"Identification of Long Non-Coding RNAs in *Arabidopsis thaliana*"

Dissertation submitted in partial fulfilment of the requirement for the degree of

MASTER OF TECHNOLOGY

IN

BIOTECHNOLOGY

BY

ANKITA SINGH

Enrollment No.-212552

Under the guidance of

Dr. Shikha Mittal

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology

Waknaghat, Solan, (H.P.), India.

CERTIFICATION OF ORIGINALITY

This is to certify that the thesis titled **"Identification of Long Non-Coding RNAs in *Arabidopsis thaliana*"** is an original work of the student and is being submitted in Partial fulfillment for the award of the Degree of **Master of Technology (Biotechnology)**.

This dissertation has not been submitted earlier either to this University or to any other Institution for the fulfillment of the requirement of any course of study.

Signature of Supervisor

Dr. Shikha Mittal

Assistant Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P.

Signature of Student

Ankita Singh

UNDERTAKING FROM THE CANDIDATE

I, Ankita Singh, present the project entitled **"Identification of Long Non-coding RNAs in *Arabidopsis thaliana*"**. Though care has been taken while writing this report, there may still be some errors that are inadvertent on my part.

Place: Jaypee University of Information Technology, Waknaghat, Solan.

Signature of Candidate

Ankita Singh
Enrollment No.-212552
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P.

ACKNOWLEDGMENT

I would like to express a special Thanks of Gratitude to my project guide, **Dr. Shikha Mittal**, for her guidance and support. Her guidance helped me during the time of research and writing of this project report. The accomplishment of the task of writing this project report seemed impossible without the support, motivation, and facilitation of my guide and the cooperation of the other faculty members.

My heartfelt gratitude also goes to my Head of the Department **Prof. Dr. Sudhir Kumar** for allowing me to work on such a diversified project.

Ankita Singh (212552)

Date:

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

Abstract

Long non-coding RNAs (lncRNAs) are a kind of RNA molecule that plays important functions in many physiological processes and has been found to be linked to a variety of diseases/disorders. To understand the functional importance of long non-coding RNAs and determining how they interact with numerous biological systems, requires their identification and characterization depending upon various parameters.

In this thesis, we present a comprehensive identification and analysis of lncRNAs in *Arabidopsis thaliana*. In current analysis, firstly we identified 1284 significant differentially expressed genes using reference based assembly having log2FC>2 and p-value <0.05. Then we identified 473 lncRNAs by using multiple linux based tools like PLEK, CNCI, and CPC2. Further various parameters were used for filtration of lncRNAs like length >200, non-protein coding etc. Then we annotated the identified lncRNAs and did their expression based analysis to have a better idea of their possible functional roles. We also performed synteny analysis and found that the identified lncRNAs showed similarities with *Brassica rapa* and *Brassica napus*. Further we also identified 44 SSRs using Krait tool and also designed primers for SSRs using Primer-BLAST which will be validated in our future study.

Overall, this research provides a solid computational framework for finding and characterizing long non-coding RNAs using NGS technique. The findings help to enrich the catalogue of annotated lncRNAs and improve our understanding for their biological roles. The pipeline built can be a great resource for future studies on lncRNA research, allowing researchers to investigate their involvement in development, illness, and other biological processes.

Chapter-01 INTRODUCTION

In most cases RNA is a single stranded moiety. The existence of own-complementary sequences in the RNA results in the fold-change of the ribonucleotide chain into very complex structural shapes such as helices, bulges and also includes intra-chain base-pairing. The 3D structure of ribonucleic acid is crucial due to its stability, also it offers modification in nitrogenous bases and ribose sugar with the help of enzymes that attach with the methyl groups of the chain. These modifications facilitate bond formation between distant regions in the strand leading to a stabilized structure of RNA [1]. RNA is a molecule found in most organisms and viruses. It is also basic genetic material of various viruses [1]. It consists of nucleotide, which is a ribose sugar attached to a nitrogenous base and a phosphate group. Basically, RNAs exist primarily in a single-stranded form, but there are certain RNA viruses that are double-stranded [1]. The main job of RNA is to make proteins through translation. It provides genetic information that transcribes the ribosome into various proteins required for biological function. Other functions include gene regulation, RNA interference, and RNA editing [2]. These actions are carried out by small-regulatory RNAs, including microRNAs (miRNAs)/small-interfering RNAs (siRNAs), and small nuclear RNAs [3].

There are different types of RNAs depending upon different criteria, they are as follows.

Translated:

Messenger-RNA also termed as mRNAs, is a RNA which carries information from DNA to ribosome, the site of protein synthesis in the cells. Messenger-RNA is a copy of DNA. The amino acid sequence of the resulting protein can be determined by the help of the mRNA coding sequence.

Transfer-RNA also known as tRNA is also an adapter molecule that is composed of the RNA that serves a role of a physical link between the mRNA and the amino acid sequence of a protein. It is usually 7690 nucleotides in length (in eukaryotes). During translation, specific amino acids are added to the developing poly-peptide chain at the site of protein synthesis. It has an amino-acid attachment region and also an anti-codon region to recognize other codons that can form hydrogen bonds with specific sequences that are located on the messenger RNA strand. Ribosomal-RNA also called as rRNA is the most common type of RNA which is present in most of the cells, making-up 80% of all cellular RNA, regardless of never being translated into protein. Ribosomal-RNA is a non-coding RNA that is a major component of ribosomes required by all cells. Protein synthesis is carried out in the ribosome by rRNA. Ribosomal RNA should be translated from ribosomal DNA before it can be used to make tiny and very large ribosomal subunits (rDNA). The mechanical and the physical component of the ribosomes that allows messenger RNA (mRNA) and transfer-RNA (tRNA) to be processed and translated into proteins is called Ribosomal RNA (rRNA).

By length:

RNAs are classified into two types, short RNAs and long RNAs, based on the length of the RNA chain. Small-RNAs are usually less than 200 nucleotides in length, whereas long non-coding RNAs are more than 200 nucleotides in length. Non-coding RNA consists mainly of long non-coding RNA, lnc-RNA and mRNA. Small-RNAs are mainly small interfering RNA (siRNA), transfer RNA (tRNA), ribosomal-RNA (rRNA), mi-RNA, pivi-interacting RNA (piRNA), small RNA derived from t-RNA (cRNA), small nucleolar RNA (snoRNA) [4].

Back-ground on Long Non-coding RNAs (lncRNAs)

We know that Protein and DNA are the major components of chromatin. However, a growing body of research demonstrates that ribonucleic acid occupies a significant portion of the chromatin and controls nuclear architecture [5]. The ENCODE-project, estimates that most commonly non-coding transcripts cover about of 62-75% of genome, and contribute to the overall estimation of 80% of potentially functional sequences in DNA. Transcripts that are from the non-coding regions also dominate populations of the non-ribosomal, & non-mitochondrial RNAs in human cells. From Genome-wide Association studies, it is concluded that non-coding RNAs have emerged as an important source of biomarkers, therapeutic-targets, & putative explanations for the function of noncoding mutations. Previously considered as "junk", the potential roles of these molecules have received a great deal of attention and new discoveries. Two main members of nc-RNAs, are microRNAs (miRNAs) and long Coding RNA (lncRNAs) [6].

These are the transcripts that are longer than 200 nucleotides and are not translated into protein. Long non-coding RNAs have the ability to regulate gene expression at various levels, including transcription, and post-transcription and chromatin levels. They influence longevity pathways by regulating cell proliferation, differentiation, apoptosis, and senescence [7]. A notable number of long noncoding RNAs reside preferentially in nucleus and co-operate with the protein complexes to regulate phase separation, compartment formation, epigenetic-regulation, and nuclear organization.

These were once considered as 'transcriptional noise' because they lack sufficient sequence conservation that is characteristic of a protein-coding gene. It turns out that the secondary structure but not the primary, is evolutionarily conserved and may serve as the major functional unit [8].

Long non-coding RNAs are also known to act in various human-diseases such as cancer. HOTAIR is a lncRNA, that plays a role as an oncogenic molecule in various cancer cells such as breast cancer/colon cancer/gastric cancer and cervical cancer, and also analyzed the expression of various long ncRNAs in various tumors [6].

It has also been observed that long noncoding RNAs can modulate programmed cell death. *Nostoc punctiforme* is the best-studied plant commensal cyanobacterial species. *N. punctiforme* is a heterocyst species capable of promoting growth and colonizing a variety of host plants, including Arabidopsis, suggesting the evolution of plant–cyanobacterial interactions.

Slight observation has been paid to the immunological-dynamics of symbiosis between plants and cyanobacteria. The lipopolysaccharide (LPO) cell wall peptidoglycan in the outer membrane of *Nostoc spp*. Is a typical microbe-associated molecular pattern (MAMP). Although they appear to have no LCO biosynthetic pathway, they can form close relationships with the genera Rhizobia and Frankia. In some studies, they are noticed to form intracellular associations without inducing conflicting responses.

Active inhibition of plant immune-responses, such as programmed cell death (PCD) and *Nostoc* LPO to promote commensal *Nostoc spp.* Colonization, are not recognized as MAMPs by plant receptors [9].

Chapter-02 LITERATURE REVIEW

As already studied, non-coding lncRNAs accounts for a significant percentage of transcripts generated from transcription but does not encodes for any protein. These once formerly was regarded as the 'junk units' but their potential functions have prompted a lot of curiosity, and further discoveries have been made. The two primary members of ncRNAs, firstly, long non-coding RNA and secondly, microRNA they have shown participation in gene expression and other many biological and physiological mechanism which has now been expanded and includes programmed cell death. The interrelated processes and mediators of programmed cell death are greatly modulated by both miRNA and lncRNA, either singly or in conjunction with one another [10].

As studied long non-coding RNAs can control gene expression through variety of methods. To begin, they can either act directly on genome DNA to influence its expression, in order to hire complexes for chromatin modification that can regulate the establishment of heterochromatin, which can repress transcription. Second, lncRNA can bind with protein units, specifically transcription factors and few binding proteins that can indirectly regulate the process of transcription [10].

Programmed cell death, in short called as PCD, is a complex process involving many gene regulation pathways. Programmed cell death (PCD), which is essentially comprised of the processes of (i) apoptosis, (ii) autophagy, (iii) necroptosis, and (iv) ferroptosis, is governed by gene expression. The PCD process can be directed by a number of factors, such as mitochondrial failure, altered protein volume and functionality of macro-molecules components, abnormality in amount of glucose, and oxidative stress [10].

Long non-coding RNAs have significant functions in gene regulation. A detailed understanding of lncRNAs indicates that miR-873 prevents receptor-interacting protein kinase 1/RIPK3 from being translated in cardiomyocytes. These results support the notion that long non-coding RNAs can interact with PCD pathway proteins. Both intrinsic and extrinsic apoptosis pathways can be directed by Caspases, which serve as an important apoptosis mediator. Demise receptors such tumor factor responsible for necrosis (TNFR), mixed lineage kinase domain like_[MLKL], receptor-interacting-protein kinase-3 [RIPK3], receptor-interacting protein kinase-1 [RIPK1], are initiated while caspase activity is blocked, resulting in the creation of necrotic bodies, which results in necroptosis [11].

In research, of Highly Upregulated lncRNAs in Liver Cancer (HULC), it was discovered that miR-9 expression is regulated by HULC, which in turn controls TNF-induced apoptosis. Research of osteoarthritis found that by inhibiting chondrocyte death via HIF1 (Hypoxia-inducible factors 1) and p53 can be done by cutting down the lncRNA reprogramming (ROR). Ferroptosis is primarily brought on by the production of ROS i.e. iron-dependent reactive oxygen species and consumption polyunsaturated fatty acids of plasma membrane. Auto-phagosomes are made by cells under conditions like famine, hypoxia, and hormone signaling. Auto-phagosomes and lysosomes interact to destroy macromolecular junctions and produce macro-autophagy. Autophagy is the most pervasive of his PCD processes [10].

Several apoptotic pathways can be impacted by lncRNAs. For instance, the lncRNA TUG1 suppresses miR-29b37 and modifies the extrinsic apoptotic pathway to prevent apoptosis. Overexpression of lncRNA may reduce the expression of receptors on membrane surfaces. In few cases, Autophagy occasionally causes cell death in addition to helping cells survive. Autophagy can be aided by lncRNAs activating key enzymes. In this regard, MALAT1 overexpression reduced the rate of apoptosis. Also, MALAT1 expression was reduced by the

autophagy inhibitor 3-mA. With the help of these results it is implied that MALAT1 may encourage growth of the cell by preventing cell death by CRC inducing autophagy. Additionally, long ncRNAs can also shield tumor cells from necrosis by preventing the synthesis of a number of associated proteins [10].

Additionally, a number of studies have shown that lncRNAs have an impact on the proliferative assemblage of cancer patients, which is essential for clinical care and prognosis. Long non-coding RNAs in short lnc-RNAs have been linked to medication resistance in malignancies such as breast, stomach, and lung. MALAT1, for example, is a lncRNA that may hinder breast cancer metastasis. Also, MALAT1 overexpression has been linked to development of tumours and metastasis. In a variety of cancers, including breast cancer [10].

Non-coding RNAs are the transcripts that cannot be translated into proteins yet are essential in regulating the proliferation of stable tumors, cardiac dysfunction, and inflammatory and infected tissue. It has been discovered that a numerous non-coding RNAs, specifically long non-coding RNAs and micro-RNAs, bind with PDCD4 (Programmed Cell Death 4) to affect its expression via transcriptional control and to operate as either oncogenes or tumor suppressor [12]. PDCD4, a tumor suppressor is linked to various cellular processes, including biological processes, apoptosis, and direction of many signaling pathways. It also causes inflammation because it is an inflammatory agent. The overexpression of tumor suppressor gene reduces cancer cell invasion, expansion and migration of cancerous cells, induces apoptosis or stoppage of cell cycle, and makes cancer cells more susceptible to anticancer therapies. As a result, aberrant PDCD4 expression levels are associated with the development of numerous diseases [12].

Recently, it has also been shown that non-coding RNAs control the availability of PDCD4 in a multitude of ways, including miRNAs directly aiming for PDCD4 to the 3′-UTR, miRNA-sponges, epigenetic modifications etc. are mediated through lncRNAs on PDCD4. Non-coding

RNAs, such as mi-RNAs and lncRNAs, play a key role in supervising the PDCD4 dosage and it also have the potential to precisely alter PDCD4 expression. CASC15, known as the Cancer Susceptibility Candidate 15, likely to be termed as linc00340, is a highly active long non-coding RNA. Various processes such as melanoma, gastric cancer etc. all have shown increased CASC15 expression. It has the ability to induce cancer development and phenotypic shifting by acting as an oncogene. It has been observed when CASC15 was activated in melanoma cells, her EZH2 can directly bind to the PDCD4 promoter and further block the production of PDCD4 [12]. By cutting the recruitment of downstream molecules like EZH2 and LSD1, it has been discovered that downregulation of long non-coding RNA HOTAIR (GeneID:100124700) stimulates the transcriptional production of PDCD4 in glioma stem cells [12]. Likewise, in esophageal squamous cell carcinoma, epigenetic modification of the Taurine Upregulated gene 1 (TUG1) (Gene ID: 55000), lncRNA, significantly reduced PDCD4 by promoting its EZH2. Numerous miRNAs particularly target PDCD4 to reduce its amount and function, which encourages the emergence and development of tumors [12]. LncRNAs may inhibit PDCD4 expression by luring EZH2 to the gene's promoter and boosting H3K27me3 accumulation there via epigenetic changes. Additionally, lncRNAs can also act as miRNA-sponges, counterbalance miRNA effects on targeted mRNAs [12].

Non-small cell lung cancer, the most prevalent histological form, is the primary cause of cancer-related fatalities globally [13]. It is crucial to find new therapeutic targets for NSCLC due to the constrained efficacy of therapy and adverse effects of currently used medicines. As previously studied, long non-coding RNAs are the transcripts with no coding-potential and carry a length greater than 200 nucleotides which play crucial part in the development & progression of numerous malignancies, including non-small cell lung cancer. The primary mechanism taking place in most of the cancer treatments worldwide is due to the induction mechanism of the

Programmed Cell Death (PCD). Recent studies have demonstrated that non-coding transcripts specifically lncRNAs are related to PCD mechanism and its various types such as ferroptosis, autophagy, necrosis etc. and also have the ability to regulate their death pathways which in response affects the NSC lung cancer progression and its clinical treatment efficacy. PCD pathways of various types are responsible for the origination and progression of NSCLC, and also participate in the abnormal regulation of PCD which result in tumor formation. Apoptosis is a form of programmed cell death which works in two mechanisms i.e. intrinsic pathway and extrinsic pathway which is responsible for cellular destruction. In case, a cell undergoes apoptosis or retains survival depends on a dynamic balance between pro- and anti-apoptotic proteins [13].

The dysregulation caused due to the imbalanced Bcl-2 family members during apoptosis, overexpression of inhibitors of apoptotic proteins (IAPs), caspase downregulation, or inadequate death receptor 10 esemblan has been linked to several types of cancer, including NSCLC. Chemo-resistance and tumor development in cancer are connected with the pathophysiology of NSCL to decreased levels of the pro-apoptotic protein Bax and increased amounts anti-apoptotic protein. Bcl-2 levels of the Patients diseased with non-small cell lung cancer (NSCLC) are found in high concentrations of oncogenic and carcinogenic lncRNAs which are highly upregulated in NSCLC patients, and they may inhibit lung carcinogenesis and cancer progression by blocking apoptosis via modulation of related protein synthesis. It has been reported that highly expressed AFAP1-AS1 has the ability to activate Bcl-2 protein expression and inhibits the apoptotic mechanism along with cell proliferation. In reverse by blocking the expression of the same gene, it decreased the impression of Bcl-2. Another study stated that higher expression of AFAP1-AS1 gene results in large tumor size, high tumor node metastasis etc. and results in critical survival in patients. Higher expression of AFAP1AS1 resulted in lowering cell apoptosis yet is

responsible for increased expansion and migration of lung cancerous cells by inhibiting HBP1 impression when conjugated with LSD1 to HBP1 promoter region. With the help of this context we can conclude that long noncoding RNA MVIH is responsible in the resistance of the drug therapy in NSCLC, but its decreased level restored its sensitivity towards drugs by activating apoptosis with the help of upregulation of Caspase 3/6 and PARP cleavage. Another lncRNA MINCR exerts inhibitory actions on apoptosis in cancer by activation of the oncogene named c-Myc and also regulates the level of Bcl-2 gene, cleavage of PARP and Bax which results in the progression of NSCLC and accelerates the process. ANRIL a highly expressed long non-coding RNA responsible for large tumor size, poor response of patients to clinical therapy etc. It was found that overexpression of ANRIL is also found to be responsible for cell expansion in NSC lung cancer and repression in apoptosis by knocking p21 and KLF2 transcription which directly conjugates with EZH2. lncRNAs that act as tumor suppressors play crucial role in inhibiting and avert lung carcinogenesis and cancer expansion via promoting cell death through alteration of 11 esemblan pathways. Long non-coding RNA MEG3 was seen to be dramatically downregulated in NSCLC, and its up-regulation resulted in apoptosis, anti-tumor properties, escalated impression of the tumor suppressor p53 and diminishes NSCLC development by activating the p53 signaling pathway [5]. But it was also recorded that higher expression of MEG3 could help in suppressing the tumor growth in cancer. Long Non-coding RNAs becoming a target unit has shown great potential in NSCLC. As previously stated, a large amount of long non-coding RNAs are implicated in lung carcinogenesis and show resistance to cancer therapies, implying that if they could be used as predictive-biomarkers of cancer chemotherapeutic sensitivity and clinical care and prognosis. The circulating long non-coding RNAs SOX2OT and ANRIL, which were previously discovered to be considerably over-expressed in the sera of non-small-cell lung cancer patients, has been considered as the suitable biomarker for detecting lung

cancer and predicting overall survival in NSCLC patients. Its exceptional sensitivity and selectivity have the potential to identify cancer patients from healthy controls [13].

Increase in soil salinity is an important environmental factor that is responsible for pernicious effects on world-wide agricultural productivity. Excessive salt-stress is responsible for disruptive normal plant growth and metabolism that induces various other stress conditions such as osmotic, ionic and nutrient stress [14].

Chickpea, scientific name *Cicer arietinum L.* is a leguminous plant which is most regularly cultivated legume and also, is advised as a very beneficial source of fibers and diet proteins. Chickpea earlier was considered as an orphan-crop but now with advanced genomic research resources it has been reported to engage in multiple mechanism responses against various stress conditions [14]. As reported salt-stress condition have the ability to induce osmotic and ionic imbalances which is responsible for the retarded plant growth and metabolism. But according to the new studies made it has been found that by regulating post-transcriptional and transcriptional elements, the plants can endure and revamp under the stress conditions. It has also been marked that lncRNAs can regulate growth and development, and also physiological metabolism by involving in nucleic acid structural modifications, histone modifications, and RNA interactions. The cis-regulating transcripts and eTMs of identified long non-coding RNAs shows involvement in several abiotic stress response [14]. As already remarked, a group of transcription-factors such as *AP2/NAC/ERF* and *WRKY* etc. have important roles in numerous physio-logical processes and they act as plant regulators in various stress responses [14].

With the advancement in the field of research it has been proved that Long Non-coding RNAs are the type of non-coding RNA which also plays role in several plant biological activities such as flowering time, male sterility and nodule development. In exceptional cases it has been

reported that lncRNAs can be species specific. They are usually illustrated at low-levels and indicate no sequence-similarity among species [15].

In various literature it is also mentioned that lncRNAs also participated in complex biological phenomena's such as transcription in gene, gene translation, heat shock response and protein localization & cellular structure integrity. Various of long-ncRNAs have been identified in animals, but only a handful have been discovered in plants [15].

Modern technology, such as next-generation sequencing, makes analysis easier of hundreds and thousands of long non-coding RNAs in model organism i.e. *A.thaliana* which has been reported to be involved in regulation and dictation of flowering-time due to the association of long-ncRNAs such as cold-assisted intronic non-coding RNA [COLDAIR] and cool-assisted intronic non-coding RNA [COOLAIR], responsible for repression in flowering locus C (FLC), and specific male fertility associated RNA (LDMAR), it was discovered to be responsible in photo-period regulated male sterility in rice crop and also in ripening in tomato [15].

Through various studies it is also been acknowledged that Simple Sequence Repeats (SSR) are distributed evenly throughout the entire non-coding and coding regions which are co-dominant and hyper variable. These are micro-satellites bio-markers which carries extremely important and valuable information on the genetic-diversity of the plants. SSRs exhibit a complicated pattern of the occurrence, evolution, function, and mutability. Among them there is further classification of repeats such as mono, di, tri and tetra etc. repeats [15].

It has also been reported that the introduction of the Transposable Element (TE) influences the origin, evolution & function of lncRNAs present in cells. As we already know that LncRNAs are the transcripts with a length of >200 nucleotides. These transcripts can either play role as primary or splicing transcripts independent of their length threshold, for example, lncRNAs like BC1 and snaR found to have a length of less than 200 nucleotides [16]. Literature review reports,

informs the identification of long-ncRNAs in plants like Arabidopsis, wheat, maize, rapeseed, rice and cassava, showing that lncRNAs have a role in a variety of biological processes that contribute to plant development and their ability to respond to challenges. Several studies have also demonstrated the involvement of long-ncRNAs in abiotic-stress, particularly drought-stress, but these studies are sparse in number. Several RNAs such as siRNAs and miRNAs also influence gene expression by a variety of ways, including suppressing gene transcription, degrading the target mRNA, and preventing translation of the target mRNA. Long Non-coding RNAs, on the other hand, affect gene expression differently, either by chromatin modification, or by translation modification [4]. With the help of comparative analysis of lncRNAs it has been revealed that Long Non-coding transcripts show limited conservation between animals and plants [16]. Long Non-coding RNAs as a result of their length, function in a broad spectrum of manners generating a secondary structure that provides a huge number of binding sites for RNA, Protein, DNA and TFs, as well as it has the ability to recruit several types of regulatory components. As acknowledged before with the help of previous studies, lncRNAs can also regulate gene-expression at transcriptional level or post-transcriptional level. In context with transcriptional regulation, lncRNA behaves as a mediator to help in the recruitment of chromatin modifiers, protein activators, transcriptional factors, or repressor proteins for the modification of the chromatin which results in suppression or activation of the target genes and their expression. And it has also been showed that absence of lncRNAs is responsible for the increased number of miRNAs which is also responsible for the suppression in expression of targeted genes [16].

OBJECTIVES

- Identification of Differentially Expressed Genes and Long Non-coding RNAs.
- Functional enrichment of differentially expressed lncRNAs
- Identification of SSRs
- Primer designing for the identified SSRs

Chapter-03 METHODOLOGY

1. DATA COLLECTION

To identify the lncRNAs, SRA data was obtained from NCBI [Accession: PRJNA656261].Our data had total six replicates. Out of which three replicates of *Arabidopsis thaliana* suspension cell cultures were incubated with conditioned-medium (CM) from *N. punctiforme* liquid cultures. The other replicates were control-treatment, cells which are instead incubated with *N. punctiforme* growth-media (BG) (growth medium in which no cells were previously grown). The size of data is approximated 15 GB in zipped format and 80 GB in fastq format.

PRE-PROCESSING OF DATA

2. QUALITY FILTERING AND TRIMMING OF THE DATA RETRIEVED

- FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) seeks to build quality control tests on raw sequences of the data from high-throughput sequencing workflows. It offers a versatile set of analyses that can be used to quickly determine whether data has any concerns that we should be aware of before continue with additional investigation workflow. Also in order to store our raw data for future research the raw reads are converted to FASTQ format [17].
- Trimmomatic was used for read trimming and filtering. However, the quality filtering and adapter sequence identification are the main algorithmic innovations in Trimmomatic. Low quality reads, overlapping sequences and attached adapters were filtered out using Trimmomatic [18].

3. **INDEXING AND MAPPING**

Bowtie2 (https://bowtie-bio.sourceforge.net/index.shtml) is a memory-efficient and quick tool for matching sequencing reads to lengthy reference sequences. Bowtie 2 enables paired-end, local, and gapped alignment modes. In order to verify the overall alignment rate between the two samples with regard to the reference genome, we mapped our two samples data to the reference genome [19].

4. **ESTIMATION OF GENES**

- By conducting Differential gene expression analysis, we found the total number of down and upregulated genes in both the samples.
- In RNA-Seq samples, Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/manual/) assembles transcripts, calculates their abundances, and checks for differential expression and regulation. We used Tool Cufflinks to estimate the counting of fragments that originated from the transcripts and gene in each sample [20].
- Cufflinks provides a cuffmerge script that can be used to join different Cufflinks assemblies of transcripts obtained. Cuffmerge was used to analyze and ensure that the transcriptome assembly retained is distinct and consists of groups of non-overlapping transcripts.

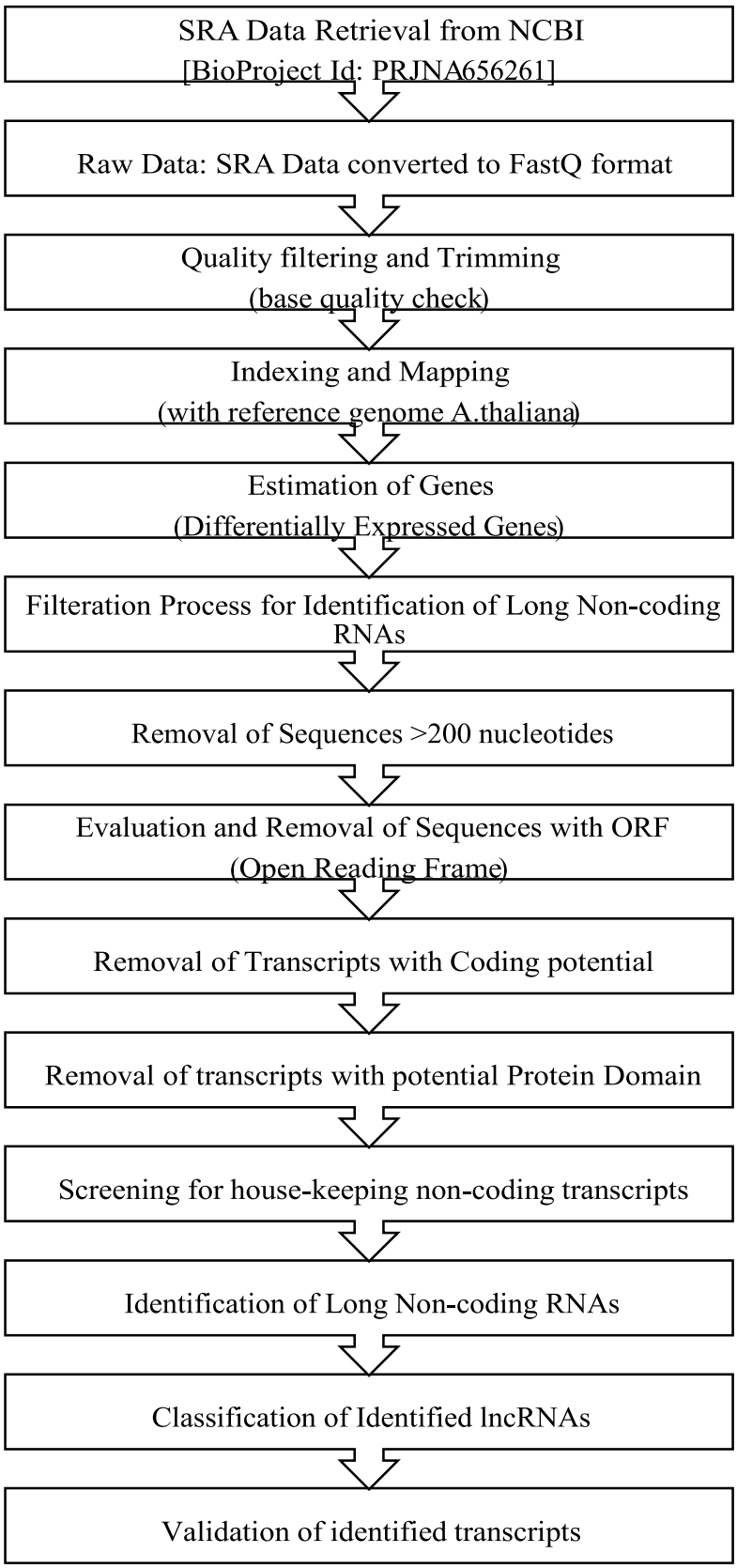

Fig.1 Pipeline for identification of lncRNAs

FILTERATION PIPELINE FOR SIGNIFICANT LONG NON-CODING RNAs

5. IDENTIFICATION OF LncRNAs BASED ON VARIOUS PARAMETERS

ON THE BASIS OF LENGTH

The transcripts obtained with the pre-processing of data were then filtered and the transcripts <200 nucleotides of length were removed. As we already know the long non-coding RNAs are the transcripts having a length of greater than [>200 nucleotides], were filtered out and were selected for further processing.

ON THE BASIS OF POTENTIAL ORF

The ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/), a graphical-analytic tool that locates all open-reading frames in either sequences supplied by the user or a sequence already available in the database. The presence of ORF in a sequence defines that the sequence encodes for some protein [21].

The transcripts showing potential ORF were evaluated and removed with the help of ORF Finder and Transdecoder. Open Reading Frame (ORF) length of >100 amino acid were removed from transcripts.

REMOVAL OF TRANSCRIPTS WITH CODING POTENTIAL

Biological transcripts that are similar with each another are found via BLAST. The program calculates the statistical significance by comparing protein sequences, nucleotide sequences to sequences already present in the databases. BLAST was performed against databases such as UNIPROT and COG in order to filter out transcripts with coding potential.

UNIPROT (https://www.uniprot.org/) is a freely available database which contains of protein-sequences and their functional information.

COG-Database (https://www.ncbi.nlm.nih.gov/research/cog/) stands for Clusters of Orthologous Genes. It reflects the relationships of orthologous groups of proteins.

- **REMOVAL OF SEQUENCES ENCODING FOR PROTEIN DOMAIN**

The obtained transcripts were then utilized for further filtration-process, omitting out the transcripts with coding potential. These transcripts were deployed through CPC2 tool, CNCI and hmm scan in Pfam to locate for sequences which encodes for protein.

CPC2 (http://cpc2.gao-lab.org/) is a quick and precise coding potential calculator based on inherent sequence features.

Coding-Non-Coding Index (CNCI) (https://github.com/www-bioinfo-org/CNCI) effectively distinguishes protein-coding and non-coding regions regardless of known annotations. CNCI is a very effective tool which can be used for the classification of transcripts into sense-antisense pairs [22].

Pfam (http://pfam.xfam.org/) is a collection of a large number of protein libraries and domains. In hmm scan it performs sequence alignment against the domain libraries already present in the Pfam database.

- **PREDICTION OF Long-ncRNAs**

Long noncoding RNA (lncRNAs) predictor PLEK was also deployed to eliminate transcripts with coding-potential. PLEK (https://sourceforge.net/projects/plek/files/) is a fast and accurate alignment-free computational method for distinguishing. Long non-

coding RNAs (lncRNAs) from mRNAs in RNA transcripts of species with no reference genomes.

- **HOMOLOGY SEARCH AGAINST LNC-RNA DATABASE**

Homology search of the differentially-expressed long non-coding RNAs of *A.thaliana* was performed against the known plant long non-coding RNA database, that is, CANTATAdb. CANTATA Database (http://cantata.amu.edu.pl) is a collection of identified Plant Long Noncoding RNAs.

6. **SCREENING AND IDENTIFICATION OF LONG-NON-CODING RNAs**

The identified differentially expressed RNAs were screened for the chromosomal distribution throughout the genome of *A. thaliana*. The graph generated defines the density of all the chromosomes distributed throughout. The screening process also helped to describe the over-expressed and down-regulated identified long non-coding transcripts.

With the above mentioned Homology process the obtained output helped to generate the synteny data which expresses the conservation of long non-coding RNAs to other identified lnc-RNAs of other species available in the database.

7. **IDENTIFICATION OF SSRs**

Simple Sequence Repeats are micro-satellites markers that participate in molecular characterization and is useful for providing information regarding the genetic diversity in plants. The SSRs are evenly dispersed throughout the genome. With the help of Krait tool (https://krait.biosv.com) we identified the SSRs present in the identified long

non-coding RNAs. Using the same we also analyzed the distribution rate and quality parameters etc. of the identified SSRs [23].

8. PRIMER DESIGNING OF THE IDENTIFIED SSRs

Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast) is a tool which is useful to design specific primers for PCR. Primer-BLAST is used to design primers for the identified SSRs which act as biomarkers and may be found useful in further analysis of the long non-coding RNAs [24].

CHAPTER-04 RESULTS

Identification of lncRNAs in *A.thaliana* using available specific transcriptomic information

To identify the long non-coding RNAs present in the genome of Arabidopsis thaliana, we used the transcriptomic data that is available on the NCBI database [Accession number: PRJNA656261]. Our data had a total of six replicates. Out of which three replicates of *Arabidopsis thaliana* suspension cell cultures were incubated with conditioned-medium (CM) from *Nostoc punctiforme* liquid cultures. The other replicates are control treatment, cells which were instead incubated with fresh *N.punctiforme* growth-medium (BG) (growth medium in which no cells were previously grown).

Table 1 Summary of Raw Data used for this study

Condition	SRA ID	No. of Reads	No. of bases	Size in (Gb)
A.thaliana_BG_rep_1	SRR12424063	20,208,158	6.1G	1.7Gb
A.thaliana_BG_rep_2	SRR12424062	20,813,658	6.2G	1.8Gb
A.thaliana_BG_rep_3	SRR12424061	22,850,526	6.9G	2Gb
A.thaliana_CM_rep_1	SRR12424060	23,531,573	7.1G	2Gb
A.thaliana_CM_rep_2	SRR12424059	26,502,573	8G	2.3Gb
A.thaliana_CM_rep_3	SRR12424058	24,883,300	7.5G	2.1Gb

The size of the data obtained was approximately 15 GB in zipped format and 80 GB in fastq format. The raw data obtained was further processed for the quality check analysis of the data.

Table 2 FASTQC Report of raw data i.e., before trimming.

Sample Name	Base sequence Quality	Sequence Duplication level	Overrepresented Sequences	GC% content
A.thaliana_BG_rep_1	PASS	FAIL	PASS	44%
A.thaliana_BG_rep_2	PASS	FAIL	PASS	44%
A.thaliana_BG_rep_3	PASS	FAIL	PASS	44%
A.thaliana_CM_rep_1	PASS	FAIL	PASS	45%
A.thaliana_CM_rep_2	PASS	FAIL	PASS	45%
A.thaliana_CM_rep_3	PASS	FAIL	PASS	45%

After FASTQC-analysis, clean reads were obtained which were used for further analysis and identification of long non-coding transcripts, and etc. After trimming, low quality reads, overlapping sequences, attached adapters were filtered out using Trimmomatic tool. A minimum quality score of >30 was set in order to improve the quality of reads and reliability of data.

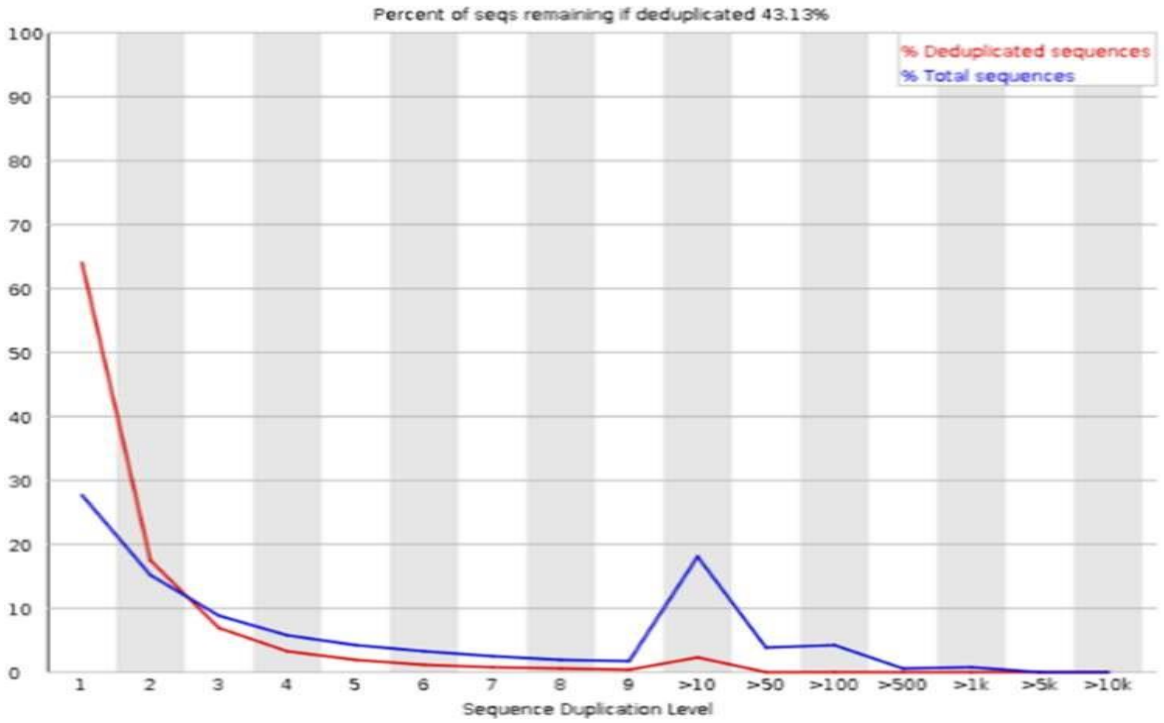

Fig.2 Quality check using FASTQC representing duplication levels in Sequences

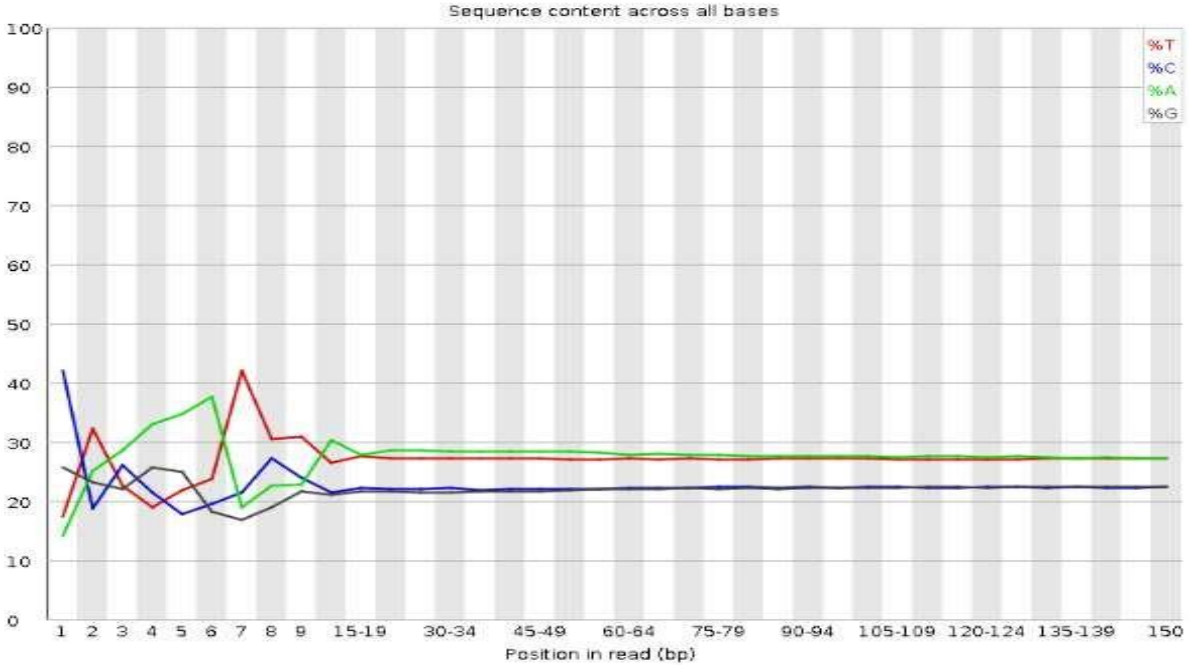

Fig.3 Quality check using FASTQC representing sequence content across all bases

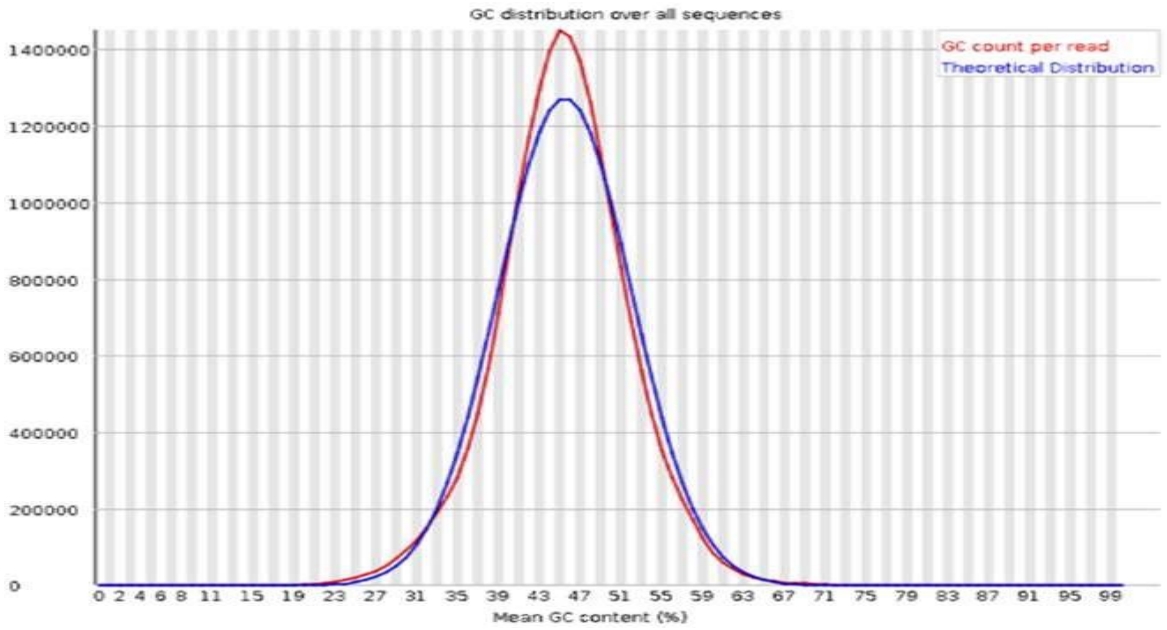

Fig.4 Quality check using FASTQC representing GC content distribution all over sequences

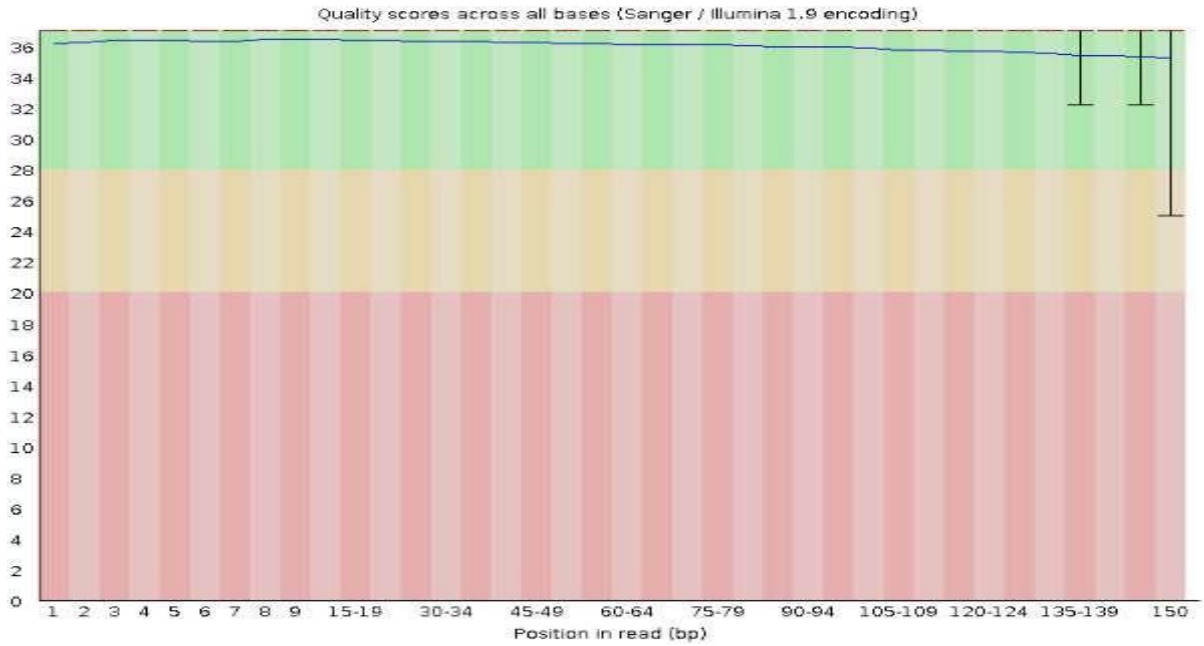

Fig.5 Quality check using FASTQC representing quality scores across sequences

Table 3 Summary statistics of clean data

Name	Total Sequence
A.thaliana_BG_rep_1	5712631
A.thaliana_BG_rep_2	5203414
A.thaliana_BG_rep_3	5052039
A.thaliana_CM_rep_1	6220825
A.thaliana_CM_rep_2	6625643
A.thaliana_CM_rep_3	5882974

The obtained clean data was mapped on the reference genome i.e. *Arabidopsis thaliana* using the Bowtie2 program in order to check the overall alignment of both our samples with the reference genome. The reads from the conditioned media sample (74.83%) successfully mapped with the reference genome. Similarly, the reads from the growth media sample (70.93%) successfully mapped with the reference genome.

Table 4: Percentage of clean reads mapped on to the reference genome.

Sample Name	Overall Alignment Rate (%)
BG_paired.fastq	70.93%
CM_paired.fastq	74.83

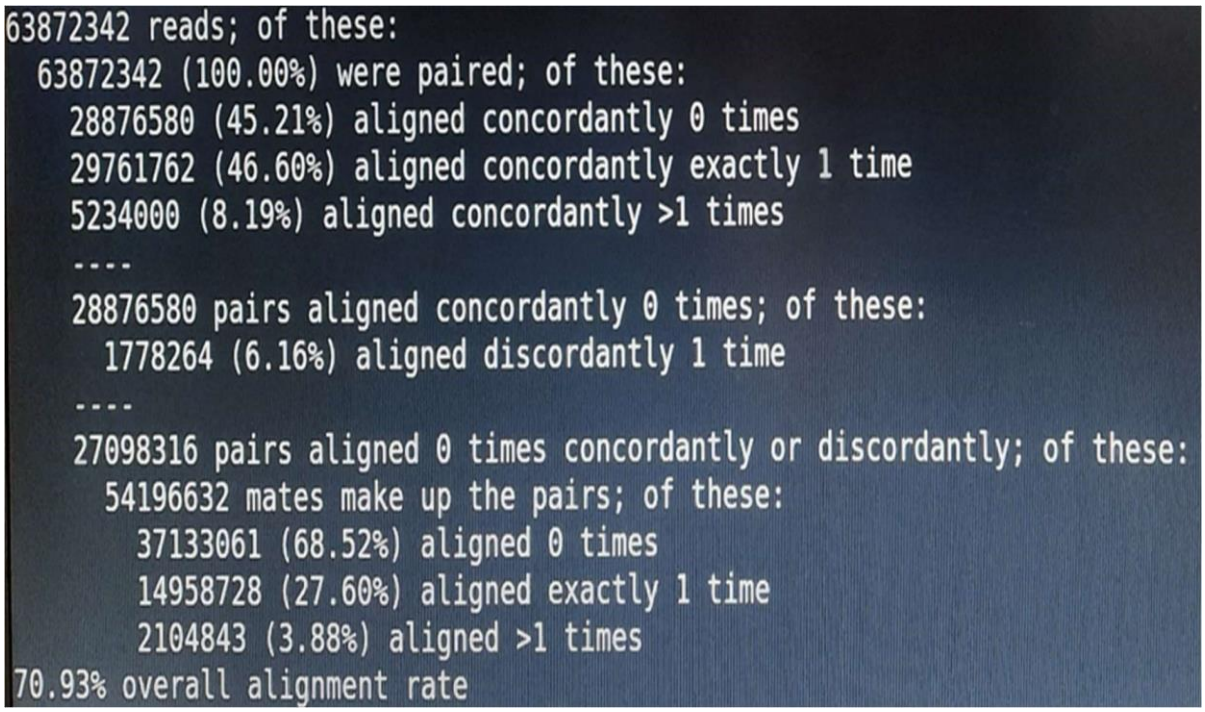

Fig. 6 Mapping results of BG sample with the reference genome

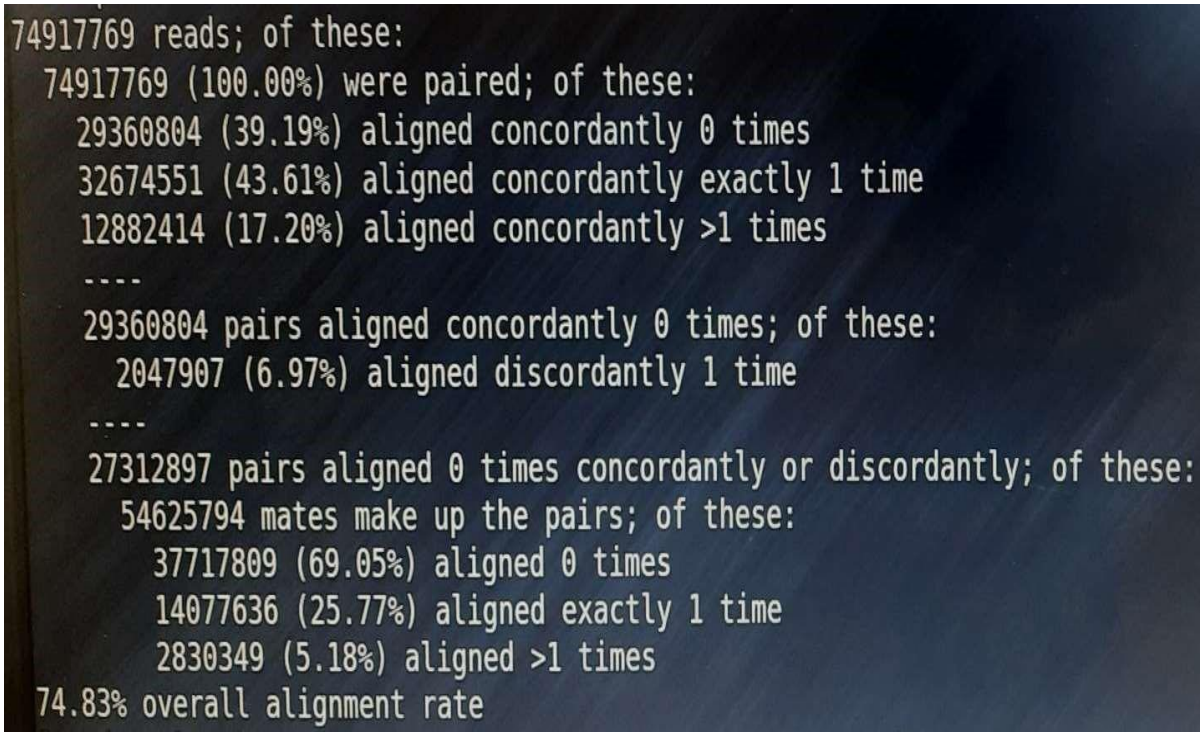

Fig. 7 Mapping results of CM sample with the reference genome

The FPKM value of each gene and its isoforms were calculated using Cufflinks program. This tool was used to estimate the number of fragments obtained through each transcript and gene sample. The cuff-merge tool analysis gave the unique and non-overlapping transcripts. Further, we located differentially expressed genes present in the sample. By performing Differential Analysis we concluded the number of total differentially expressed genes, number of down regulated genes and upregulated genes in both our sample. The total number of significant DEGs obtained was 1284 among which the total 507 were found to be highly expressed and 777 were found to be down-regulated in the sample.

Table 5 Differential Analysis Report showing number of Significant DEGs, Up-regulated DEGs and Down-regulated DEGs

Total DEGs	Significant DEGs	Up-regulated genes	Down-regulated genes
18293	1284	507	777

The total transcripts obtained were processed through a filtration pipeline which distinguished the lncRNAs based on various characteristic parameters. Transcripts in length more than 200 nucleotides were filtered out and chosen for additional processing.

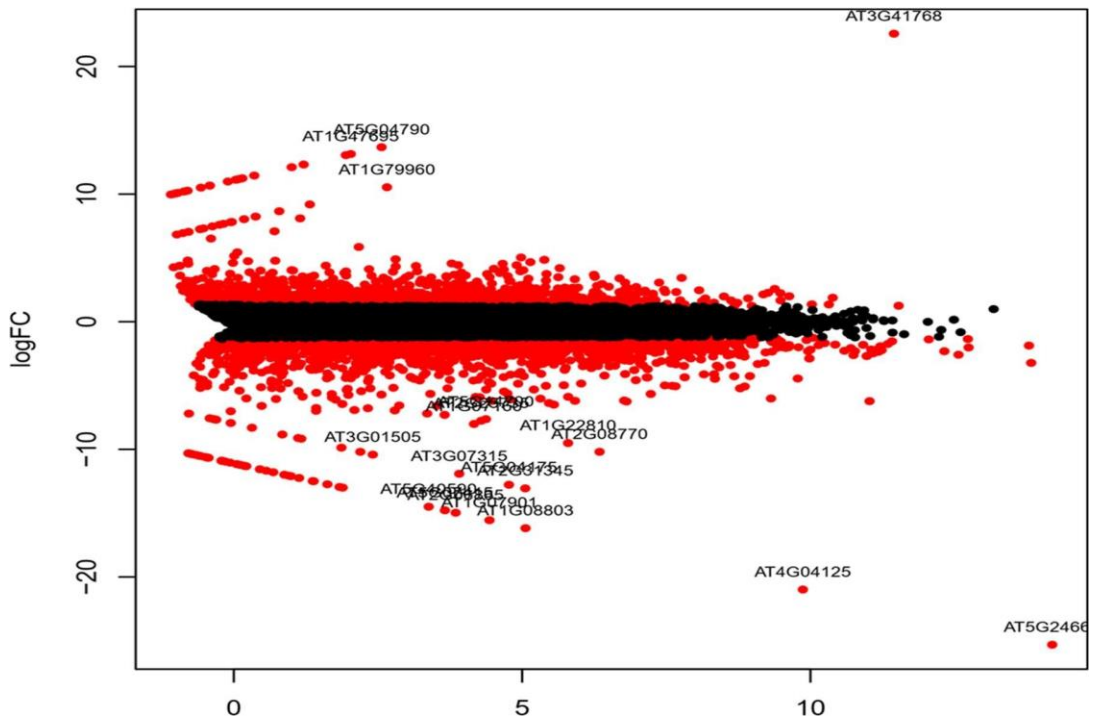

Fig 8: MA plot representing Differentially Expressed Genes between BG and CM conditions

A total of 19171 of transcripts obtained which were processed for the ORF prediction using the ORF finder software. The transcripts consisting of an ORF were excluded and 4601 total number of transcripts were left.

Then using other various tools such as PLEK, CPC2 etc. were used to diagnose the transcripts with coding potential. The results obtained using CPC2 tool, we obtained 660 non-coding transcripts. On the other-hand, using PLEK we obtained a total of 1475 non-coding transcripts. Further calculating both the results we obtained common non-coding transcripts of total of 637.

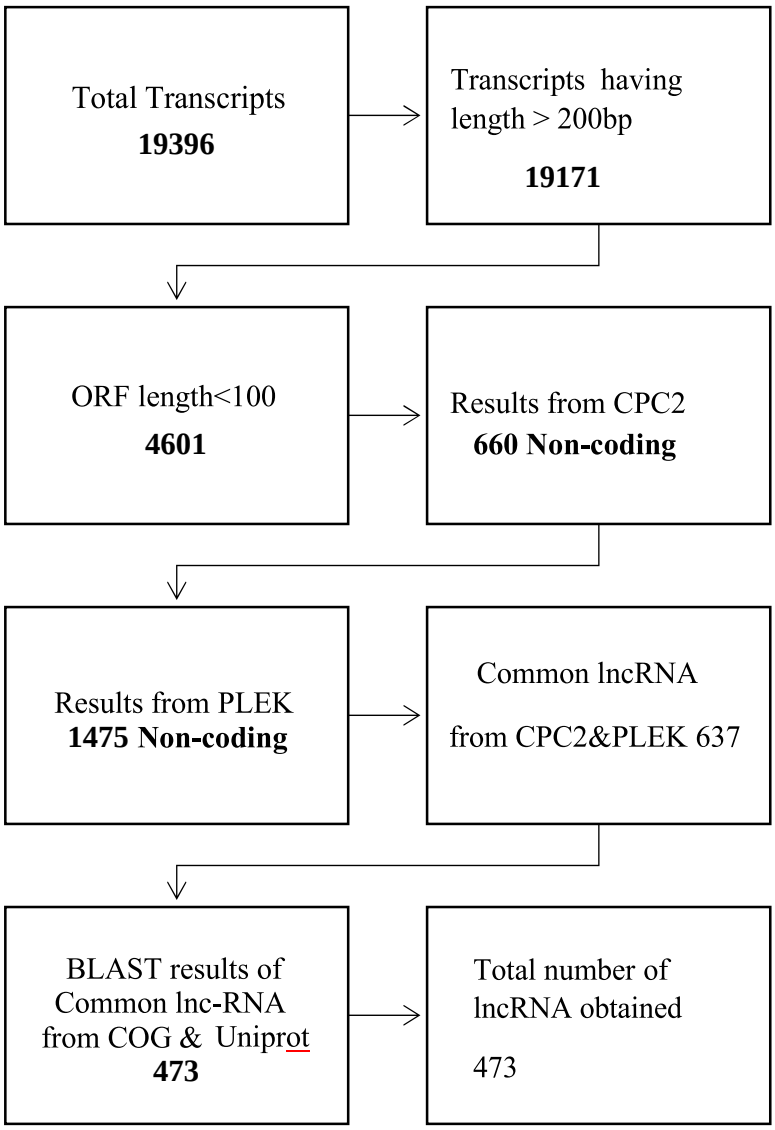

Fig. 9 Flowchart representing the output result through each step of the lncRNA identification pipeline

The obtained common non-coding transcripts were run BLAST against the protein databases such as Uniprot, COG database etc. Also, homology search was performed against the available lncRNA database i.e. CANTATAdb. It was shown that our sample transcripts does not show any resemblance with the already existing lncRNA data in the database which concluded that 473 transcripts obtained are novel long non-coding RNAs.

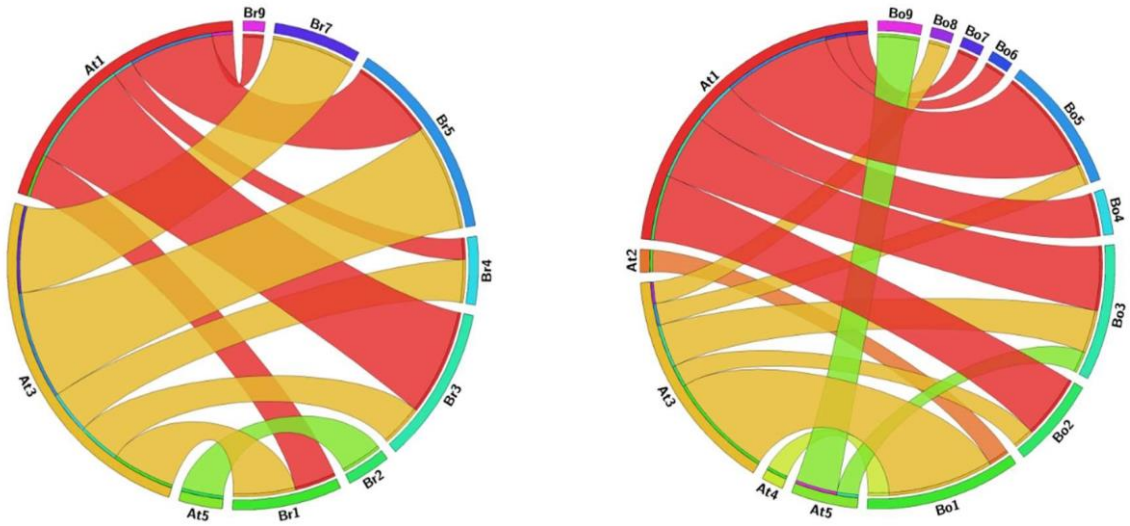

Fig. 10 Synteny Between (a) *A.thaliana* & *Brassica rapa* (b) *A.thaliana* & *Brassica napus*

We performed synteny analysis which showed that *A. thaliana* showed similarity with *Brassica napus* and *Brassica rapa*. Further, the obtained long non-coding RNAs were screened and classified, which displayed the distribution of the chromosomes in the identified long non-coding sequences, the expression level of up-regulated and down-regulated lncRNAs. It was seen that chromosome 5 represents the highest number of lncRNAs contained. While screening it was also analyzed that the total number of Differentially expressed lncRNAs obtained were 94 which showed large number of down-regulated DElncRNAs i.e. 86 (91.5%) and 8(8.5%) up-regulated DElncRNAs.

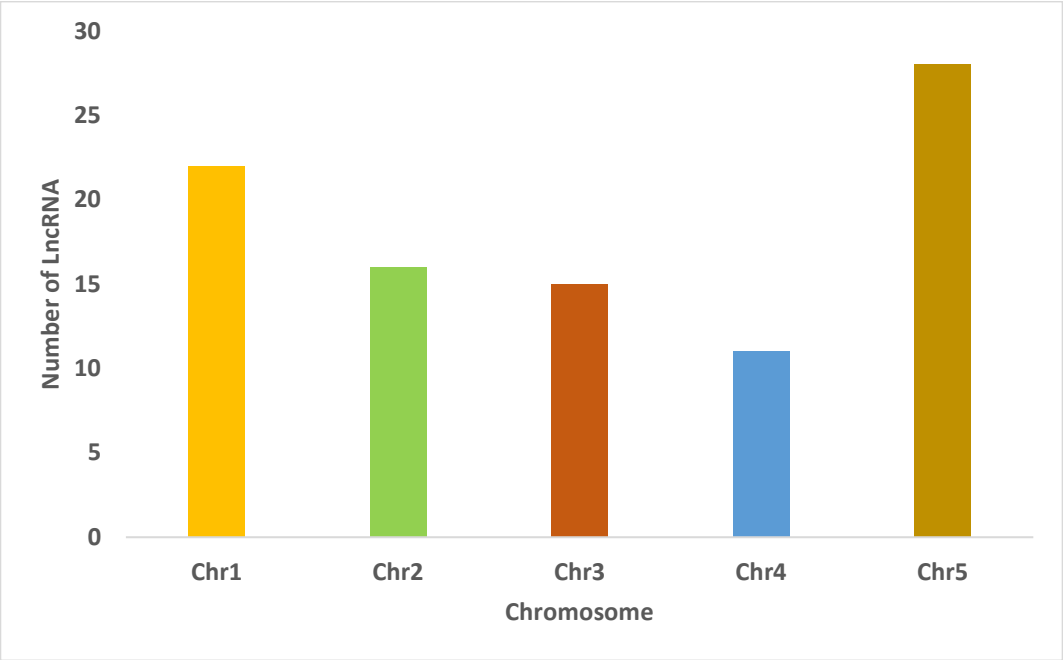

Fig. 11 Chromosomal Distribution of Differentially Expressed lncRNAs

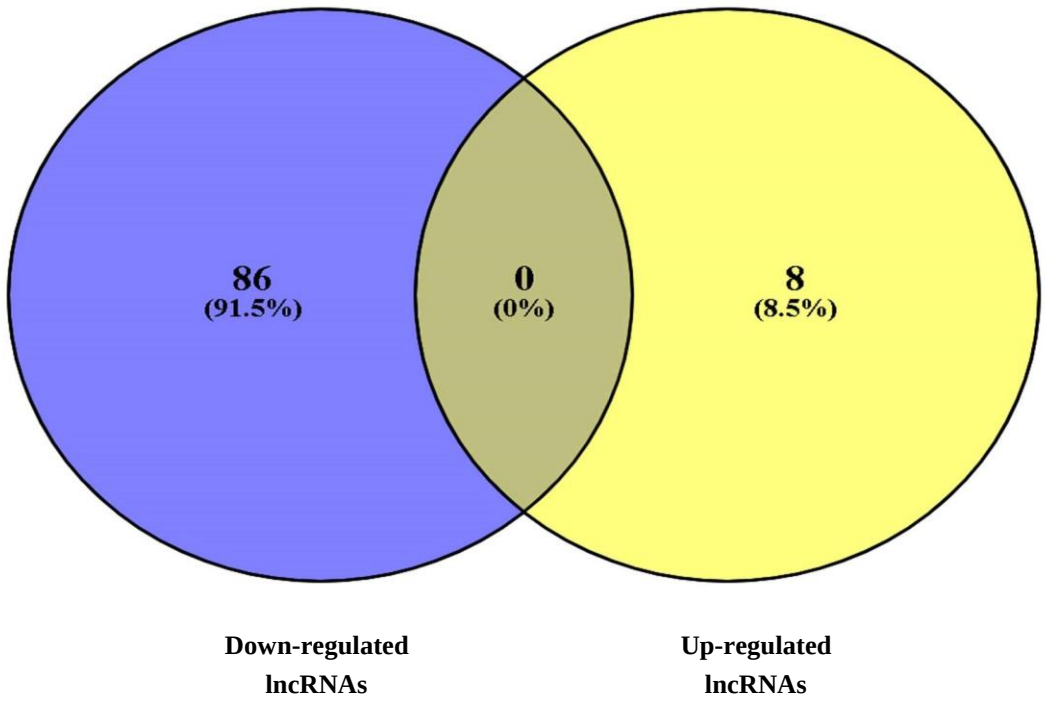

Fig. 12 Venn Diagram showing Significant Up-regulated and Down-regulated lncRNAs

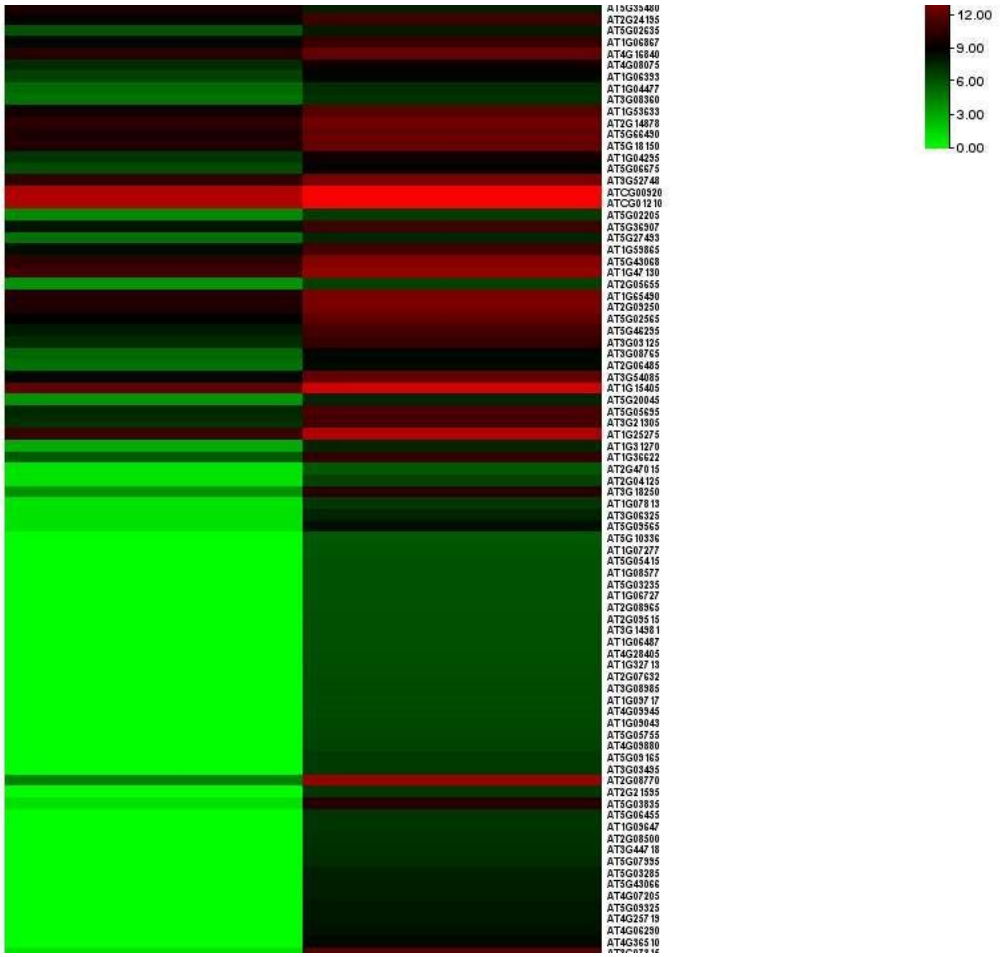

Fig. 13 Heat Map of Differentially Expressed lncRNAs

With the help of krait tool it was found that among the identified 473 long non-coding RNAs, only 44 long non-coding RNAs consisted of SSRs micro-satellites. The statistical report showed their distribution and other quality parameters.

Table 6: List of Identified SSRs

ID	Sequence	Standard	Motif	Type	Repeat	Start	End	Length
4	AT1G08723|AT1G08723.1	AG	TC	2	7	349	362	14
5	AT1G33615|AT1G33615.3	AG	CT	2	11	1032	1053	22
9	AT1G06407|AT1G06407.1	AG	AG	2	7	225	238	14
23	AT3G03825|AT3G03825.1	AG	CT	2	7	289	302	14
27	AT3G60480|AT3G60480.2	AC	TG	2	11	316	337	22
32	AT4G20470|AT4G20470.1	AG	TC	2	7	218	231	14
1	AT1G28250|AT1G28250.1	AAG	GAA	3	6	30	47	18
3	AT1G08723|AT1G08723.1	AAG	AGA	3	6	1	18	18
7	AT1G09993|AT1G09993.1	AAG	CTT	3	6	615	632	18
8	AT1G09247|AT1G09247.1	AAG	TCT	3	5	38	52	15
11	AT1G28395|AT1G28395.3	AAG	AAG	3	5	721	735	15
12	AT1G52550|AT1G52550.1	ATC	TGA	3	5	477	491	15
13	AT1G77880|AT1G77880.2	AAG	GAA	3	5	67	81	15
14	AT1G09713|AT1G09713.1	AAG	TTC	3	9	1	27	27
15	AT1G09297|AT1G09297.1	AAG	TTC	3	6	313	330	18
16	AT1G80890|AT1G80890.1	AAG	AGA	3	16	14	61	48
17	AT3G04285|AT3G04285.1	AAG	AAG	3	6	350	367	18
18	AT2G08425|AT2G08425.1	AAG	TTC	3	7	3	23	21
20	AT2G06485|AT2G06485.1	AAG	AAG	3	6	153	170	18
21	AT2G38450|AT2G38450.2	ACC	CCA	3	5	218	232	15
24	AT3G07568|AT3G07568.1	AAG	GAA	3	5	52	66	15
25	AT2G42395|AT2G42395.1	AAC	AAC	3	5	386	400	15
26	AT4G08785|AT4G08785.2	ACC	GGT	3	5	404	418	15
28	AT5G66490|AT5G66490.1	AAG	GAA	3	5	251	265	15
31	AT4G20470|AT4G20470.1	AAG	TTC	3	5	187	201	15
33	AT5G03210|AT5G03210.1	AAC	TTG	3	5	301	315	15
34	AT4G04055|AT4G04055.1	AAG	TCT	3	5	96	110	15
39	AT4G32590|AT4G32590.2	AAG	TCT	3	6	188	205	18
42	AT4G08230|AT4G08230.2	ACC	GGT	3	5	186	200	15
43	AT4G08035|AT4G08035.1	AAG	AAG	3	6	769	786	18
35	AT3G52748|AT3G52748.1	AAAG	TTTC	4	4	345	360	16
40	AT5G06445|AT5G06445.1	AAAG	TCTT	4	4	262	277	16
41	AT5G43068|AT5G43068.1	AAAAG	GAAAA	5	4	370	389	20
29	AT5G06235|AT5G06235.1	AACGGT	ACGGTA	6	4	522	545	24

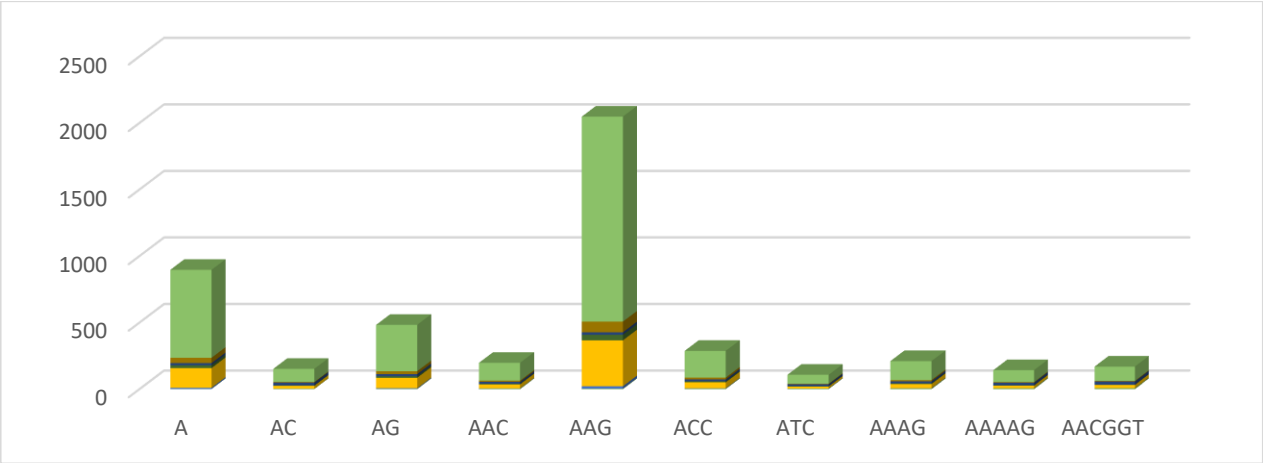

Fig.14 Graph displaying abundant motif categories

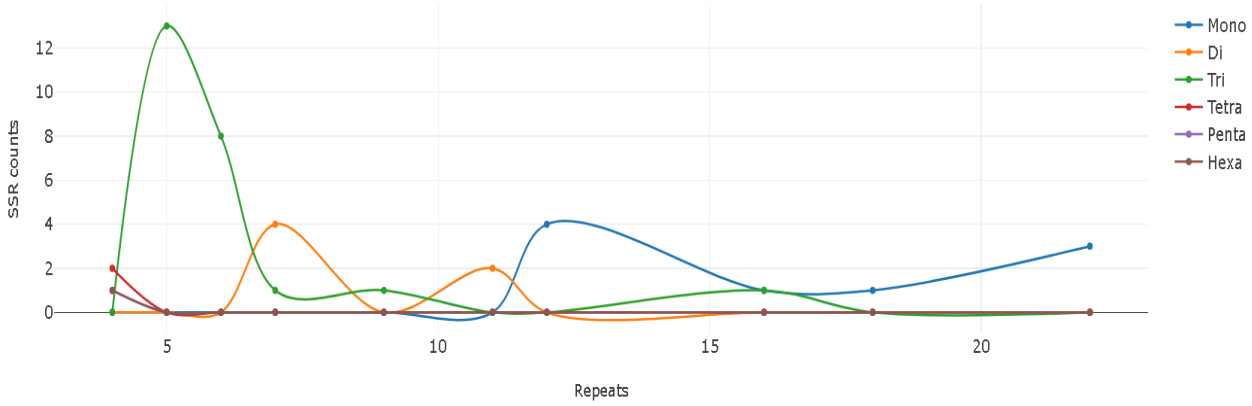

Fig.15 Graph representing the distribution of SSRs for each type

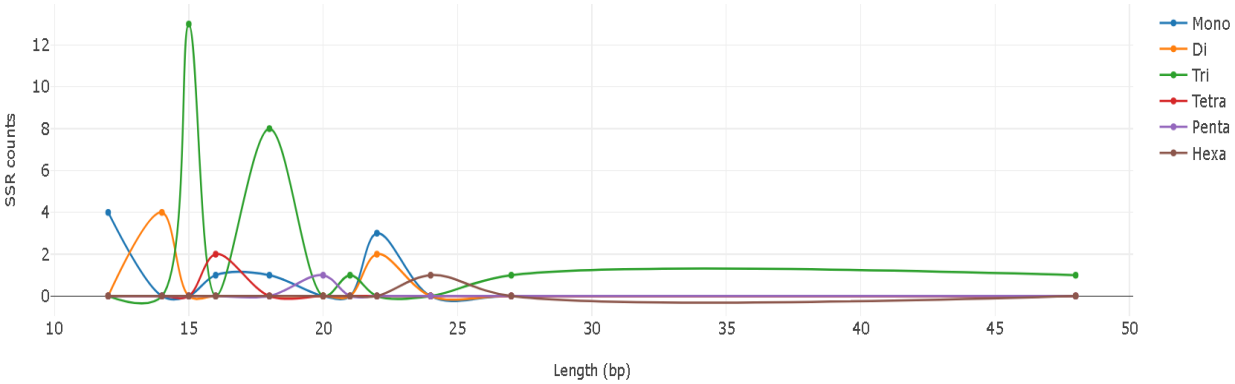

Fig. 16 Graph displaying SSRs length distribution for each type

The statistical report generated displays the abundance of tri-nucleotide micro-satellites present in the long non-coding RNAs.

Table 7 Primer Designed for di, tri, tetra, penta and hexa SSRs

Gene Id	SSR motif	Type	Primer Sequence	Primer length	Tm	GC %
AT3G03825	CT	**Fwd**	AGTCCCCCAAGCATTCAAAGA	21	59.57	47.62
		Rev	CCCTCGAGAGTCGCATACAC	20	59.97	60
AT1G52550	TGA	**Fwd**	CCATCGCTGCCGTAAAAACC	20	60.18	55
		Rev	GGAGTGGGGACGAGGAAAAG	20	60.04	60
AT3G52748	TTTC	**Fwd**	TAAGACCCTCCCCACCATGT	20	59.88	55
		Rev	GGAGGTGGTCTTGGAATGGG	20	60.03	60
AT5G06445	TCTT	**Fwd**	CAGCAGCATCGTGCAAACTA	20	59.2	50
		Rev	GGTGCAAGAAGCCGATGAAAC	21	60.4	52.38
AT5G43068	GAAAA	**Fwd**	GCCATGATCAATCGTCGGTG	20	59.42	55
		Rev	GCTCTTCCCCAAAGGACGTT	20	60.25	55
AT1G06407	AG	**Fwd**	CATGCGGCTCTGCAAATGAT	20	59.62	50
		Rev	AAACATGCGATGCCGATTGC	20	60.53	50
AT1G80890	AGA	**Fwd**	AAGCTCGAAAAGCTCGTCTCT	21	59.73	47.62
		Rev	CGAAACTCACAGCTCACACA	20	58.43	50
AT4G04055	TCT	**Fwd**	GCGAGGAGCTGTTGAAGAGT	20	60.04	55
		Rev	GCCTCACGGGTCTCTCTATT	20	58.6	55

CONCLUSION

In conclusion, our research identified the long non-coding RNAs (lncRNAs) in model plant *Arabidopsis thaliana* with the help of Next Generation Sequencing technology. We have effectively characterized a considerable number of unique and differentially expressed long non-coding RNAs.

By using advanced NGS tools and methodology, we annotated and classified total 473 differentially expressed lncRNAs along with their expression values in two different media conditions. It is concluded that the identified lncRNA transcripts are mostly down-regulated in the provided conditions. Furthermore, we have also identified 44 SSRs containing long non-coding RNAs and also designed primers for these SSRs.

REFERENCES

1. Y. Wan and K. Chatterjee, “RNA,” Encyclopedia Britannica. 29-Aug-2022.

2. D. Wang and A. Farhana, “Biochemistry, RNA Structure,” PubMed, 2021. https://www.ncbi.nlm.nih.gov

3. “RNA- Properties, Structure, Types and Functions | Molecular Biology,” Microbe Notes, Jan. 14, 2020. https://microbenotes.com/rna-properties-structure-types-and-functions

4. Wikipedia Contributors, “RNA,” Wikipedia, Feb. 25, 2019. https://en.wikipedia.org/wiki/RNA

5. C.-Y. Guh, Y.-H. Hsieh, and H.-P. Chu, “Functions and properties of nuclear lncRNAs—from systematically mapping the interactomes of lncRNAs,” Journal of Biomedical Science, vol. 27, 1, Mar. 2020, doi: 10.1186/s12929-020-00640-3

6. Q. Tang and S. Hann, “HOTAIR: An Oncogenic Long Non-Coding RNA in Human Cancer,” Cellular Physiology and Biochemistry, vol. 47, 3, pp. 893–913, 2018, doi: 10.1159/000490131.

7. X. Zhang et al., “Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels,” International Journal of Molecular Sciences, vol. 20, 22, Nov. 2019, doi: 10.3390/ijms20225573.

8. Y. Su et al., “Regulatory non-coding RNA: new instruments in the orchestration of cell death,” Cell Death & Disease, vol. 7, 8, pp. e2333–e2333, Aug. 2016, doi: 10.1038/cddis.2016.210.

9. M. Zhao et al., “The Regulatory Role of Non-coding RNAs on Programmed Cell Death Four in Inflammation and Cancer,” Frontiers in Oncology, vol. 9, p. 919, Sep. 2019, doi: 10.3389/fonc.2019.00919.
10. N. Jiang, X. Zhang, X. Gu, X. Li, and L. Shang, “Progress in understanding the role of lncRNA in programmed cell death,” Cell Death Discovery, vol. 7, 1, Feb. 2021, doi: https://doi.org/10.1038/s41420-021-00407-1.
11. Y. Su et al., “Regulatory non-coding RNA: new instruments in the orchestration of cell death,” Cell Death & Disease, vol. 7, 8, pp. e2333–e2333, Aug. 2016, doi: https://doi.org/10.1038/cddis.2016.210.
12. M. Zhao et al., “The Regulatory Role of Non-coding RNAs on Programmed Cell Death Four in Inflammation and Cancer,” Frontiers in Oncology, vol. 9, p. 919, Sep. 2019, doi: https://doi.org/10.3389/fonc.2019.00919.
13. Y. Luo et al., “Targeting lncRNAs in programmed cell death as a therapeutic strategy for nonsmall cell lung cancer,” Cell Death Discovery, vol. 8, 1, Apr. 2022, doi: https://doi.org/10.1038/s41420-022-00982-x.
14. N. Kumar et al., “Genome-wide identification and functional prediction of salt- stress Related long non-coding RNAs (lncRNAs) in chickpea (Cicer arietinum L.),” Physiology and Molecular Biology of Plants, vol. 27, 11, pp. 2605–2619, Nov. 2021, doi: https://doi.org/10.1007/s12298-021-01093-0.
15. B. Kumar et al., “Genome-Wide Identification of Long Non-Coding RNAs in Pearl Millet (Pennisetum glaucum (L.)) Genotype Subjected to Drought Stress,” Agronomy, vol. 12, 8, p. 1976, Aug. 2022, doi: https://doi.org/10.3390/agronomy12081976.

16. A. Das et al., “Expressivity of the key genes associated with seed and pod development is highly regulated via lncRNAs and miRNAs in Pigeonpea,” Scientific Reports, vol. 9, 1, Dec. 2019, doi: https://doi.org/10.1038/s41598-019-54340-6
17. “Babraham Bioinformatics-FastQCAQuality Control tool for High Throughput Sequence Data,”Babraham.ac.uk,2019.http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
18. A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” Bioinformatics, vol. 30, 15, pp. 2114–2120, Apr. 2014, doi: https://doi.org/10.1093/bioinformatics/btu170.
19. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” Genome Biology, vol. 10, 3, p. R25, 2009, doi: https://doi.org/10.1186/gb-2009-10-3-r25.
20. C. Trapnell et al., “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,” Nature Biotechnology, vol. 28, 5, pp. 511–515, May 2010, doi: https://doi.org/10.1038/nbt.1621.
21. I. T. Rombel, K. F. Sykes, S. Rayner, and S. A. Johnston, “ORF-FINDER: a vector for highthroughput gene identification,” Gene, vol. 282, 1–2, pp. 33–41, Jan. 2002, doi: https://doi.org/10.1016/s0378-1119(01)00819-8.
22. L. Sun et al., “Utilizing sequence intrinsic composition to classify protein-coding and long noncoding transcripts,” Nucleic Acids Research, vol. 41, 17, pp. e166–e166, Jul. 2013, doi: https://doi.org/10.1093/nar/gkt646.
23. L. Du, C. Zhang, Q. Liu, X. Zhang, and B. Yue, “Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design,” *Bioinformatics*, vol. 34, 4, pp. 681–683, Oct. 2017, doi: https://doi.org/10.1093/bioinformatics/btx665.

24. J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden, “Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction,” *BMC Bioinformatics*, vol. 13, 1, Jun. 2012, doi: https://doi.org/10.1186/1471-2105-13-134.