quality value larger than 20 and GC% is the proportion of guanidine and cytosine bases among the total bases. Almost 30% of the reads were discarded which left us with total 78,326,009 clean reads that were further utilized for analysis.

4.1.2 Comparison of *de novo* and reference genome based analysis

This section compares the tools used in two different methods of transcriptome analysis i.e., *de novo* and reference genome based methods.

4.1.2.1 Assembly of RNA-Seq data using *de novo* analysis

For *de novo* assembly, comparison was made between *CLC bio* and *Velvet-Oases* assembler. 78,326,009 reads were de novo assembled to yield contigs. Assembly by CLC was executed by following the default parameters with word size or k-mer of length 25.

While using *Velvet*, reads were assembled at four different k-mers (25, 27, 29 and 31) (Table 4.2). After running this process, the results of all the assemblies were

merged into a single non-reductant assembly using *Oases* which processes the contigs into locus and their associated transcripts. K-mer length of 23, 25, 27 and 29 was used to merge the assemblies (Table 4.3). Merging of assemblies resulted in large and robust contigs.

Table 4.2 Assembly statistics of *Velvet*

S.No.	k-mer	N50	Maximum length	No. of contigs
1.	25	420	1273	30,080
2.	27	426	1258	29,675
3.	29	430	1009	28,453
4.	31	436	918	28,051

Table 4.3 Assembly statistics by *Oases*

S.No.	k-mer	N50	Maximum length	No. of transcripts
1.	23	735	3162	23,897
2.	25	741	2050	24621
3.	27	739	2185	29182
4.	29	735	2183	29205

Out of the four merged assemblies, assembly at k-mer of size 25 yielded transcripts with larger N50 value. Assembly having greater N50 value was regarded as the best possible assembly.

Table 4.4 Comparing results of *CLC* and *Velvet-Oases* assembler

Statistics	CLC	Velvet-Oases
Number of contigs	6,168	32,543
N50	282	741
Maximum contig length	1115	2050
Minimum contiglength	100	100
Average contiglength	565	640

Maximum numbers of contigs were formed when assembly was carried out by *Velvet-Oases* with higher N50 value and average length was also found to be greater than CLC. Since larger N50 is most desirable for the optimum assembly, *Velvet-Oases* was preferred over *CLC bio.*

4.1.2.2 Mapping of RNA-Seq data using reference genome approach

Comparison was made between *CLC bio* and *TopHat* which uses *Bowtie* as its alignment engine. In reference-genome guided method, percentage of reads aligned against reference genome plays a very critical role that determines the further downstream analysis stating from reference-genome guided assembly to differential expression analysis. Higher the alignment percentage, better it is.

Table 4.5 Comparing results of *CLC* and *TopHat*

Statistics	**CLC**	**Tophat**
Total no. of reads	36,557,140	
No. of reads mapped	10, 235,999	17,181,855
Mapping (%)	46	59

47%of the reads were aligned by *TopHat* whereas only 28% of the reads could be aligned by CLC bio (Table 4.5). Hence, *TopHat* was preferred over CLC when reference-genome was available to carry out the further trancriptome analysis.

4.2 Transcriptome analysis of *Lathyrus sativus* RNA-Seq data

4.2.1 Quality assessment

The single-end reads of moisture stressed and control sample of *Lathyrus sativus* were generated using AB SOLiD 4 System. Total of 64,246,685 and 64,640,515 raw reads of control and stress samples were generated respectively with the read length of 50 bp. After pre-processing of these data sets, 27689545 and 25662149 poor quality reads of control and stress sample respectively were removed (Table 4.6). These cleaned reads were used further for *de novo* assembly and reference-genome trancriptome analysis of *Lathyrus sativus*.

Table 4.6 Statistics after quality assessment and pre-processing

S. No.	Samples	Reads before cleaning	Reads after cleaning	Q20 %	GC%
1.	Control (LS_C)	64,246,685	36,557,140	56.9	45.07
2.	Drought(LS_D)	64,640,515	38,978,366	54.64	43.59

4.2.2 Transcriptome assembly

4.2.2.1 *De novo* assembly

The high quality reads were assembled into reference transcriptome using *Velvet-Oases*by following the same protocol which was used to assemble Soybean reads. The assembly resulted in generation of 24621 transcripts with an average length of 524nt and N50 of 428 nt. 7.5% of the total contigs were of length less than 300 nt and 88.9% of the contigs lied between the range of 300 to 600 nt. Only 0.2% of the total contigs exceeded the length of 1200 nt. The largest contig was of length 2652 nt (Figure4.1).

Table 4.7Assembly statistics of *Velvet-Oases*

S.No.	k-mer	N50	Maximum length	No. of loci
1.	25	459	1887	19162
2.	27	451	1885	19182
3.	29	445	1883	19205
4.	31	443	1881	19256

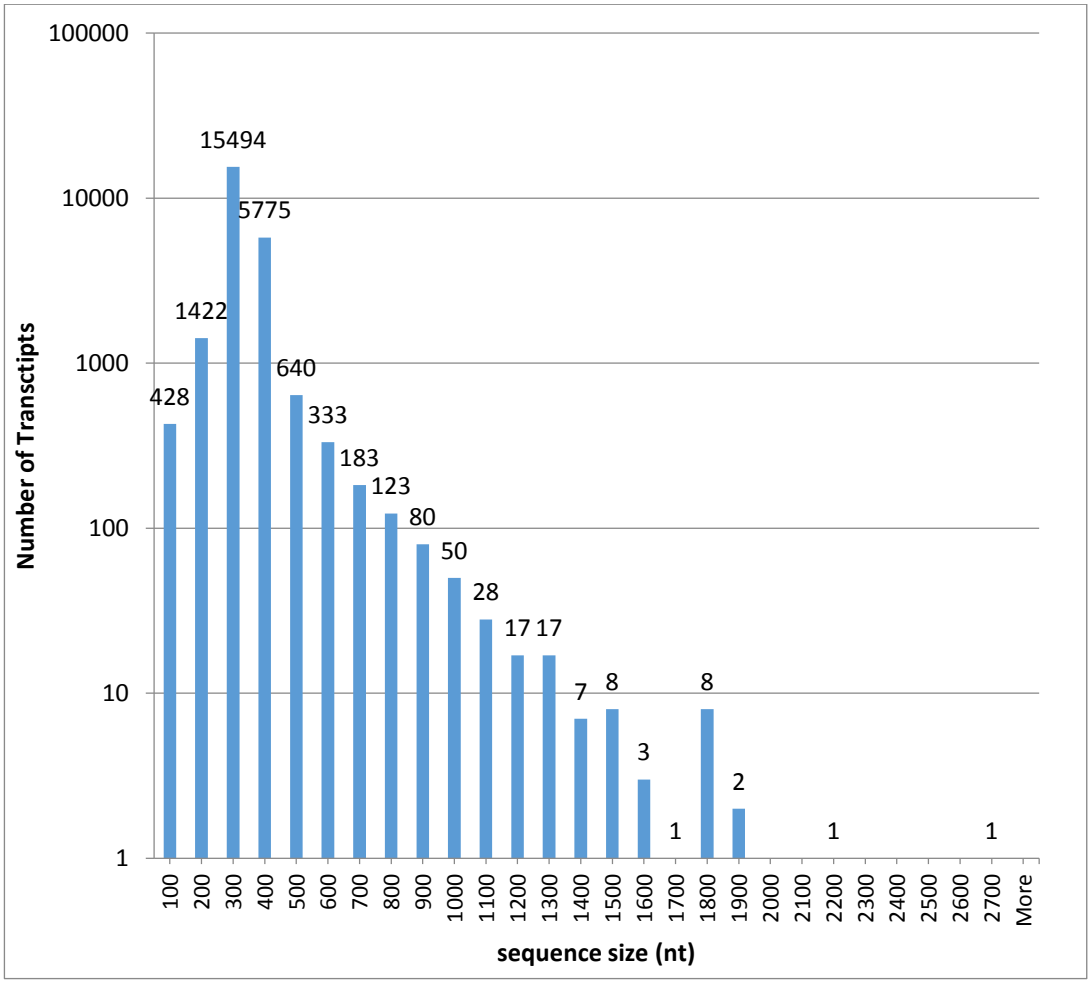


Figure 4.1 Length distribution of *Lathyrus* contigs using *de novo* assembly

4.2.2.2 Assembly by reference-genome based approach.

*TopHat*was used to align reads of the sample of Grass pea against reference-genome of Soybean. The mapping statistics of both the samples are given in Table 4.7.

Table 4.8 Mapping statistics

S.No.	Sample	No. of reads	Mapping (%)
1.	Control	36,557,140	47
2.	Drought	35,332,372	43

The aligned reads mapped against reference genome were individually assembled using *Cufflinks.* Assemblies of both the samples were merged together to form the final assembly consisting of 96,028 contigs using *Cuffmerge*.

4.2.3 Functional annotation and analysis of *Lathyrus sativus* transcriptome data

Contigs resulted from both *de novo* and reference-genome guided assembly were blasted and functionally annotated.Out of 21,621 transcripts which were *de novo* assembled 15,562 (63.21%) showed significant similarity to known proteins in non-redundant (NR) database of NCBI. Blast2GO was employed to identify functional categories of these transcripts. In total, 56,084 GO IDs were assigned to 13,723 transcripts. Maximum number of transcripts showed resemblance with *Glycine max*, followed by *Medicago truncatula* and *Trifolium subterranean*(Figure 4.2).

On the other hand, 93.53% of the total 96,028 reference-genome based assembled transcripts shared the similarity with the nr-database. 4,43,062 GO Ids were assigned to 89,087 transcripts.These IDs are associated to three principal categories: biological process (BP), cellular component (CC) and molecular function (MF).

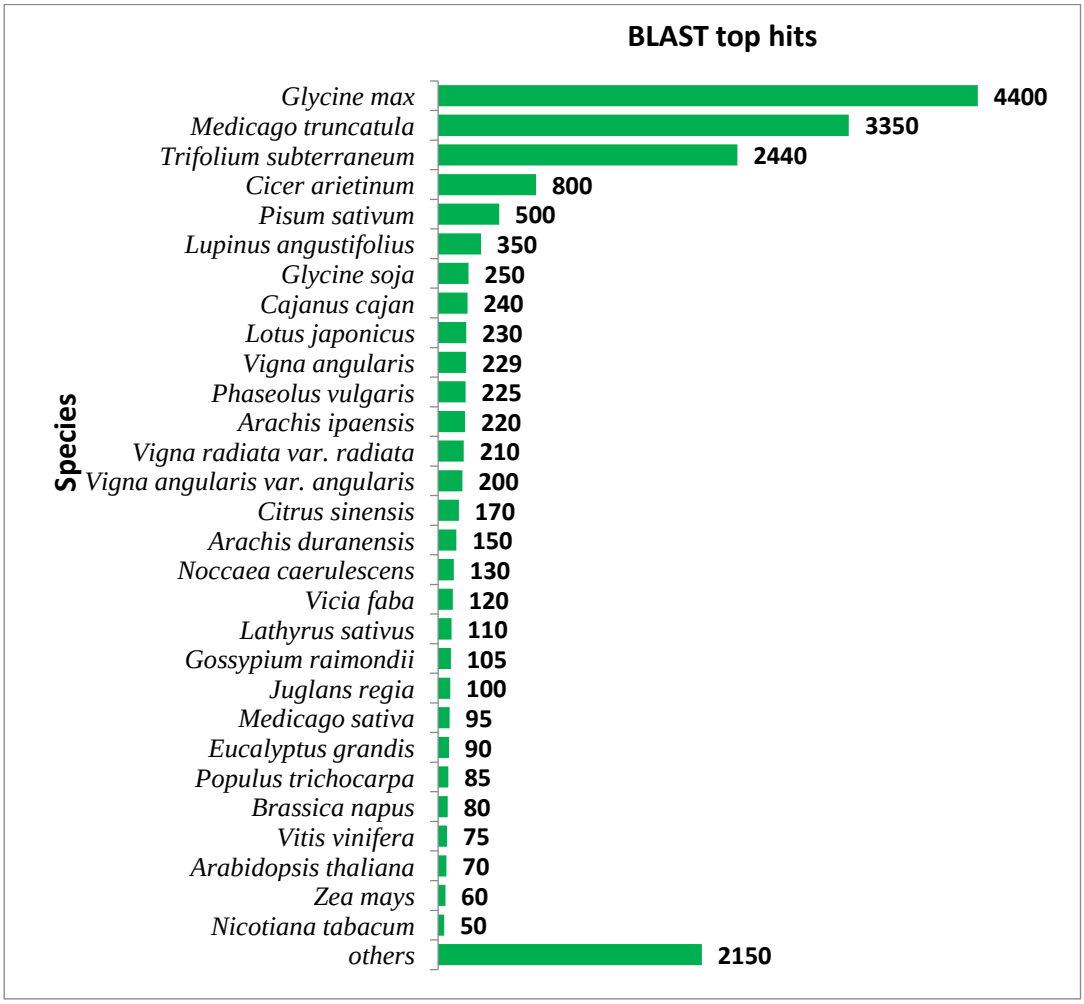


Figure 4.2 Graphical representation of species distribution according to BLAST hits

In *de novo* based approach for the biological process majorityof the transcripts were involved in metabolic process and followed by cellular process and response to stimulus (Figure 4.3a). Similar distribution of gene ontology terms was found in reference-genome based approach. Other biological process included biological regulation, signaling, growth and reproduction (Figure 4.3b).

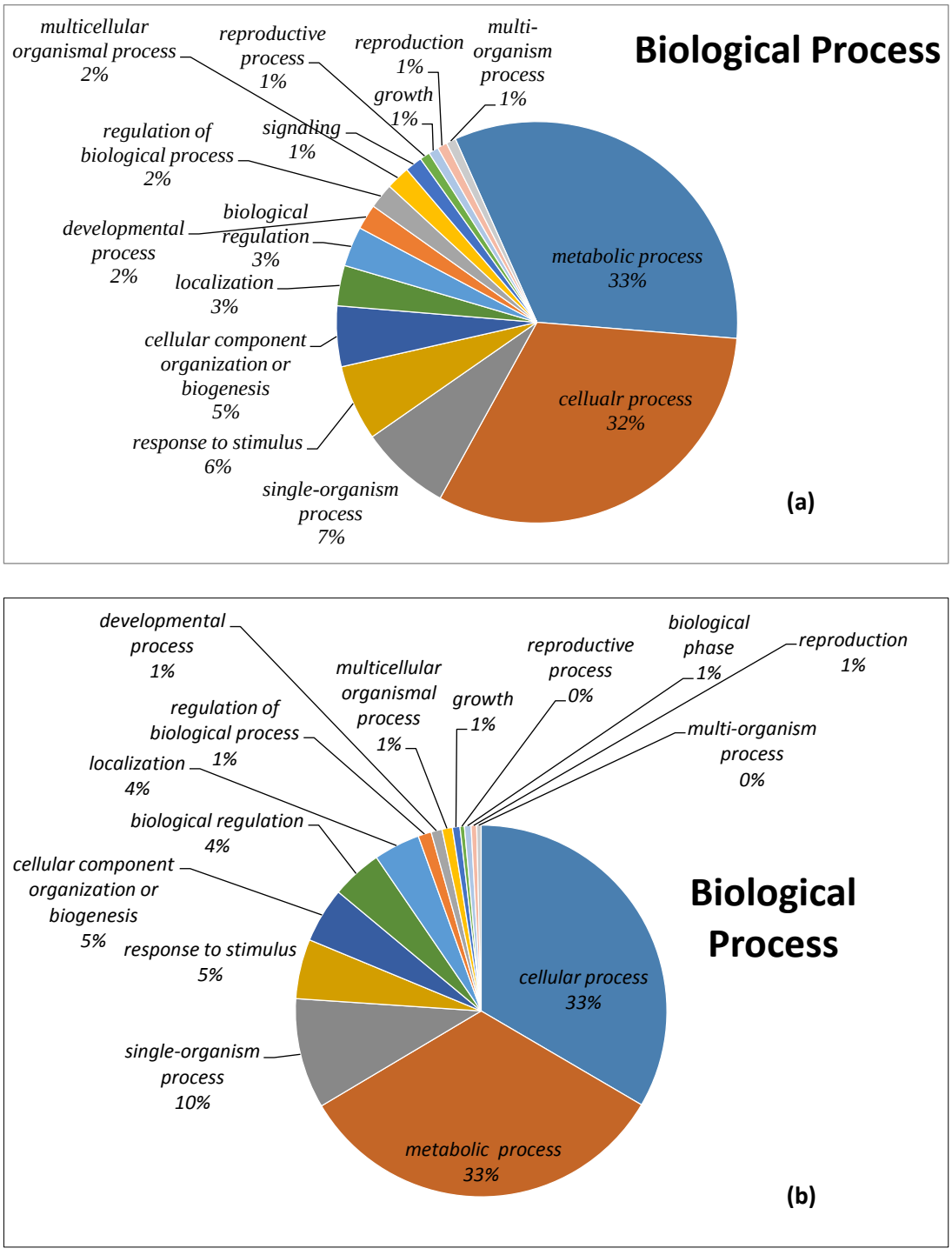


Figure 4.4 Graphical representations of GO terms for biological process in a)*de novo* based approach and b) reference based approach

For the molecular function 43% were related to nucleotide binding and 36% were associated with hydrolase activity. Reasonable number of transcripts were also involved in catalytic activity, transferase and kinase activity (Figure 4.4a).

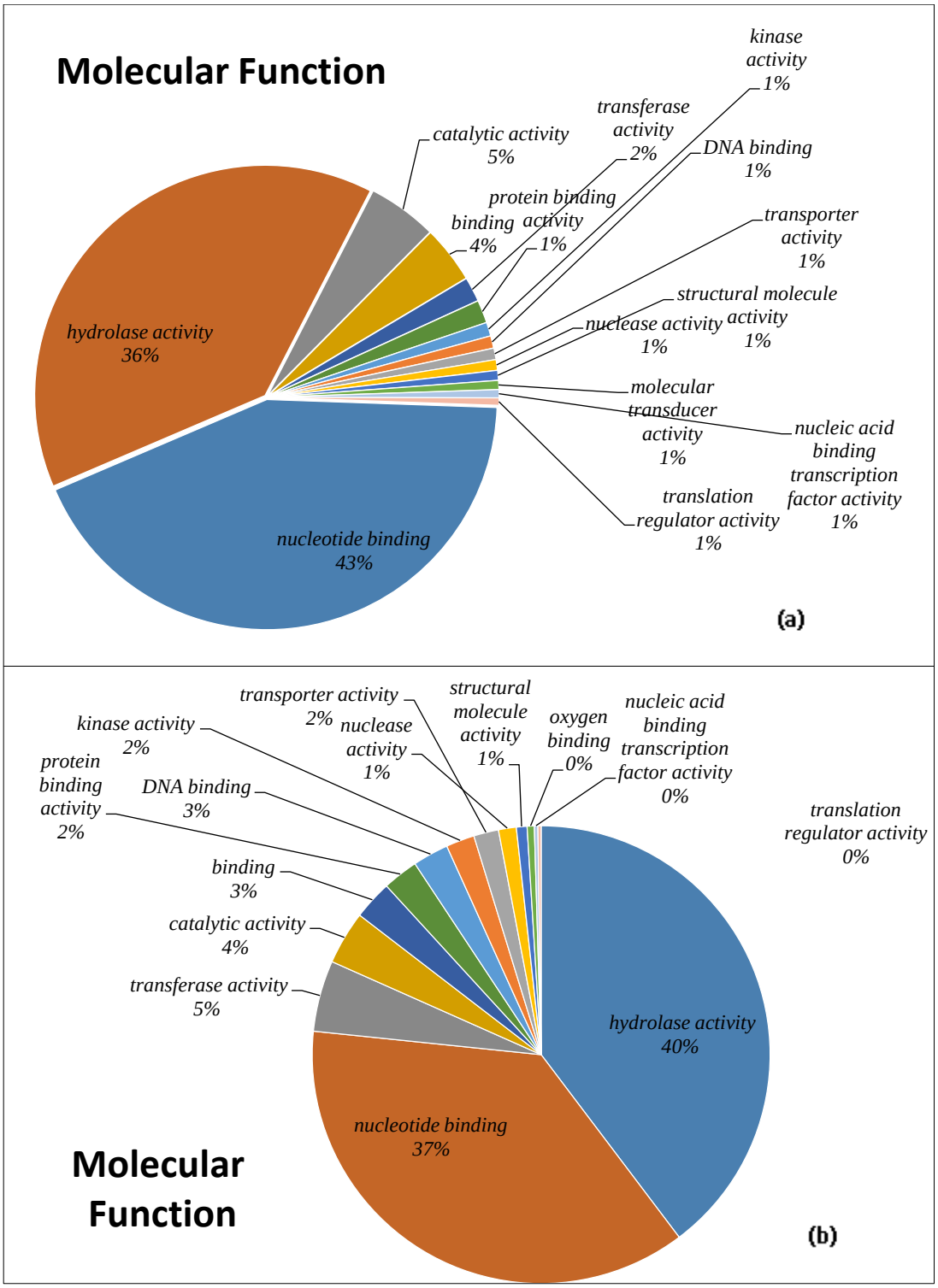


Figure 4.4 Graphical representations of GO terms for molecular function in a) *de novo* based approach and b) reference based approach

In reference-genome based approach, majority of the transcripts were involved in hydrolase activity (40%) followed by nucleotide binding (37%). Other functions found in common with *de novo* based approach were binding activities of protein and DNA, kinase activity, transferase and catalytic activity (Figure 4.4b).

About 36% of transcripts were active organelle and 28% in membrane. Other active cellular components were macro-molecular complex, organelle part, cell wall, nucleolus and endoplasmic reticulum (Figure 4.5a). In reference-genome based approach, four most active cellular components were cell part, organelle, membrane and macromolecular complex (Figure 4.5b).

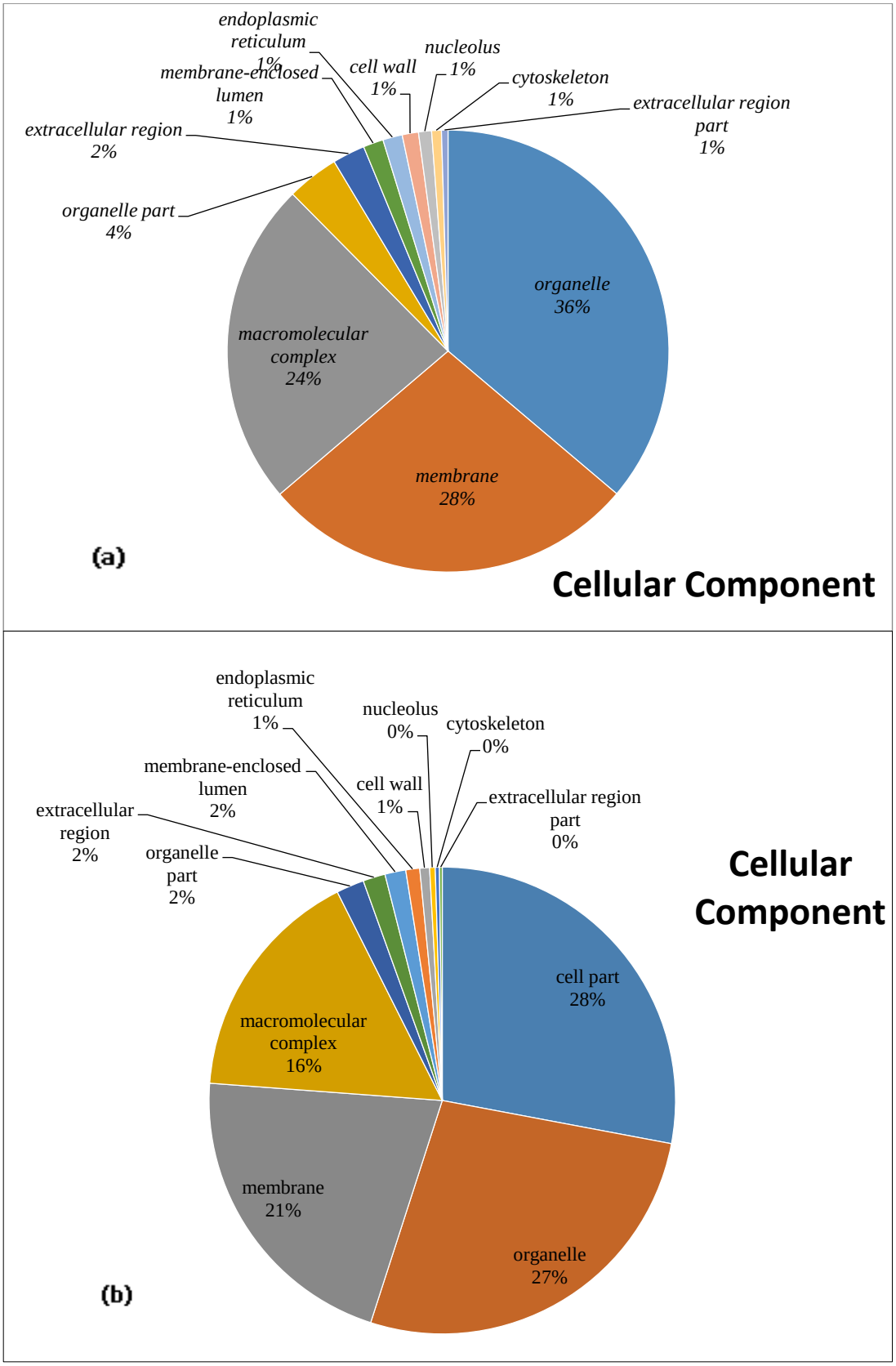


Figure 4.5 Graphical representations of GO terms for cellular component in a) *de novo* based approach and b) reference based approach

Pathway analysis was carried out on KEGG Pathway database. Total of 144 KEGG pathways were identified. The top five pathways were “Biosynthesis of antibiotics”, “Purine metabolism”, “Amino sugar and nucleotide sugar metabolism”, “Cysteine and methionine metabolism” and “Starch and sugar biosynthesis”.

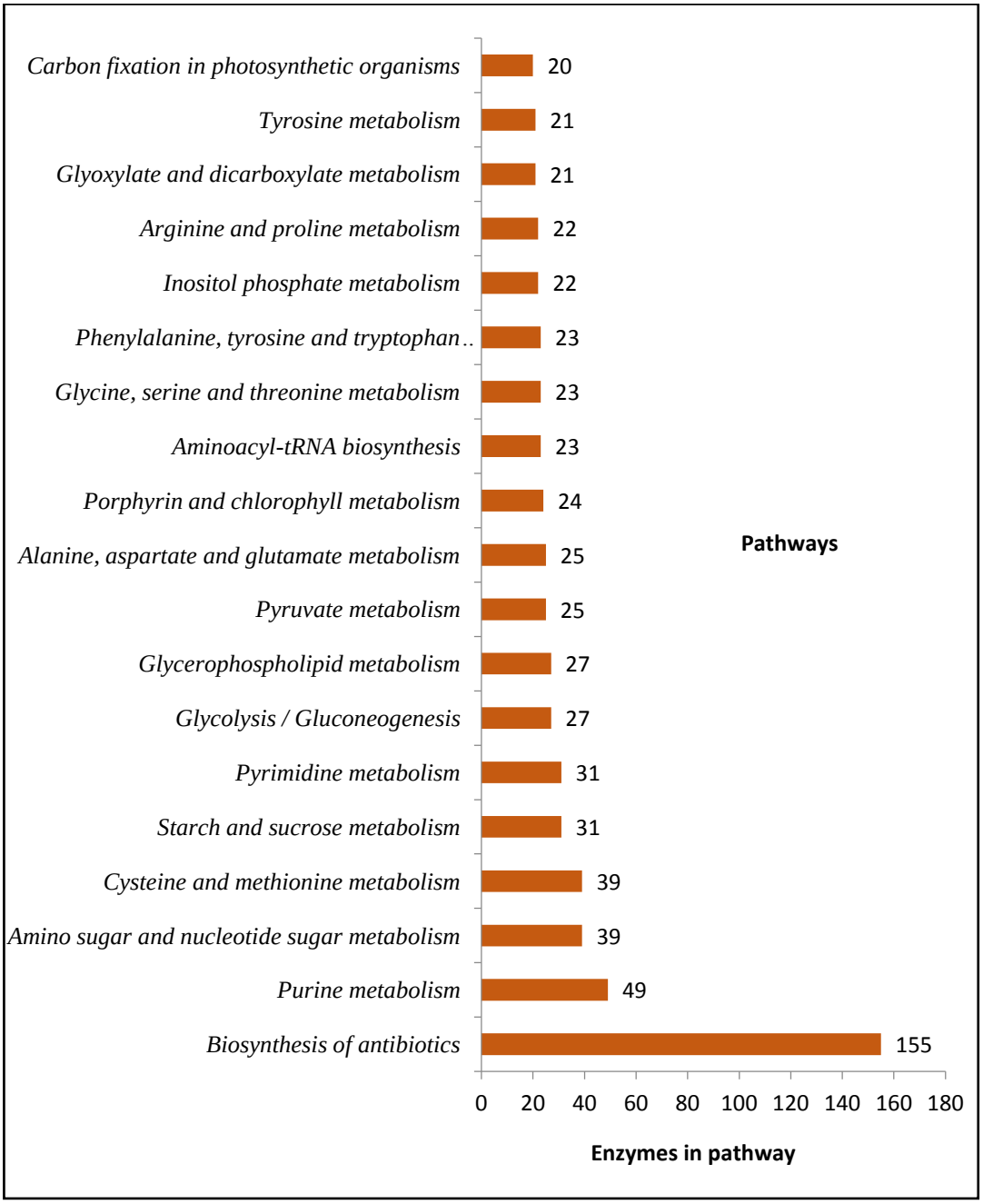


Figure 4.6 Top 20 KEGG pathways identified in *Lathyrus sativus* transcripts (*de novo* based approach)

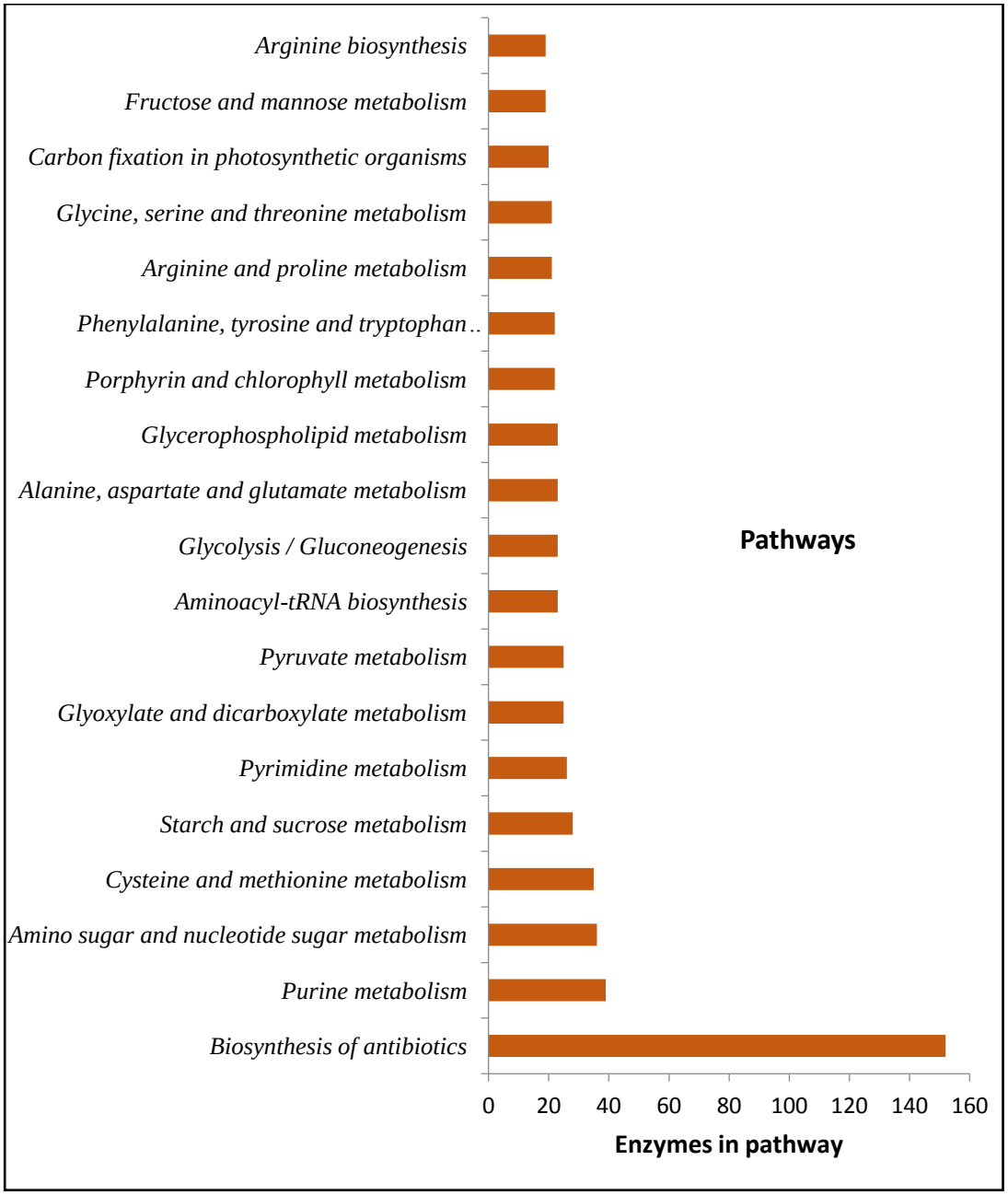


Figure 4.7 Top 20 KEGG pathways identified in *Lathyrus sativus* (reference genome based)

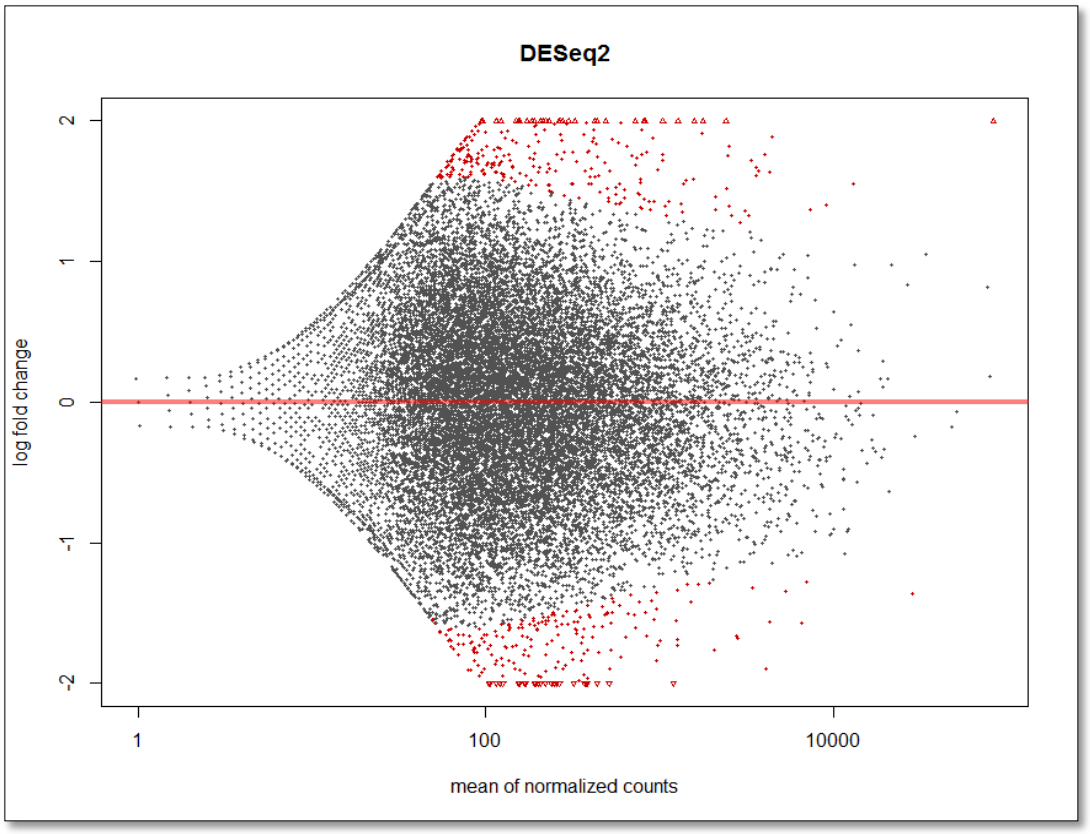
Top 5 metabolic pathways identified in reference-genome based approach was similar to *de novo* based approach. Rest of the pathways include glyoxylate and dicaroxylate metabolism, pyruvate metabolism, glycolysis/gluconeogenesis and aminoacyl-tRNA biosynthesis.

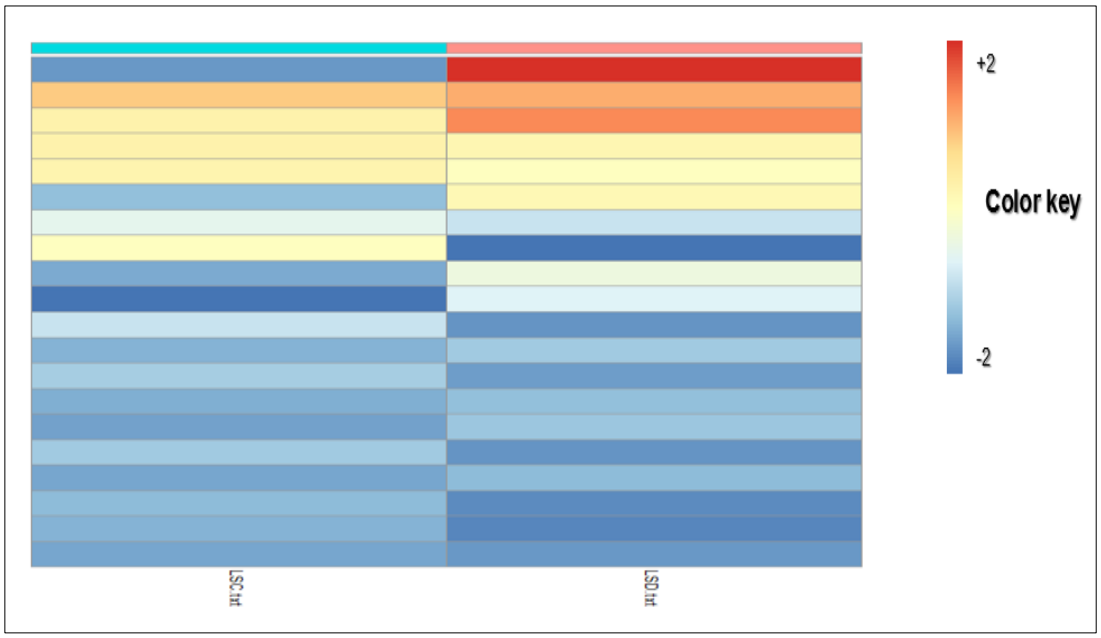
4.3 Analysis of differential expression of assembled transcripts under moisture-stress

57 transcripts were classified as differentially expressed genes (DEGs) in moisture stressed sample when compared to control. Out of DEGs, 32 transcripts were found to be up-regulated whereas, 25 transcripts were found to be down-regulated (Figure 4.8b). Upregulated genes included cysteine protease, ascorbate oxidase and heat shock protein. Down-regulated genes included NAC domain-containing 72-like protein and Zinc Finger 512B.

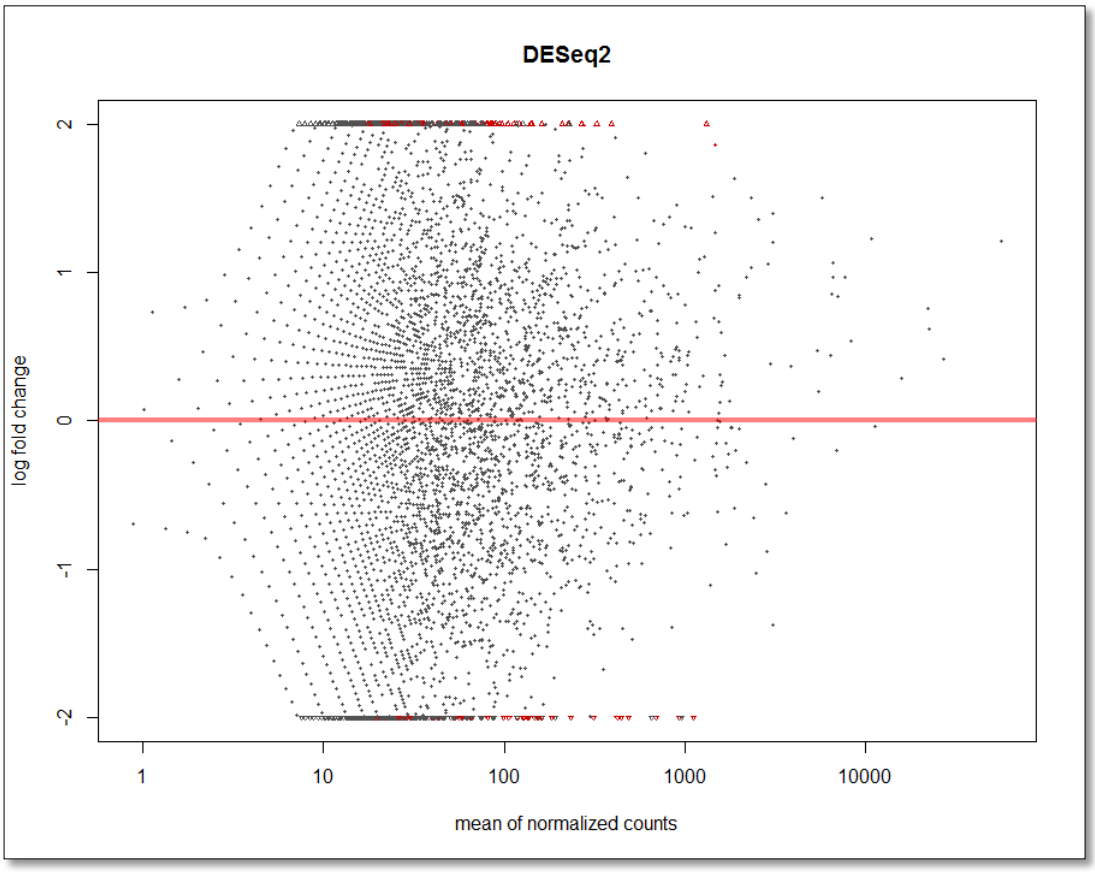
23 out of 57 transcripts did not share any GO terms and these novel transcripts are unclassified. Many of the remaining transcripts had more than one GO annotation.

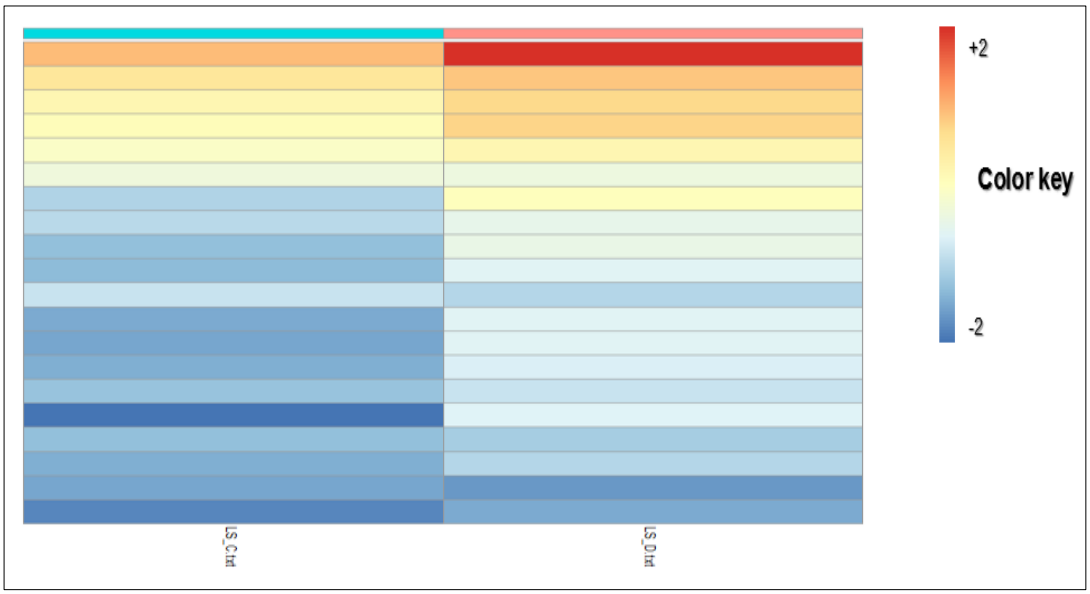
In case of reference-genome based approach, 140 transcripts were classified as differentially expressed genes (DEGs) in moisture stressed sample. Out of DEGs, 74 transcripts were found to be up-regulated whereas, 66 transcripts were found to be down-regulated (Figure 4.8 b).Upregulated genes includedsuccinate dehydrogenase, GTP-binding homolog, histone H4 and sucrose synthase 2. Down-regulated genes included asparticase 2 and BSD domain. 25 out of 140 transcripts did not share any GO terms and these novel transcripts are unclassified.


DESeq2
log fold change
2
1
0
-1
-2
1
100
10000
mean of normalized counts


+2
Color key
-2
LSC1#
LSC1#

(a)





(b)

Figure 4.8 Graphical representation of DEGs through MA plot and heat map in a) *de novo* based approach and b) reference-genome based approach. The red dots signifies differentialy expressed transcripts.

Discussion

CHAPTER V
DISCUSSION

RNA-Seq makes use of Next Generation Sequencing (NGS) technology to precisely determine the expression level of several genes in particular time and in particular cell by sequencing RNA. The recent advancement in the next-generation sequencing technology has made the characterization of transcriptome cost effective (Vera *et al.,* 2008; Ekblom and Galindo, 2011). Together with the growing popularity of RNA-Seq, a number of data analysis methods and pipelines have already been developed for this task. There is a common assumptions that substantial gains occur in the quality of the results as read length increases and when paired-ends (PE) are used. The current read length that is standard for many expeiments is PE 100 bp reads (Chhangawala *et al.,* 2015). Currently, however, there are no clear consensus about the best practices for SOLiD short read single end data, which makes the choice of an appropriate method a daunting task especially for a basic user. The performance of the bioinformatics RNA-Seq analysis tools are influenced by the type of dataset. Hence, to select an optimum performing tool for transcriptome analysis of *Lathyrus sativus*, we first attempted to optimize the same protocol using a well annotated sequenced genome of Glycine max, which is also a close relative of *Lathyrus sativus*. The reference transcriptome data generated form same sequencing technology (ABI-SOLiD) as that of *Lathyrus* and was of the same length (50bp) and type (single-end). Comparison was made between commercialized and open source software for the de novo and reference based methods, with the aim to understand and assist the choice of selection of such methods for SOLiD transcriptome data. *CLC Genomics Workbench,* a commercialized integrated software was compared with *Velvet-Oases* for *de novo* assembly and with *TopHat-Cufflinks* for reference genome based approach. Assembly and read alignment in case of *de novo* and reference-genome based approach respectively is the first crucial step on which the rest of the downstream analysis lies upon. *Velvet-Oases* and *Tophat-Cufflinks* yielded optimum assembly and mapping result over *CLC Genomics Workbench* in *de novo* and reference-genome based methods respectively based on the transcriptome data of *Glycine max*. Hence based on the assembly and mapping statistics (presented in section 4.1.2) of the RNA-Seq reads *Velvet-Oases* and *TopHat-Cufflinks* were chosen

to carry out the transcriptome analysis of *Lathyrus sativus* based on their performance against *Glycine max* test dataset.

Moisture stress is one of the most important environmental stresses all around the world. Drought tolerance is a complex trait controlled by many genes. It is important to understand how plants respond to drought stress at the molecular level for developing improved genotypes which would perform well under water limited conditions. *Lathyrus sativus* offers an attractive choice for sustainable food production, owing to its intrinsic properties including limited water requirement and drought tolerance. It is important to mine candidate genes and unravel molecular mechanisms in response to drought stress in *Lathyrus,* which would accelerate genetic improvement through technologies like marker-assisted selection.

Lathyrus sativus (Grass pea) is a dual purpose legume crop. Its seed are used for human consumption and fodder for consumption by livestock. It is well adapted to the arid conditions and is one of the hardiest pulses known till date. It contains a high level of protein (25-30%). It holds the tag of "insurance crop" and hence take a special importance in the light of climate change. Drought negatively impacts plant growth and the productivity of crops around the world. Understanding the molecular mechanisms in the drought response is important for improvement of drought tolerance using molecular techniques. The genes responsible for the plant's remarkable environmental tolerance are unknown (Yan *et al.,* 2006). The biotechnological potential of grass pea as a source of stress tolerance genes for general crop improvement remains to be exploited. A solution to the current stagnation is expected from high technological advancement such as transcriptome analysis providing insight into stress related gene activity which may accelerate knowledge based breeding. With the aim to elucidate the underlying molecular basis of grass pea response to drought stress, comparative trancriptome analysis was conducted between cultivars grown under two different conditions, control and drought stressed.

Since the genome of *Lathyrus sativus* (grass pea) is not yet sequenced, this study was carried out to identify the novel transcripts as well as differential expressed genes from both the *de novo* and reference-genome based method. Hence, assembly of clean reads of *Lathyrus sativus* was carried out by *Velvet/Oases* Assembler. A total of 64,246,685 and 64,640,515 raw reads were generated using ABI-SOLiD 4

sequencing technology, were assembled into 24621 contigs with N50 equals to 459 bp and 44.35% GC content value. It was found that around 63.21% transcripts showed similarity with known sequences when against searched non-redundant (nr) database, which implies that the other unknown, uncharacterized and hypothetical genes may be the novel transcripts which requires further experimental validation. Maximum number of contigs matched with *Glycine max* which signifies its close relationship with *Lathyrus sativus.* Hence, making its genome most suitable to be used as a reference genome for reference-genome based method of RNA-Seq analysis.

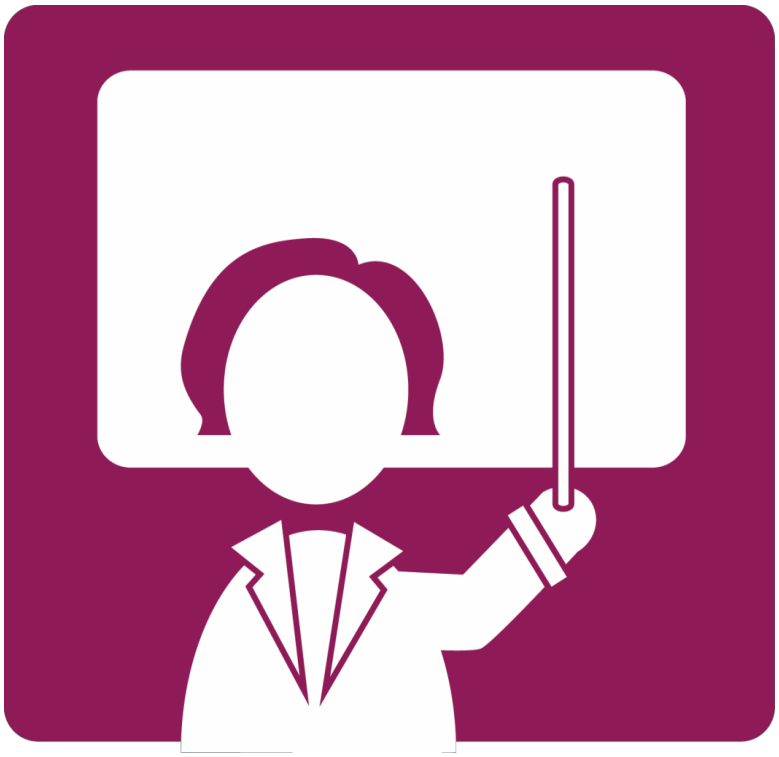
In case of reference-genome based approach, *TopHat-Cufflinks* was used for aligning reads to the reference-genome and assembling the mapped reads into a contigs. It generated 96,028 contigs, out of which 89,087 showed significant similarity with the known sequences the function of which could be annotated by searches of public databases. These functions were classified by GO and the metabolic pathways were assigned using the KEGG database.

The identified differential expressed genes in our study in control and moisture-stressed *Lathyrus* cultivars may be responsible to play supportive role in plant tolerance mechanism against drought condition. Through a comparative transcriptome analysis, we identified several moisture-stress responsive genes encoding moisture-responsive transcription factors and anti-oxidant enzymes. In *de novo* based approach, 57 transcripts were classified as differentially expressed genes (DEGs) in moisture stressed sample. Out of DEGs, 32 transcripts were found to be up-regulated whereas, 25 transcripts were found to be down-regulated. Upregulated genes included cysteine protease, ascorbate oxidase and heat shock protein. Down-regulated genes included NAC domain-containing 72-like protein and Zinc Finger 512B. 23 out of 57 transcripts did not share any GO terms and these novel transcripts are unclassified. Many of the remaining transcripts had more than one GO annotation which indicates one gene can give rise to more than one enzymes with different functions. Total 21 number of transcription factors were found in de novo based study which includes bZIP (basic leucine-zipper), C2H2-like zinc finger, C3H, EIL, NAC and MYB.

Cysteine protease is found to be involved in signaling pathway and in response to biotic and abiotic stress. Previous study has shown its accumulation in leaf tissue of

drought-stresses tomato (Harrak *et al.*, 2001) and *Arabidopsis thaliana* (Koizumi *et al.*, 1993). NAC is plant specific-proteins constituting a major transcription factor family, renowned for its role in several developmental programs and plays an important role in regulating stress responses. NAC are novel stress responsive genes that are activated and play a positive role in regulating the abscisic acid (ABA) mediating pathway (Hong *et al.*, 2016). It has received special attention due to its regulating mechanism in stress signaling pathways (Puranik *et al.*, 2012). Zinc finger are the class of domains which regulates the mechanism against biotic and abiotic stress. We also found these genes highly upregulated in our study.

In case of reference-genome based approach, 140 transcripts were classified as differentially expressed genes (DEGs) in moisture stressed sample. Out of DEGs, 74 transcripts were found to be up-regulated whereas, 66 transcripts were found to be down-regulated. Upregulated genes included succinate dehydrogenase, GTP-binding homolog, histone H4 and sucrose synthase 2. Succinate dehydrogenase was found to be differentially expressed during drought condition in in kernels of two maize lines. This enzyme participated in carbohydrate metabolism (Yang *et al.*, 2014). Sucrose synthase 2 is supposed to regulate nitrogen-fixation in nodules of Soybean in drought condition (González *et al.*, 1995). Down-regulated genes included aspartic are 2, BSD domain, Glyceraldehyde-3-phosphate cytosolic and Alcohol dehydrogenase 1. Glyceraldehyde-3-phosphate dehydrogenase has been found to interact with plasma membrane-associated phospholipase D to induce hydrogen peroxide signal during stress condition in *A. thaliana* (Guo *et al.*, 2012). Out of 140 transcripts, 25 did not share any GO terms and these novel transcripts are unclassified. Transcription factors included bHLH, NAC, S1Fa-like, LBD, TCP, ERF and WRKY. Identification of these differentially expressed genes may help in understanding the molecular mechanism of the drought resistant trait. Ethylene Response Element binding Factors (ERF) performs several diverse roles in abiotic and biotic stress (Sharoni *et al.*, 2011). WRKY is one of the largest transcription factor families and actively participate in drought stress condition (Qin *et al.*, 2013). These TFs have been identified in this study.



Summary and Conclusion

CHAPTER VI

SUMMARY AND CONCLUSION

Lathyrus sativus, commonly known as Grass pea belongs to family Fabaceae and sub-family Papilionoideae and is the only cultivated species of genus Lathyrus. This plant shows high resilience to moisture stress conditions i.e., both drought and flood. Drought is the period during which the amount of moisture in the soil no longer meets the needs of the particular crop. The struggle to grow food crops during such scarcity of moisture is therefore a big issue. It has also very high nutritive value. Grass pea (*Lathyrus sativus* L.) is a crop of immense economic and agronomic importance which has multiple uses as food, feed and fodder and hence it is used for both human and livestock consumption. But in spite of possessing such agronomical valuable traits it has remained underused and neglected species and very limited research has been devoted to this crop. Physiological studies that could aid our understanding the mechanisms and traits resulting in drought resistance are scarce and also not well understood. Hence this study was undertaken to unravel the genes and its expression along with their metabolic pathway that assist the crop to withstand moisture stress condition.

The protocol of transcriptome analysis of Lathyrus was first optimized by the use RNA-Seq SOLiD data of *Glycine max* which has a well annotated sequenced genome. Comparative study was made between commercial and open-source software for both the approaches of transcriptome analysis and the optimum performing tool was selected to carry out the RNA-Seq analysis of *Lathyrus sativus* SOLiD data. Velvet-Oases was selected for de novo assembly and TopHat-Cufflinks was used for reference-genome guided assembly. Further steps for downstream analysis for detecting the differential expression of genes were kept same in both the approaches so that only difference between the result of *de novo* and reference-genome based approach is the presence of annotated genome. Assemblies of different k-mer length were merged to non-redundant assembly at four different k-mers. The k-mer size at which the value of N50 and average transcripts length was higher, assembly at that k-mer size was considered to be the optimum assembly one could achieve using data of such short read length (50 bp). All the assembled transcripts were blasted against nr-database of NCBI to functionally characterize

those transcripts. KEGG pathway analysis was done to know the metabolic pathway in which the transcripts are involved, that would help us to understand the molecular mechanism of drought-resistance trait. DESeq2 was used for testing the differential expression of the genes. Genes having absolute log2 fold change greater than 2 and p-value less than 0.05 was considered to be differentially expressed.

The assembly resulted in 21,621 and 96,028 contigs in *de novo* and reference-genome based approach. Total 57 genes were found to be differentially expressed in case of de novo based approach out of which 32 genes were up regulated and 25 genes were down regulated. Upregulated genes included cysteine protease, ascorbate oxidase and heat shock protein. Down-regulated genes included NAC domain-containing 72-like protein and Zinc Finger 512B. 16 of the up regulated genes and 7 of the down regulated genes did not share any GO terms which means these genes are yet to be annotated. 63.21% and 93.53% of the total assembled contigs showed significant similarity with the known sequences of the NCBI database. Maximum transcripts got blast hit with *Glycine max* which also justifies our choice of reference-genome in case of reference-genome guided assembly. Total 21 number of transcription factors were found in de novo based study which includes bZIP (basic leucine-zipper), C2H2-like zinc finger, C3H, EIL, NAC and MYB.

Total 140 genes were found to be differentially expressed in case of reference-genome based approach out of which 74 genes were up regulated and 66 genes were down regulated. Upregulated genes included succinate dehydrogenase, GTP-binding homolog, histone H4 and sucrose synthase 2. Down-regulated genes included asparticase 2 and BSD domain. 16 of the up regulated genes and 7 of the down regulated genes did not share any GO terms. Transcription factors included bHLH, NAC, S1Fa-like, LBD, TCP and WRKY.

The most dominant biological process was “metabolic process” in both the *de novo* and reference based approach. The dominant molecular function in case of *de novo*was “nucleotide binding” and in case of reference-based approach it was “hydrolase activity”. In case of *de novo* the dominant cellular component was “organelle” and in other method it was “cell part”.

The findings of this study is expected to facilitate the decision of choosing an optimal tools for the analysis of short read SOLiD transcriptome data. The result is

also expected to provide an improved understanding and identification of sources for resistance against moisture stress for future genetic research in this hitherto under-researched, valuable legume crop.

ABSTRACT

Lathyrus has a great agronomic importance as it is grown for both human consumption and livestock feed and is well adapted to the arid conditions and is one of the hardiest pulses known till date. Together with the growing popularity of RNA-Seq, a number of data analysis methods and pipelines have already been developed for transcriptome analysis. There is a common assumption that substantial gains occur in the quality of the results as read length increases and when paired-ends (PE) are used. Currently, however, there are no clear consensus about the best practices for SOLiD short read single end data, which makes the choice of an appropriate method a daunting task especially for a basic user. Hence a comparative study of RNA-Seq analysis tools, in this study commercialized *CLC bio Genomics Workbench* vs open-source software like *Velvet-Oases* and *TopHat-Cufflinks* for *de novo* and reference-genome based approach respectively, was made with the aim to understand and assist the choice of selection of such methods for SOLiD transcriptome data. *Velvet-Oases* and *TopHat-Cufflinks* were chosen to carry out the transcriptome analysis of *Lathyrus sativus* based on their performance against *Glycine max* test dataset. Drought negatively impacts plant growth and the productivity of crops around the world. Understanding the molecular mechanisms in the drought response is important for improvement of drought tolerance using molecular techniques. In this study, we found 57 differentially expressed genes in case of *de novo* based approach and 140 in case of reference-genome based approach. The findings of this study is expected to facilitate the decision of choosing an optimal tools for the analysis of short read SOLiD transcriptome data. The result is also expected to provide an improved understanding and identification of sources for resistance against moisture stress for future genetic research in this hitherto under-researched, valuable legume crop.

सार

लैथाइरस का एक विशेष कृषि महत्व है क्योंकि यह मानव उपभोग और पशुओं के भोजन दोनों के लिए उगाया जाता है और यह शुष्क क्षेत्रों के लिए अनुकूल है और आज तक ज्ञात सबसे सख्त दालों में से एक है। आर.एन.ए.-सैक की बढ़ती लोकप्रियता के साथ, ट्रांस्क्रिप्टोम विश्लेषण के लिए पहले से ही कई डेटा विश्लेषण विधियों और पाइपलाइनों का विकास किया गया है। एक आम धारणा है कि परिणामों की गुणवत्ता में पर्याप्त लाभ होने के कारण रीड्स की लंबाई बढ़ जाती है और जब पेयर्ड-एण्ड (पीई) का उपयोग किया जाता है। वर्तमान में, हालांकि, एस.ओ.ली.ड. छोटे रीड्स वाले एकल अंत डेटा के लिए सर्वोत्तम अभ्यासों के बारे में कोई स्पष्ट सहमति नहीं है, जो विशेष रूप से उपयुक्त विधि का चयन एक मूल उपयोगकर्ता के लिए एक चुनौतीपूर्ण कार्य बनाता है। अतः आरएनए-सैक विश्लेषण उपकरणों का एक तुलनात्मक विश्लेषण, इस अध्ययन में सीएलसी बायो जीनोमिक्स वर्कबेन्च बनाम खुले स्रोत सॉफ़्टवेयर जैसे वेलवेट-ओसेस और टॉप-हेट-कफलिंक के लिए क्रमशः डे नोवो और संदर्भ-जीनोम आधारित दृष्टिकोण क्रमशः समझने के उद्देश्य से बनाया गया था और एस.ओ.ली.ड. ट्रांस्क्रिप्टोम डेटा के लिए इस तरह के विधियों के चयन के विकल्प की सहायता करें। ग्लाइसीन मैक्स टेस्ट डाटासेट के खिलाफ उनके प्रदर्शन के आधार पर लैथाइरस सटाइवस के ट्रांस्क्रिप्टोम विश्लेषण के लिए वेलवेट-ओसेस और टॉपहैट-कफ़लिंक को चुना गया। सूखा नकारात्मक रूप से पौधों के विकास और दुनिया भर में फसलों की उत्पादकता पर असर डालता है। सूखे की प्रतिक्रिया में आणविक तंत्र को समझना आणविक तकनीकों का उपयोग करके सूखा सहिष्णुता के सुधार के लिए महत्वपूर्ण है। इस अध्ययन में, हमें संदर्भ-जीनोम आधारित दृष्टिकोण के मामले में 140 और डे नोवो आधारित दृष्टिकोण के मामले में 57 अलग-अलग व्यक्त जीन मिले। इस अध्ययन के निष्कर्षों से उम्मीद है कि लघु-रीड्स एस.ओ.ली.ड. ट्रांस्क्रिप्टोम डेटा के विश्लेषण के लिए एक इष्टतम उपकरण चुनने के निर्णय की सुविधा होगी। नतीजतन, भविष्य में आनुवंशिक अनुसंधान के लिए नमी तनाव के विरुद्ध बेहतर स्रोतों की पहचान और पहचान प्रदान करने प्रतिरोध की भी आशा है।

BIBLIOGRAPHY

Allkin R., Macfarlane T.D., White R.J., Bisby F.A. and Adey M.E. 1985. The geographical distribution of Lathyrus. Vicieae Database Project Publication No. 6. University of Southampton.

Almeida, N. F., Leitão, S. T., Krezdorn, N., Rotter, B., Winter, P., Rubiales, D., and Patto, M. C. V. (2014). Allelic diversity in the transcriptomes of contrasting rust-infected genotypes of *Lathyrus sativus*, a lasting resource for smart breeding. *BMC plant biology*, *14*(1), 376.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25-29.

Bharati, M. P. (1986). Status of Lathyrus sativus among grain legumes cultivated in Nepal.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E. and Horsman, D.E. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872-2877.

Boyer, J. S. (1982). Plant productivity and environment. *Science*, *218*(4571), 443-448.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*, *100*(7), 3960-3964.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. and Roth, R., (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, *18*(6), 630-634.

Chhangawala, S., Rudy, G., Mason, C. E., and Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome biology*, *16*(1), 131.

Collins, L. J., Biggs, P. J., Voelckel, C., and Joly, S. (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome informatics*, *21*, 3-14.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674-3676.

Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1-15.

Gan, Q., Chepelev, I., Wei, G., Tarayrah, L., Cui, K., Zhao, K., and Chen, X. (2010). Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell research*, *20*(7), 763-783.

González, E. M., Gordon, A. J., James, C. L., and Arrese-Lgor, C. (1995). The role of sucrose synthase in the response of soybean nodules to drought. *Journal of Experimental Botany*, *46*(10), 1515-1523.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. and Conesa, A., (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, *36*(10), 3420-3435.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. and Chen, Z., (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), 644-652.

Guo, L., Devaiah, S. P., Narasimhan, R., Pan, X., Zhang, Y., Zhang, W., and Wang, X. (2012). Cytosolic glyceraldehyde-3-phosphate dehydrogenases interact with phospholipase Dδ to transduce hydrogen peroxide signals in the Arabidopsis response to stress. *The Plant Cell*, *24*(5), 2200-2212.

Gusmao, M., Siddique, K. H. M., Flower, K., Nesbitt, H., and Veneklaas, E. J. (2012). Water deficit during the reproductive period of grass pea (*Lathyrus sativus* L.) reduced grain yield but maintained seed size. *Journal of agronomy and crop science*, *198*(6), 430-441.

Harrak, H., Azelmat, S., Baker, E.N., and Tabaeizadeh, Z. (2001). Isolation and characterization of a gene encoding a drought-induced cysteine protease in tomato *(Lycopersicon esculentum). Genome*, 44**,** 368–374.

Hiremath, P.J., Farmer, A., Cannon, S.B., Woodward, J., Kudapa, H., Tuteja, R., Kumar, A., BhanuPrakash, A., Mulaosmanovic, B., Gujaria, N. and Krishnamurthy, L., (2011). Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant biotechnology journal*, *9*(8), 922-931.

Hong, Y., Zhang, H., Huang, L., Li, D., and Song, F. (2016). Overexpression of a stress-responsive NAC transcription factor gene ONAC022 improves drought and salt tolerance in rice. *Frontiers in plant science*, *7*.

Jackson, B. G., Schnable, P. S., and Aluru, S. (2009). Parallel short sequence assembly of transcriptomes. *BMC bioinformatics*, *10*(1), S14.

Jeuffroy, M. H., and Ney, B. (1997). Crop physiology and productivity. *Field Crops Research*, *53*(1), 3-16.

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, *45*(D1), D1040-D1045.

Joshi, K. D., Subedi, M., Rana, R. B., Kadayat, K. B., and Sthapit, B. R. (1997). Enhancing on-farm varietal diversity through participatory varietal selection: a case study for Chaite rice in Nepal. *Experimental Agriculture*, *33*(3), 335-344.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M. and Hayashizaki, Y. (2006). CAGE: cap analysis of gene expression. *Nature methods*, *3*(3), 211-222.

Koizumi, M., Yamaguchi-Shinozaki, K., Tsuji, H., and Shinozaki, K. (1993). Structure and expression of two genes that encode distinct drought-inducible cysteine proteinases in Arabidopsis thaliana. *Gene*, *129*(2), 175-182.

Kumar, S., Bejiga, G., Ahmed, S., Nakkoul, H., and Sarker, A. (2011). Genetic improvement of grass pea for low neurotoxin (β-ODAP) content. *Food and Chemical Toxicology*, *49*(3), 589-600.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics*, *24*(5), 713-714.

Lister, R., Gregory, B. D., and Ecker, J. R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current opinion in plant biology*, *12*(2), 107-118.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M., (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, *458*(7234), 97-101.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, *9*, 387-402.

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560-564.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-628.

Nunes, C., Araújo, S. S., Silva, J. M., Fevereiro, P., and Silva, A. B. (2009). Photosynthesis light curves: a method for screening water deficit resistance in the model legume Medicago truncatula. *Annals of applied biology*, *155*(3), 321-332.

Okoniewski, M. J., and Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics*, *7*(1), 276.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome biology*, *11*(12), 220.

Puranik, S., Sahu, P. P., Srivastava, P. S., and Prasad, M. (2012). NAC proteins: regulation and role in stress tolerance. *Trends in plant science*, *17*(6), 369-381.

Qin, Y., Tian, Y., Han, L., and Yang, X. (2013). Constitutive expression of a salinity-induced wheat WRKY transcription factor enhances salinity and ionic stress tolerance in transgenic Arabidopsis thaliana. *Biochemical and biophysical research communications*, *441*(2), 476-481.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q. and Griffith, M., (2010). De novo assembly and analysis of RNA-seq data. *Nature methods*, *7*(11), 909-912.

Sanger, F. (1975). The Croonian Lecture, 1975: Nucleotide Sequences in DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, *191*(1104), 317-333.

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086-1092.

Seyednasrollah, F., Laiho, A., and Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*, *16*(1), 59-70.

Sharoni, A.M., Nuruzzaman, M., Satoh, K., Shimizu, T., Kondoh, H., Sasaya, T., Choi, I.R., Omura, T. and Kikuchi, S., (2010). Gene structures, classification and expression models of the AP2/EREBP transcription factor family in rice. *Plant and cell physiology*, *52*(2), pp.344-360.

Shinozaki, K., and Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. *Journal of experimental botany*, *58*(2), 221-227.

Talukdar, D. (2013). Bioaccumulation and transport of arsenic in different genotypes of lentil (Lens culinaris Medik.). *International Journal of Pharma and Bio Sciences*, *4*(1), 694-701.

Trindade, I., Capitão, C., Dalmay, T., Fevereiro, M. P., and Dos Santos, D. M. (2010). miR398 and miR408 are up-regulated in response to water deficit in Medicago truncatula. *Planta*, *231*(3), 705-716.

Tripp, R., and Heide, W. M. (1996). *The erosion of crop genetic diversity: challenges, strategies and uncertainties*. Overseas Development Institute.

Tyagi, A., Santha, I. M., and Mehta, S. L. (1999). Effect of water stress on proline content and transcript levels in Lathyrus sativus.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., and Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular ecology*, *17*(7), 1636-1647.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, *55*(4), 641-658.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

Wu, T. D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873-881.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S. and Zhou, X., (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660-1666.

Yan, Z. Y.; Spencer, P. S.; Li, Z. X.; Liang, Y. M.; Wang, Y. F.; Wang, C. Y.; Li, F. M., 2006. *Lathyrus sativus* (grass pea) and its neurotoxin ODAP. Phytochemistry, 67 (2): 107-121

Yang, L., Jiang, T., Fountain, J.C., Scully, B.T., Lee, R.D., Kemerait, R.C., Chen, S. and Guo, B., (2014). Protein profiles reveal diverse responsive signaling pathways in kernels of two maize inbred lines with contrasting drought sensitivity. *International journal of molecular sciences*, *15*(10), 18892-18918.

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), 821-829.

Zivcak, M., Brestic, M., Olsovska, K., and Slamka, P. (2008). Performance index as a sensitive indicator of water stress in Triticum aestivum L. *L. Plant Soil Environ*, *54*(4), 133-139.

APPENDIX

1. Command used for DESeq2 in R-3.4.0

```
> utils:::menuInstallLocal()
> library("DESeq2")
> directory <- "C:\\Users\\Ishu\\Desktop\\denovo_htseq"
> sampleFiles <- grep( "treated" , list.files(directory),
value=TRUE)
> sampleCondition <- sub( "(. *treated). *" , "\\1"
,sampleFiles)
> sampleTable <- data.frame( sampleName = sampleFiles,
fileName = sampleFiles, condition = sampleCondition)
> colnames<-c("id", "count")
> ddsHTSeq <- DESeqDataSetFromHTSeqCount( sampleTable =
sampleTable, directory = directory, design= ~ condition)
> ddsHTSeq
> directory <- system.file( "extdata" , package="pasilla"
, mustWork=TRUE)
> library("DESeq2")
> directory <- "C:\\Users\\Ishu\\Desktop\\denovo_htseq"
> sampleFiles <- grep( "treated" , list.files(directory),
value=TRUE)
> sampleCondition <- sub( "(. *treated). *" , "\\1"
,sampleFiles)
> sampleTable <- data.frame( sampleName = sampleFiles,
fileName = sampleFiles, condition = sampleCondition)
> ddsHTSeq <- DESeqDataSetFromHTSeqCount( sampleTable =
sampleTable, directory = directory, design= ~ condition)
> ddsHTSeq
> dds <- DESeq(ddsHTSeq)
> head(ddsHTSeq)
> head(dds)
> dds <- ddsHTSeq[ rowSums( counts(ddsHTSeq)) > 1, ]
> dds <- DESeq(dds)
> res <- results(dds)
> register( SnowParam( 4))
> resOrdered <- res[ order(res$padj),]
> sum(res$padj < 0.1, na.rm=TRUE)
> dds
> res05 <- results(dds, pvalue=0.05)
> res05 <- results(dds, pAdjustMethod = "FDR", alpha=.05)
```

2. Installation procedure of required software

All the executable programs were stored in a common directory before installing and the same directory was added to the PATH environment variable.

```
$ mkdir $HOME/bin
$ export PATH = $HOME/bin:$PATH
```

To install the SAM tools, SAM tools (http://samtools.sourceforge.net) was downloaded and unpacked the SAM tools tarball and cd to the SAM tools source directory:

```
$ tar jxvf samtools-0.1.17.tar.bz2
$ cd samtools-0.1.17
$ cp samtools $HOME/bin
```

To install Bowtie, the latest binary package for Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) was downloaded and unpacked the Bowtie zip archive and cd to the unpacked directory:

```
$ unzip bowtie-0.12.7-macos-10.5-x86_64.zip
$ cd bowtie-0.12.7
$ cp bowtie $HOME/bin
$ cp bowtie-build $HOME/bin
$ cp bowtie-inspect $HOME/bin
```

To install TopHat, the binary package for version 1.3.2 of TopHat (http://tophat.cbcb.umd.edu/) was downloaded and unpacked the TopHat tarball and cd to the unpacked directory:

```
$ tar zxvf tophat-1.3.2.OSX_x86_64.tar.gz
$ cd tophat-1.3.2.OSX_x86_64
cp * $HOME/bin
```

To install Cufflinks, the binary package of version 1.2.1 for Cufflinks (http://cufflinks.cbcb.umd.edu/) was downloaded and unpacked the Cufflinks tarball and cd to the unpacked directory:

```
$ tar zxvf cufflinks-1.2.1.OSX_x86_64.tar.gz
$ cd cufflinks-1.2.1.OSX_x86_64
$ cp * $HOME/bin
```

To install CummeRbund, we need to start an R session:

```
$ R
```

And then CummeRbund package was installed using the following commands:

```
> source('http://www.bioconductor.org/biocLite.R')
> biocLite('cummeRbund')
```

3. Up-regulated differentially expressed genes in case of *de novo* based aaproach

Transcripts	Description	log2FoldChange	pvalue
Locus_10015_Transcript_1/1_Confidence_0.000_Length_111	NA	2.428785191	0.008112385
Locus_10019_Transcript_1/1_Confidence_0.000_Length_231	papain family cysteine protease	2.391488221	0.003788892
Locus_10037_Transcript_1/1_Confidence_0.000_Length_132	heat shock 70 kDa	2.347678474	0.037957849
Locus_10058_Transcript_1/1_Confidence_0.000_Length_251	L-ascorbate oxidase homolog	2.324160459	0.027984583
Locus_10245_Transcript_1/1_Confidence_0.000_Length_108	NA	2.318997942	0.03221003
Locus_10262_Transcript_1/1_Confidence_0.000_Length_150	NA	2.309930524	0.040400342
Locus_10296_Transcript_1/1_Confidence_0.000_Length_142	xyloglucan endotransglucosylase hydrolase family	2.302225339	0.033197155
Locus_10516_Transcript_1/1_Confidence_0.000_Length_108	NA	2.286281088	0.049422817
Locus_10545_Transcript_1/1_Confidence_0.000_Length_130	NA	2.282775042	0.019241313
Locus_10550_Transcript_1/1_Confidence_0.000_Length_130	NA	2.263249186	0.043242757

Transcripts	**Description**	**log2FoldChange**	**pvalue**
Locus_10554_Transcript_1/1_Confidence_0.000_Length_181	ribulose bisphosphate carboxylase small chloroplastic-like	2.249559249	0.03934212
Locus_10648_Transcript_1/1_Confidence_0.000_Length_197	NA	2.239075009	0.017407539
Locus_10677_Transcript_1/1_Confidence_0.000_Length_185	NA	2.236679763	0.020853637
Locus_1073_Transcript_2/10_Confidence_0.143_Length_262	Histone	2.203005639	0.023316637
Locus_10776_Transcript_1/1_Confidence_0.000_Length_113	NA	2.198346473	0.042102855
Locus_10849_Transcript_1/1_Confidence_0.000_Length_106	Signal recognition particle 14 kDa	2.150845819	0.030108115
Locus_10899_Transcript_1/1_Confidence_0.000_Length_159	rhodanese-related sulfurtransferase	2.150769889	0.008708429
Locus_10908_Transcript_1/1_Confidence_0.000_Length_133	probable inactive receptor kinase At1g48480	2.147717827	0.034726978
Locus_1090_Transcript_1/1_Confidence_0.000_Length_110	NA	2.138235964	0.015766526
Locus_11340_Transcript_1/1_Confidence_0.000_Length_159	NA	2.073032116	0.028408358

Transcripts	Description	log2FoldChange	pvalue
Locus_10954_Transcript_1/1_Confidence_0.000_Length_409	heat shock factor HSF30	2.136203052	0.037650248
Locus_11064_Transcript_1/1_Confidence_0.000_Length_210	65-kDa microtubule-associated 1-like	2.117991947	0.030107676
Locus_11100_Transcript_1/1_Confidence_0.500_Length_119	NA	2.117689436	0.023581589
Locus_11152_Transcript_1/1_Confidence_0.000_Length_178	TSS-like isoform X1	2.096469221	0.031035381
Locus_1135_Transcript_1/1_Confidence_0.000_Length_102	NA	2.060752437	0.03741885
Locus_11401_Transcript_1/1_Confidence_0.000_Length_101	NA	2.046242313	0.010029157
Locus_11413_Transcript_1/1_Confidence_0.000_Length_109	uncharacterized GPI-anchored At1g61900-like	2.045895207	0.024739139
Locus_11430_Transcript_1/2_Confidence_0.333_Length_228	methionine gamma-lyase-like	2.040941412	0.015930081
Locus_11487_Transcript_1/1_Confidence_0.000_Length_108	beta-glucosidase 18-like isoform X2	2.038792028	0.011325109
Locus_11516_Transcript_1/1_Confidence_0.000_Length_102	NA	2.016481297	0.035155868
Locus_11527_Transcript_3/3_Confidence_0.500_Length_367	hydroproline-rich partial	2.009078431	0.018709526
Locus_11567_Transcript_1/1_Confidence_0.000_Length_119	NA	2.008589243	0.021013529

4. Down-regulated differentially expressed genes in case of *de novo* based aaproach

Transcripts	Description	log2FoldChange	pvalue
Locus_8708_Transcript_1/1_Confidence_0.000_Length_161	NA	-2.018939331	0.028346529
Locus_8769_Transcript_1/1_Confidence_0.000_Length_157	probable methyltransferase PMT3	-2.036520926	0.045873136
Locus_8776_Transcript_1/1_Confidence_0.000_Length_192	NA	-2.04448145	0.011098058
Locus_8797_Transcript_1/1_Confidence_0.000_Length_173	aspartyl protease family At5g10770-like	-2.045904385	0.028552096
Locus_880_Transcript_1/1_Confidence_0.000_Length_313	patellin-4	-2.060360868	0.031895717
Locus_8816_Transcript_1/1_Confidence_0.000_Length_107	splicing partial	-2.067698159	0.01662676
Locus_8830_Transcript_1/1_Confidence_0.000_Length_120	16S rRNA processing	-2.071216776	0.041607877
Locus_889_Transcript_1/2_Confidence_1.000_Length_157	heat shock cognate 70	-2.07267341	0.049418002
Locus_894_Transcript_3/3_Confidence_0.200_Length_134	NA	-2.073475343	0.049796156
Locus_8954_Transcript_1/1_Confidence_0.000_Length_100	NA	-2.080333842	0.021868902

Transcripts	Description	log2FoldChange	pvalue
Locus_8993_Transcript_1/1_Confidence_0.000_Length_148	rhodanese-related sulfurtransferase	-2.101038794	0.048273259
Locus_9060_Transcript_1/1_Confidence_0.333_Length_169	NA	-2.108745537	0.015482402
Locus_9163_Transcript_1/1_Confidence_0.000_Length_174	ASPARTIC PROTEASE IN GUARD CELL 2-like	-2.109996104	0.005441022
Locus_9172_Transcript_1/1_Confidence_0.000_Length_125	NADPH:quinone oxidoreductase	-2.118901409	0.038604774
Locus_9175_Transcript_1/1_Confidence_0.000_Length_184	group 3 LEA	-2.127034834	0.039970524
Locus_9315_Transcript_1/2_Confidence_0.667_Length_343	photosystem I reaction center subunit chloroplastic	-2.153592068	0.043672398
Locus_9450_Transcript_1/1_Confidence_0.000_Length_174	boron transporter 1-like	-2.180777956	0.030436599
Locus_9524_Transcript_1/1_Confidence_0.000_Length_154	group 3 LEA	-2.239674528	0.027252912
Locus_9576_Transcript_1/1_Confidence_0.000_Length_104	NA	-2.250347238	0.04789871
Locus_96_Transcript_1/4_Confidence_0.278_Length_198	ADP-ribosylation partial	-2.290539479	0.044397594

Transcripts	**Description**	**log2FoldChange**	**pvalue**
Locus_982_Transcript_1/1_Confidence_0.000_Length_186	ribonucleoside-diphosphate reductase small chain	-2.343214759	0.041537121
Locus_98_Transcript_1/3_Confidence_0.364_Length_185	carbonic chloroplastic isoform X2	-2.406938388	0.027798238
Locus_9903_Transcript_2/2_Confidence_0.000_Length_170	NA	-2.415889052	0.032205063
Locus_993_Transcript_1/2_Confidence_0.333_Length_202	Zinc finger 512B	-2.423706229	0.029773343
Locus_9960_Transcript_1/1_Confidence_0.000_Length_214	NAC domain-containing 72-like	-2.478428163	0.029036794

5. Up-regulated differentially expressed genes in case of reference-genome based approach

Transcripts	Description	Fold Change value	p-value
TCONS_00000009	NA	5.024835172	0.004188072
TCONS_00000017	NA	4.657367364	0.048244107
TCONS_00000018	NA	4.453849024	0.025361956
TCONS_00000026	ATP synthase CF0 subunit III (chloroplast)	4.17420523	0.024563496
TCONS_00000064	NA	4.082825715	0.036880963
TCONS_00000079	NA	4.000236509	0.023607359
TCONS_00000091	NA	3.993708452	0.00397087
TCONS_00000100	NA	3.983291485	0.025143159
TCONS_00000949	metacaspase-1	3.961204521	0.028437859
TCONS_00001667	succinate dehydrogenase [ubiquinone] flavo subunit mitochondrial	3.850586043	0.038278267
TCONS_00002110	nucleotide-diphospho-sugar transferase superfamily	3.812024818	0.028437859

Transcripts	Description	Fold Change value	p-value
TCONS_00002536	plant F10M23-360	3.812024818	0.046015071
TCONS_00005556	Photosystem II reaction center W chloroplastic	3.772384308	0.02119161
TCONS_00005952	GTP-binding homolog	3.731610358	0.049491739
TCONS_00006151	Cellulose synthase A catalytic subunit 7 [UDP-forming]	3.731610358	0.020250037
TCONS_00008337	40S ribosomal S5	3.731610358	0.034080994
TCONS_00009311	Molybdopterin biosynthesis CNX1	3.731610358	0.024370143
TCONS_00010724	glycosyltransferase-like At2g41451	3.689645277	0.027407094
TCONS_00010958	deoxyhypusine hydroxylase	3.689397941	0.006413101
TCONS_00011069	histone H4	3.604080761	0.035182829
TCONS_00011111	thioredoxin chloroplastic	3.602423301	0.045234353
TCONS_00011166	probable xyloglucan endotransglucosylase hydrolase 32	3.601461974	0.038456978

Transcripts	Description	Fold Change value	p-value
TCONS_00011580	YIPF1 homolog	3.601461974	0.032311259
TCONS_00012785	E3 ubiquitin- ligase RNF25	3.592736048	0.024864802
TCONS_00012902	NRT1 PTR FAMILY -like isoform X1	3.57453212	0.041790708
TCONS_00013678	sucrose synthase 2	3.57453212	0.024434907
TCONS_00014188	phosphatidylinositol phosphatidylcholine transfer SFH6	3.564052539	0.024016918
TCONS_00014595	Transcription initiation factor TFIID subunit 5	3.555626722	0.013645427
TCONS_00016247	uncharacterized protein LOC100500559	3.555448572	0.034625827
TCONS_00016501	PREDICTED: uncharacterized protein LOC100775631	3.548174853	0.043642368
TCONS_00016567	MLO 1	3.519659448	0.030681086
TCONS_00018165	DNA (cytosine-5)-methyltransferase 1-like	3.512960937	0.047813919
TCONS_00018533	PLASTID TRANSCRIPTIONALLY ACTIVE 7	3.501990575	0.038456978

Transcripts	Description	Fold Change value	p-value
TCONS_00019662	cysteine ase COT44-like	3.462428691	0.037682012
TCONS_00019686	F-box-like WD repeat-containing TBL1XR1	3.459443589	0.006216155
TCONS_00019913	dynamin ARC5	3.459443589	0.044864138
TCONS_00020204	chalcone synthase	3.459443589	0.013391656
TCONS_00020504	tRNA uridine 5-carboxymethylaminomethyl modification enzyme	3.448624574	0.022028337
TCONS_00020556	signal recognition particle subunit SRP68	3.446170126	0.035075047
TCONS_00020724	Luminal-binding 5	3.408932104	0.0354892
TCONS_00021040	Universal stress A	3.408932104	0.048125824
TCONS_00021260	Ribosomal RNA small subunit methyltransferase G	3.408932104	0.041341968
TCONS_00022276	K-box region and MADS-box transcription factor family partial	3.365711914	0.047813919
TCONS_00022322	kinesin KIN12B	3.325931546	0.03288699

Transcripts	Description	Fold Change value	p-value
TCONS_00022583	alanine--glyoxylate aminotransferase 2	3.325931546	0.047813919
TCONS_00022983	probable phosphatase 2C 6	3.272684928	0.008235345
TCONS_00023173	histone deacetylase 19-like	3.272530736	0.044471041
TCONS_00023432	BRI1 kinase inhibitor 1-like	3.269234686	0.015548295
TCONS_00024788	ATP synthase subunit mitochondrial	3.25275154	0.029344409
TCONS_00025396	translation factor SUI1 homolog 1	3.24252138	0.04047731
TCONS_00025655	TMV resistance N-like	3.203387526	0.035075047
TCONS_00025928	Outer envelope pore chloroplastic	3.200633324	0.034625827
TCONS_00027179	ferrochelatase- chloroplastic	3.182539603	0.041341968
TCONS_00027280	elongation factor 1-beta	3.182539603	0.031102956
TCONS_00029048	kDa class III heat shock	3.146285892	0.016943483

Transcripts	Description	Fold Change value	p-value
TCONS_00029558	nucleolar 56-like	3.041520324	0.048076914
TCONS_00033426	myosin heavy chain	3.028594531	0.039631342
TCONS_00034041	Methyltransferase 16	3.011017966	0.028437859
TCONS_00034749	ICE-like protease (caspase) p20 domain	2.978900989	0.005735999
TCONS_00036695	F-box SKIP28	2.95729812	0.019057006
TCONS_00037829	clathrin assembly At1g03050	2.944669026	0.003414555
TCONS_00038483	probable phosphatase 2C 33	2.913710254	0.03766964
TCONS_00039131	Chaperonin-like isoform 1	2.846613062	0.022304938
TCONS_00039906	probable proteasome inhibitor	2.728103895	0.018112891
TCONS_00041670	D-alanine-D-alanine ligase	2.708797351	0.03288699
TCONS_00044576	tetratricopeptide repeat 4 homolog	2.704491179	0.026663132

Transcripts	Description	Fold Change value	p-value
TCONS_00044810	IQ domain-containing IQM3	2.647188302	0.037149709
TCONS_00045549	dormancy-associated homolog 3 isoform X2	2.499679796	0.002539794
TCONS_00046092	mitochondrial substrate carrier family C	2.485922456	0.04611193
TCONS_00046170	Low affinity potassium transport system kup isoform 1	2.459279897	0.028925596
TCONS_00046391	suppressor of Mek1-like	2.158086389	0.047050703
TCONS_00046583	chaperone dnaJ 15	2.131571285	0.039631342
TCONS_00047261	NA	2.081625394	0.045664503
TCONS_00047756	NA	2.007513112	0.019348326

6. Down-regulated differentially expressed genes in case of reference-genome based approach

Transcripts	Description	Fold Change value	p-value
TCONS_00050118	Phospholipase D delta	-2.025987617	0.029721542
TCONS_00050425	ATP-dependent Clp protease proteolytic subunit chloroplastic	-2.173665767	0.026674761
TCONS_00050947	histone H4	-2.192128376	0.039945492
TCONS_00051078	asparticase 2	-2.216979289	0.047050703
TCONS_00051543	Glyceraldehyde-3-phosphate cytosolic	-2.300438711	0.030359757
TCONS_00051989	BSD domain	-2.363262215	0.019813815
TCONS_00052690	serine threonine- phosphatase 6 regulatory subunit 3	-2.502133217	0.040116241
TCONS_00053741	signal recognition particle 14 kDa	-2.622634529	0.012851324
TCONS_00054498	vesicle-associated membrane 714	-2.647720969	0.007291467
TCONS_00055854	Chloroplastic	-2.671798625	0.047050703
TCONS_00058497	RNA-binding 38-like	-2.714803077	0.026902819

Transcripts	Description	Fold Change value	p-value
TCONS_00059130	S-acyltransferase 24	-2.720352833	0.021922307
TCONS_00059641	trihelix transcription factor ASIL2-like	-2.746411369	0.012235732
TCONS_00059686	homeobox domain	-2.773540089	0.028437859
TCONS_00059957	DJ-1 homolog B-like	-2.804947885	0.023377872
TCONS_00060756	lys-63-specific deubiquitinase BRCC36-like	-2.804947885	0.038456978
TCONS_00060970	rRNA intron-encoded homing	-2.815813946	0.041535577
TCONS_00062577	annexin D4	-2.840965599	0.033648057
TCONS_00063026	hemiasterlin resistant 1	-2.916702509	0.005623348
TCONS_00065086	Alcohol dehydrogenase 1	-3.004463188	0.001040835
TCONS_00066056	Exosome complex component RRP42	-3.067195895	0.007118907
TCONS_00066430	sorting nexin 2B-like	-3.07339222	0.014050976

Transcripts	Description	Fold Change value	p-value
TCONS_00067265	two-on-two hemoglobin-3	-3.101183363	0.043212553
TCONS_00068354	125 kDa kinesin-related –like	-3.133445682	0.04201628
TCONS_00071012	stem-specific TSJT1-like	-3.133445682	0.004035909
TCONS_00071624	choline-phosphate cytidylyltransferase 1-like	-3.192393588	0.032685802
TCONS_00072396	Aspartic ase nepenthesin-1	-3.194300132	0.045234353
TCONS_00072756	DNA topoisomerase 2	-3.267194147	0.038456978
TCONS_00073861	histidine kinase 1-like	-3.27139712	0.014900874
TCONS_00074522	KH domain-containing At1g09660 At1g09670	-3.308124392	0.002315577
TCONS_00076566	endo-1,4-beta-glucanase	-3.330049908	0.043357754
TCONS_00076631	glucose-1-phosphate adenylyltransferase large subunit chloroplastic amyloplastic	-3.363841329	0.016827223
TCONS_00076847	Polyubiquitin	-3.363841329	0.042682118

Transcripts	Description	Fold Change value	p-value
TCONS_00077521	plant UBX domain-containing 8-like	-3.363841329	0.043148521
TCONS_00077903	PREDICTED: uncharacterized protein LOC100777538	-3.392726902	0.047813919
TCONS_00081015	PREDICTED: uncharacterized protein LOC100500344	-3.392726902	0.047813919
TCONS_00081474	phosphomethylpyrimidine chloroplastic isoform X1	-3.392726902	0.044189581
TCONS_00083099	Proline iminopeptidase	-3.392726902	0.043642368
TCONS_00083276	transcription factor MYB34-like	-3.392726902	0.028925596
TCONS_00085011	Alba C9orf23	-3.429889312	0.036259601
TCONS_00085012	F-box kelch-repeat At3g24760	-3.429889312	0.048650708
TCONS_00085377	Ataxin-10	-3.444804517	0.024864802
TCONS_00086034	serine threonine- kinase VRK1	-3.456419044	0.037021496
TCONS_00087240	alpha-L-arabinofuranosidase 1-like	-3.466124843	0.033661166

Transcripts	Description	Fold Change value	p-value
TCONS_00087445	small nuclear ribonucleo Sm D1-like	-3.501471326	0.036177642
TCONS_00087706	dnaJ P58IPK homolog	-3.535964503	0.040568022
TCONS_00087715	phenylalanine ammonia-lyase 1	-3.535964503	0.016610248
TCONS_00089293	AUXIN SIGNALING F-BOX 2	-3.535964503	0.015658738
TCONS_00089953	probable beta-1,4-xylosyltransferase IRX9	-3.535964503	0.038456978
TCONS_00090085	peptidyl-prolyl cis-trans isomerase CYP18-2	-3.535964503	0.048308329
TCONS_00090673	NA	-3.602524266	0.048308329
TCONS_00090895	NA	-3.602524266	0.047503991
TCONS_00092431	NA	-3.634653056	0.013032989
TCONS_00093170	NA	-3.634653056	0.043642368
TCONS_00094593	NA	-3.644947051	0.042823277

Transcripts	Description	Fold Change value	p-value
TCONS_00094693	NA	-3.67111246	0.014149457
TCONS_00094762	NA	-3.696751754	0.039631342
TCONS_00094892	NA	-3.756620705	0.037149709
TCONS_00095019	NA	-3.920971971	0.030458504
TCONS_00095097	NA	-3.946504797	0.000725094
TCONS_00095099	NA	-3.949303619	0.026578232
TCONS_00095121	NA	-4.0661975	0.013011323
TCONS_00095139	NA	-4.088941165	0.044471041
TCONS_00095144	NA	-4.155135278	0.0020018
TCONS_00095150	NA	-4.310234567	0.023284268
TCONS_00095613	NA	-4.358471694	0.00691928