GENOME-WIDE IDENTIFICATION OF miRNAs AND lncRNA OF *VANILLA PLANIFOLIA*

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

Master of Science in Biotechnology

By

Anuja Bharmaik

225111005

Under the Supervision of:

Dr. Shikha Mittal

(Assistant Professor)

DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HIMACHAL PRADESH

MAY, 2024

DECLARATION

I hereby declare that the work reported in the M.Sc. Biotechnology project entitled **"Genome-Wide identification of miRNAs and lncRNA of *Vanilla planifolia* "** submitted at Jaypee University of Information Technology, Waknaghat, H.P, India is an authentic record of my work carried out under the supervision of **Dr. Shikha Mittal**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my M.Sc. Project report.

Anuja Bharmaik

Enrollment Number: 225111005 Date:
Department of Biotechnology & Bioinformatics
Jaypee University of Information Technology
Waknaghat, India – 173234

CERTIFICATE

This is to certify that the work reported in the **M.Sc.** project report **"Genome-Wide identification of miRNAs and lncRNA of *Vanilla planifolia*"** submitted by Anuja Bharmaik at **Jaypee University of Information Technology**, **Waknaghat, H.P, India**, in the year from August 2023 to May 2024. It is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Signature of Supervisor: ______________________

Dr. Shikha Mittal Date:
Assistant Professor

Department of Biotechnology & Bioinformatics
Jaypee University of Information Technology
Waknaghat, India-173234

ACKNOWLEDGMENT

I would like to express my profound gratitude to my guide **Dr. Shikha Mittal** for his guidance, support, and constant encouragement throughout this project work. She has been more than just my project guide; at times a mentor to rescue me out of my doubts. She has always helped me to work hard and also taught me how to implement different ideas to deal with the problem. Moreover, she taught me to not give up and many other valuable lessons. Furthermore,

I would like to acknowledge Vice-Chancellor Prof. **(Dr.) Rajendra Kumar Sharma**, Prof. **(Dr.) Ashok Kumar Gupta**, Dean of academics & research for providing me with an opportunity to be a part of the institute and to complete my Master's Degree.

I also want to mention the HOD of Biotechnology and Bioinformatics **Prof. (Dr.) Sudhir Kumar** has been a source of immense motivation and inspiration both for my academic and personal life. He was never, and I know will never be, more than just a phone call away. He has helped me in almost every aspect I have asked him for. In addition, I would like to thank all the faculty members of the BT/BI Department of JUIT, who have helped me whenever I needed, and also would like to thank all the lab engineers especially **Ms. Somlata Sharma** for providing me with a workplace and for always motivating me.

I would also like to appreciate the part that my classmates (Kritika, Akshita, and Abhimanyu) and PhD scholar Priyanka mam, have played in shaping this project work. They have been my constant support and cheered me up at hard times. They helped me whenever I had any doubts. Thanks a lot! I would like to thank the almighty God for his grace throughout my life. Last but not least I would like to thank my mother and Father who have always supported me through thick and thin and have been a constant source of encouragement and support; also, who has never given up on me and always motivated me.

[Thanks to JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY]

Anuja Bharmaik
M.Sc. Biotechnology
JUIT, Solan

TABLE OF CONTENTS

1]. Declaration -- i

2]. Certification -- ii

3]. Acknowledgment -- iii

4]. List of Tables -- v

5]. List of Figure -- vi

6]. Abstract -- vii

7]. Introduction -- 2-5

8]. Review Of Literature -- 7- 17

9]. Material and Methodology -- 19- 36

10]. Results -- 34-42

11]. Conclusion -- 44-45

12]. References -- 46-48

List of Tables

Table no.	Title
Table 4.1	Summary of raw data collection from NCBI.
Table 4.2	FASTQC output of raw data.
Table 4.3	Descriptive statistics to clean data.
Table 4.4	Statistical analysis of Novel lncRNAs identification from various protein databases after duplicates and coding potential sequences removal.

List of figures

Figure no.	Title
Fig 1.1	*Vanilla Planifolia Plant*: Mature Pods and Blossoming Flowers.
Fig 2.1	miRNA inhibiting translation by degrading mRNA.
Fig 2.2	Role of miRNAs and lncRNAs in plant gene regulations.
Fig 3.1	Workflow for miRNA identification in Vanilla planifolia.
Fig 3.2	Output repeats derived consistently from the Repbase database.
Fig 3.4	Alignment against miRBase showing zero-overall alignment rate.
Fig 3.5	workflow for lncRNA identification in Vanilla planifolia.
Fig 4.1	The assortment of duplications among sequences using FASTQC.
Fig 4.2	Quality check for the sequence through the FASTQC program with all the bases.
Fig 4.3	Using FASTQC quality check representing quality scores across sequences.
Fig 4.4	Mapping results of a sample SRR1171643 with the reference genome.
Fig 4.5	Mapping results of a sample SRR1171644 with the reference genome.
Fig 4.6	Mapping results of a sample SRR1171645 with the reference genome.
Fig 4.7	Secondary structure prediction of a microRNA generated by RNA secondarystructure prediction tool RNAfold.
Fig 4.8	psRNATarget result for miRNA target prediction.
Fig 4.9	MA plot representing Differentially Expressed Genes.
Fig 4.10	Heatmap of expression Levels of Genes (HPP92) Across Different samples.

Abstract

RNA Sequences of an organism encode for a huge number of the ncRNAs as compared to protein-coding RNAs, small nucleolar RNAs, rRNAs, tRNAs, and small nuclear RNAs are expressed in tissues, and stage independence regulatory ncRNAs shown to be expressed in both stage and the tissue-specific manner including, small interfering RNAs, microRNAs, and long non-coding RNAs (lncRNAs). Although ncRNAs do not code for proteins they are involved in genetic regulation processes. The ncRNAs appear to be involved in regulation at transcriptional, post-transcriptional, and at least epigenetic levels and to play a role in a network of complex methylation-controlled processes.

The discovery of miRNAs and lncRNAs (lncRNAs) in the genome of *V. planifolia*, a significant orchid, is the initial foundation for the discovery of phenotypic developments, such as root formation, flowering, and even for the production of metabolites. This work involved the usage of bioinformatics techniques to detect as well as examine miRNAs, and lncRNAs which were the genetic origin of *Vanilla planifolia*. Our study showed the occurrence of the same set of miRNAs and lncRNA moreover, genes of these miRNAs and lncRNAs were studied to address their implication for different biological processes including flower development and vegetative bud development, reproductive bud, and mixed bud and scent generation. We obtained 27 novel miRNAs and 22 lncRNAs in vanilla by filtering raw data taken from NCBI (Bioproject- SRA data) leads to the extraction of transcripts and prediction of miRNAs by performing homologous sequence identification by performing BLASTn and BLASTx from which we obtained 40 sequences that were later analyzed using CPC2 with the extraction of noncoding RNAs. For lncRNA transcripts were filtered with FPKM<0.5 and <200nt length, after homology search using ORF finder, removal of coding potential transcripts, performing BLASTx against different protein databases like NCBI non-redundant protein database, Swiss-Prot database, and COGs databases. Obtained 22 novel lncRNAs.

Generation of the genome-wide miRNAs and lncRNAs of *V. planifolia* not only helps to extend the understanding of the complex genetic regulation mechanisms and pathways that drive the specific characters of the economically significant orchid but also serves as a critical resource in the future.

CHAPTER -1

INTRODUCTION

1.1 Introduction

Genome-wide identification studies, also known as GWAS or genome-wide screening, typically involve the analysis of a genome-wide set of genetic variants in more than one individual to establish if any of the variants are linked to behavior. These investigations are usually very helpful for understanding the association between the qualities including illnesses, physical attributes, genetic differences, and biochemical properties in the domains of genetics and genomics. The studies on the genome-wide identification of the genetic basis of many traits and diseases have been a great breakthrough in the way we understand the genetics of these traits and diseases. Through the studies of the interactions between genes and the environment, we can get to the point of understanding the not-so-simple concept of personalizing medicine, the creation of new and better crop varieties, and the complete description of biological processes. Nevertheless, they also have a big downside, which is data analysis and interpretation, thus requiring further research in the field of genomics.

Vanilla planifolia, known as vanilla orchid and often known for the flat-leaf shaped vanilla, is a member of the Orchidaceae family and a botanical wonder valued for its economic value and importance. In the genus, vanilla has about 110 species usually found in tropical areas [1][2].

Fig 1.1 *Vanilla Planifolia* Plant: Mature Pods and Blossoming Flowers.

Ref: https://img.freepik.com/premium-photo/vanilla-vanilla-orchids-botanical-illustration-white-paper-best-medicinal-plants-their-e_508524-810.jpg?w=1060

It is a perennial climbing plant immensely rich in the chemical profile which is primarily characterized by the presence of compound ‘vanillin’, a principal compound known for its characteristic scent. The Spanish and Portuguese brought this plant to Asia and Africa in the 16th century from the Gulf Coast of Mexico (the chief producer of vanilla) [3]. Commercial vanilla i.e. *Vanilla planifolia* Andrew is mostly known for its distinctive aroma and wide usage in flavors, perfumes, cosmetics, and medicine. Despite of centuries of utilization, much remains to be uncovered about the molecular pathway that underpins various qualities of the plant. To farmers, scientists, and environmentalists, nothing has ever been so curious than a Vanilla planifolia vine the most ubiquitous flavor on Earth. A range of subjects have been investigated because of this complicated orchid’s baffling life cycle and labor-intensive requirements; agriculture, genetics, botany, and ecology among them.

This study focuses on finding small and long non-coding RNAs in Vanilla planifolia. It aims to help us learn more about Vanilla planifolia by looking at its genetic material. This includes focusing on RNAs that are not involved in making proteins, like long non-coding RNAs and miRNAs. This area of research in plant genetics provides us with information on how plants respond to stress, and what other processes they go through. Similar studies in other plants may provide a good starting point, although there is little known about these molecules of Vanilla planifolia. For example, many microRNAs, which do different things, such as controlling growth and helping plants cope with stress, have been identified in studies of model plants and important crops 4] [5].

Although there hasn't been as much information published on the discovery and characterization of these compounds across the whole genome in Vanilla planifolia, comparable research offers a strong basis. Extensive studies conducted on model plants and commercially significant crops have unveiled a multitude of miRNAs implicated in many tasks, including growth control and adaptability to stress. RNA interference mechanisms, including microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), are the key biological functions in higher plants. These are some of the most basic functions that small regulatory microRNAs and long noncoding RNAs perform at a transcriptional and post-transcriptional level. They are majorly involved in gene regulation mechanisms, developmental processes, stress adaptation mechanisms, disease resistance, etc.

MicroRNAs are small RNA regulatory molecule non-coding proteins ranging from 18-24 nucleotides in length and many are often conserved in evolution [5]. These non-coding microRNAs play a multifaceted function in transcriptional regulation at developmental processes and in response to abiotic stresses and biotic [5].

These small RNAs have been also demonstrated to have very potent regulatory functions in cells, the main role here is to inhibit the production of specific genes, which is accomplished by targeting mRNAs for degradation or translational repression [6]. Not only it, but it's to be found that it stimulates the representation of the certain lncRNAs, too. For example, in inducing plant biology, miRNAs play a key role as gene expression regulators that contribute to the control of almost everyone aspect of plant development and physiology like root and shoot development, flower and leaf morphogenesis, hormone responders, and way of attacking the stresses [7].

An example is pigeonpea, where miRNAs have been associated with many droughts tolerance-related traits suggesting potential interconnectedness within Vanilla planifolia for resilience-informed breeding [8]. The researchers identified specific miRNAs that were turned on or off during drought. This shows how miRNA profiles change in response to the environment. Learning about these processes in pigeonpeas can help us understand how to breed better, more resilient crops and it also gives clues about how to deal with challenges like drought in other crops like vanilla. Beyond just surviving droughts, though, miRNA might direct all kinds of vanilla growth stages - the way it does with those pigeon peas and a better understanding miRNA's role could mean better crops overall and new ways to tackle environmental growing challenges [8].

lncRNAs, however, can have functions that convert to small miRNAs and also function as sponges to miRNA that has inhibitory properties. Different mechanisms including the growth of cancer in mammals and possibly identical ones in plants may be simply triggered by this interaction [9][10]. Genes encoding lncRNAs are multifaceted and can act during epigenetic process and gene transcription, and transcriptional, hence influence plant development and stress response. Some recent research has also considered the MLIs (microRNA-long non-coding RNA interactions) linked to plant life cycle regulation as a possible risk for the target gene expression during some plant diseases studies. The investigation of these relationships has progressed rapidly with the creation of diverse classifiers aimed at detecting the links between miRNAs and lncRNAs, e. g. LncMirNet, and LMI-DForest [11].

The complexity and diversity of miRNAs and lncRNAs in plant species have been emphasized, with studies underscoring their critical roles in gene expression and epigenetic regulation [12]. This knowledge is crucial for using ncRNAs to improve crops and understand how plants have medicinal properties. *Vanilla planifolia* is a good example of this because it is used in both cooking and medicine [3]. Studying the genes of spice crops like *Vanilla planifolia* is also important for the food and pharmaceutical industries [10]. New technologies like high-throughput sequencing and bioinformatics have played a big role in finding miRNAs and lncRNAs. Applying this knowledge makes it possible to study non-coding RNAs in plants, which can be further applied to studying *Vanilla planifolia* or any other plant. There are also databases, like miRBase, that have a lot of information about miRNAs. These resources are helpful for comparing different species and understanding how ncRNAs control genes in *Vanilla planifolia*.

To date number of previous research have significant information about the chemistry, cultivation, and pharmacology of *Vanilla planifolia* but however not yet explored the significance of these non-coding RNA classes of this essential plant. Our research further delves deeper into better comprehending the intricate network of miRNAs, lncRNAs, and interactions between the genes, that contribute to *Vanilla planifolia* 's remarkable trait with the identification of novel non-coding RNAs. Unlocking noncoding RNA potential might not stop at hardier vanilla vines. These molecules likely also produce the enticing flavors and therapeutic compounds that make orchids so useful. Global demand for spice crops keeps rising, so genomic spice research matters economically too. The food and pharmaceutical industries could benefit from customized vanilla strands tailored through miRNA studies. Overall, noncoding RNAs offer largely untapped tools to enhance crops, meet vanilla product demands, and reveal the science underlying this orchid's charms.

Objectives

- To identify potential miRNAs and lncRNAs and their target genes in *Vanilla planifolia* using homology search.
- Annotation of miRNA and lncRNA target genes to understand the role of miRNAs in plant development.

CHAPTER -2

REVIEW OF LITERATURE

2.1 Review of literature

Vanilla planifolia is a highly valued species when it comes to issues of economics and culture because of its flavoring agents. There are other varieties of vanilla on the market, such as *V. pompona* and *V. tahitensis*, but they are not as valued because their beans are not as good. The *Vanilla planifolia* plant, where the highly sought-after vanilla beans are sourced, offers a wide range of benefits that go beyond its well-known culinary applications. Serving as the main source of natural vanilla flavor, these beans are a fundamental ingredient in desserts, pastries, and various savory and sweet dishes, imparting not just a deep, intricate taste but also a variety of aromatic compounds that elevate the dining experience. Apart from its culinary uses, *Vanilla planifolia* possesses notable medicinal properties traditionally; it has been used in herbal medicine to alleviate anxiety and depression symptoms, owing to its calming scent and potential to promote relaxation.

Additionally, vanilla extract is also been widely recognized for its antioxidant properties, which aid in combating free radicals and reducing the oxidative stress, thus providing protective effects against the cell damage and aging. Economically, vanilla beans are a valuable commodity in the global market, mainly due to the labor-intensive procedures involved in their cultivation and processing, including hand-pollination and an extended curing process to develop their distinctive flavor. The high demand for vanilla, combined with the meticulous production methods, contributes to its reputation as one of the most expensive spices worldwide, offering significant economic value to regions where it is produced. Consequently, the *Vanilla planifolia* plant emerges as a versatile and advantageous crop, valued for its culinary flexibility, therapeutic benefits, and substantial economic impact.

The chemical compounds that are majorly responsible for flavor (vanilla flavor) vary among species of vanilla and also the geographical origin. Vanillin, which is a primary flavor compound, is biosynthesized in the pods and is immensely more consolidated in the *Vanilla planifolia* as compared to other species of vanilla [3].

Research has identified specific genes, like caffeoyl CoA O-methyltransferase-like genes, that could be involved in the biosynthesis of vanillin, providing insights into the metabolic pathways of flavor production.

Gene changes in *V. planifolia* were demonstrated using small pieces of the shoot tip, showing how the technology can help improve vanilla growth A good method of gene modification with PLBs of shoot tips has been done, which can help in the rescue and improvement of *V. planifolia.* Significant progress has been made in the cryopreservation of *V. planifolia* shoot tips, which is crucial for the conservation of genetic resources [2]. Studies have compared different cryogenic techniques for vegetative growth, showing that certain methods, such as the D-cryoplate, result in stronger regenerates with less polymorphism.

The discovery of non-coding RNAs particularly lncRNAs and microRNAs (miRNAs) in *Vanilla planifolia* is crucial for unraveling the intricate gene regulatory mechanisms that drive the growth, development, and production of secondary metabolites in this economically significant plant. miRNAs and lncRNAs plays a vital role in the post-transcriptional gene expression regulation, exerting influence over a wide array of biological processes.

Moreover, the examination of miRNAs and lncRNAs in this plant (*Vanilla planifolia*) has the potential to make significant contributions to the field of plant molecular biology. By shedding light onto the evolutionary conservation as well as the diversification of these non-coding RNAs in plant species, it can provide fresh perspectives and insights. Additionally, can serve as a valuable resource for genetic markers in plant breeding and the genetic engineering endeavors, specifically aimed at enhancing the quality and yield of vanilla crops.

To summarize, the discovery and analysis of miRNAs and lncRNAs in *Vanilla planifolia* hold immense significance in advancing our knowledge of the molecular genetics of this plant. This breakthrough opens a new avenue for studying gene regulation, which in turn may have substantial implications for improving vanilla cultivation, boosting the production of secondary metabolites, and ultimately contributing to the economic value of the vanilla industry.

We can now use Next generation sequencing (NGS) technologies to identify and analyze all microRNAs (miRNAs) and long non-coding RNAs(lncRNAs) from the whole genomes of various species making it possible to understand the functions of these small molecules in plant development, metabolic processes and responses to environmental cues. *V.*

planifolia is a member of the orchid family known for its thick, succulent stems and broad pointed leaves. Stern and Judd (1999) extensively studied the stems, leaves, and roots of different vanilla species, including *V. planifolia* and *V. pompona*, to show what makes them special. *V. planifolia* is the main source of true vanilla flavor and its dried pods are used in food, fragrance, and beauty products.

MicroRNAs, also known as miRNAs, are a group of small RNA molecules that do not code for the proteins and are essential for controlling gene expression in both plants and animals. These molecules are majorly 20-24 nucleotides long and work by binding to specific sequences on the target messenger RNAs (mRNAs), which can result in either mRNA degradation or inhibition of translation. The discovery of miRNAs has greatly enhanced our comprehension of the intricate gene regulatory networks in plants, emphasizing their significance in processes such as development, responses to stress, and various physiological functions [17].

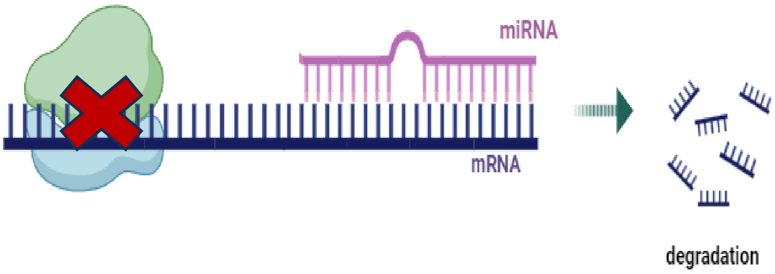


Fig 2.1 miRNA inhibiting translation by degrading mRNA.

As miRNAs are non-coding in nature and function by binding to the complementary sequences on the target messenger RNA transcripts this binding typically leads to translational repression or degradation of the target mRNA as shown in Fig 42.1. In *Vanilla planifolia,* the identification and functional analysis of miRNAs could provide valuable insights into their involvement in controlling the biosynthesis of vanillin, the primary compound responsible for the distinct vanilla flavor, as well as other valuable secondary metabolites. Understanding these regulatory mechanisms could pave the way for targeted approaches to enhance vanillin production, ultimately increasing the commercial value of the vanilla crop.

The finding or identification of let-7 and lin-4 in *Caenorhabditis elegans*, which served as the basis for initial understanding of the function and importance of small RNA molecules as gene expression regulators, paved the way for the research for miRNAs in plants. The later research work was mainly on the plant miRNAs, which resulted in the discovery of hundreds of miRNAs of various plant species [19]. Above all, J. Carrington and his team have carried out the research which is the main reason for discovering the functions of these non-coding miRNAs in plants, especially the developmental timing and pattern formation.

MiRNAs are part of several regulatory pathways, among which the ones controlling the shape of the leaves, the formation of the flower, the root formation, and the plant's response to the environmental stress and the pathogen attack [20]. The miRNAs are therefore the actors in different control systems, like those that control the formation of flowers, leaf shape, growth of the roots, and the attack of the pathogens and the responses to the environment [21]. Many of such discoveries have paved the way for the improvement of the agricultural practices, for example, crop engineering which can increase the stress resistance, yield, and quality of the crops. In short, miRNAs are the vital components of the gene regulatory network, which is very important for the correct functioning and adapting of plants. The current research and development of new sequencing techniques are still confirming the complex functions of miRNAs, which will be very helpful in solving the problems in agriculture and the science of plants.

The finding of miRNAs, which are the small non-coding RNA molecules that are very important in controlling the gene expression, has been immensely helped by the progress of bioinformatics analysis and molecular biology techniques. At very first, the miRNAs were found by the cloning the small RNA fractions and then further cloning the sequences into vectors for the sequencing. Nevertheless, this method was majorly characterized by its manpower-intensive nature and could only be able to identify a few miRNAs at once [21].

Next-generation sequencing (NGS) technologies, especially small RNA-seq, have made it possible to discover miRNAs in a completely new way. The small RNA-seq technique enables the fast and simultaneous analysis of miRNAs from different species, tissues, and developmental stages. This method of sequencing millions of small RNA molecules together at once allows for the complete spectrum of miRNA expression to be profiled and the identification of new miRNAs to be made. The sequencing process is then proceeded by the application of bioinformatics tools that can eliminate the non-miRNA sequences, predict the novel miRNAs based on the sequence and structural features, and the miRNA expression levels [22]. Besides, computational prediction methods have become an essential part of miRNA discovery which is the third arsenal aside from NGS. The used methods are the ones that apply the algorithms for scanning genomic sequences for the hairpin structures which are the usual miRNA precursors. Computational approaches are able to forecast possible miRNA genes by studying the conservation of sequence among species and predicting the secondary structure of miRNA precursors [23]. Nevertheless, the predictions made by computational tools normally need experimental verification which is usually done through Northern blotting, quantitative RT-PCR, or in situ hybridization and thus they could not confirm the expression and function of the predicted miRNAs.

Microarray technology is also the main tool for the discovery and profiling of miRNA. The miRNA microarrays consist of probes that are made to detect the known miRNAs, thus generating the miRNA expression patterns in different conditions. Although microarrays do not help much in the discovery of new miRNAs, they are very useful in the profiling of known miRNAs and their regulatory functions [25].

lncRNAs (Long non-coding RNAs) have come to the fore as the key authors of gene regulation in different organisms. While their coding RNA equivalents that translate into protein, lncRNAs do not and instead affect the gene's expression by the means of different ways, like transcription control, chromatin modification, and post-transcriptional processing [26]. In plants, lncRNAs have been proven to be the regulators of a wide range of roles in processes like developmental processes, including, among others, physiological and flowering time regulation, stress responses, and developmental patterning. The work done before has greatly improved the knowledge about the complexity of gene regulation in plants, thus, uncovering another level of gene control apart from the previously known ones.

Scientific findings into the plant lncRNAs have quickly increased, revealing their functions in abiotic and biotic stress responses, development, and signaling. In other words, it overall indicates that lncRNAs can also serve as miRNA sponges, which thus, determines the availability of miRNAs to the target mRNAs, which further in turn, has an impact on plant development, growth, and stress responses [27]. The functional characterization and discovery of lncRNAs in plants are still in progress, and on the other hand, more and more lncRNAs are being identified thanks to the new high-throughput sequencing technologies and bioinformatic analytics.

The recognition of long non-coding RNAs has become a lot better with the advancement of genomic technologies, mainly due to high-throughput RNA sequencing (RNA-seq). RNA-seq has become the main tool for the detection of lncRNAs, giving the sensitivity and specificity that are not possible with any other RNA detection method over the genome [28]. This method sequences the transcriptome in a sequence manner; hence, it is able to capture diverse RNA species, including those that were not previously annotated or classified as non-coding. The RNA-seq technique is the most thorough one, which makes it the perfect instrument for the quantitative and qualitative analysis of lncRNAs, hence, it is the most important tool for their discovery. Before the RNA-seq era, the tiling arrays method which was used for the sequencing of fixed genomic regions to identify transcribed sequences was applied, but this method lacked the resolution of RNA-seq and was dependent on the genomic sequences that were already known.

Besides the RNA-sequencing, chromatin signature mapping has been the main tool in the lncRNA discovery. The methods like Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) in the case of specific histone modifications related to active transcription, it show the genomic regions that are actively transcribed, including those that have been discovered for lncRNAs [29]. This way, even though it is not very direct, it gives us a lot of information about the transcriptional background and the possible regulatory functions of lncRNAs.

Computational predictions have also been very important in the identification of lncRNAs, they have been used for the analysis of the specific exon-intron structures and the conservation of these structures across species. On the other hand, the computational approaches are strong, yet, the validation by the experiments is necessary for the confirmation of the existence and the function of the predicted lncRNAs [30]. In recent years, scRNA-seq (single-cell RNA sequencing) has been developed as a powerful tool for the discovery of lncRNAs at the cell level, hence, it has helped the researchers to know the cellular heterogeneity and which role lncRNAs play in the cell.

The disadvantage of this method is that it can find cell-type-specific lncRNAs, thus, it can help in understanding their functions in development and disease [31].

In a nutshell, the exploration of lncRNAs was driven by the combination of the aforementioned modern genomic and bioinformatics techniques. The investigation of lncRNAs has been powered by the synergy of experimental and computational approaches, each providing unique insights into the complex scenario of the non-coding genome. The combination of these methods of study keeps on enriching our knowledge about lncRNAs, thus uncovering their different functions in life processes and diseases.

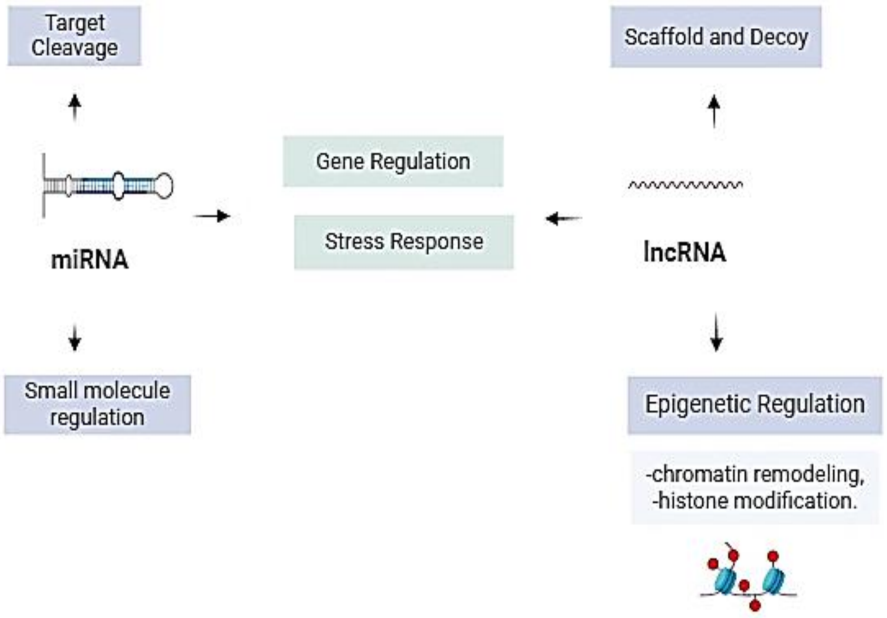


Fig 2.2 Role of miRNAs and lncRNAs in plant gene regulations.

The main points of the figure (fig2.2) are the similarities and differences in the miRNA and lncRNA profiles. The comparison of lncRNAs between G. hirsutum and G. barbadense shows that lncRNAs have a lower sequence conservation rate and a significant number of the genomes contain transposable elements which affect the miRNA and lncRNA genes.

In the vertebrates, miRNAs control the important reproductive processes, a function that is to some extent of the non-mammalian species indicates that a fundamental role of this exists among different taxa. The investigation of lncRNAs in rice immunity indicates that the species-specific responses to the pathogen in rice differ from the broader roles in mammalian innate immunity, thus, the research demonstrates the functional diversities.

Evolutionary Insights, the participation of transposable elements in the evolution of non-coding RNAs in cotton indicates a distinct evolutionary pathway where TEs are the regulatory signals for lncRNA genes and "sponges" for miRNAs. The conservation of lncRNAs in *A. thaliana* and their systematic annotation show that the biological functions of lncRNAs have been maintained even during the long period of evolution so, they give us a clear idea about the evolutionary conservation of the ncRNA functions. Applications in crop improvement and breeding strategies are just some of the areas in which this interdisciplinary science is widely used. The discovery of miRNAs and lncRNAs in plants such as *A. thaliana* and rice has made it possible to apply them in crop improvement, especially through techniques like STTM, which enable the regulation of miRNA accumulation in conditions of stress [31].

Information from the conservation and functionality of lncRNAs in different species enables us to understand their possible roles in crop resilience and productivity, which means that these ncRNAs could be a source of future breeding strategies.

A. thaliana, a plant model organism that is extensively studied in plant biology, has been the subject of much research in the characterization as well as identification of miRNAs. miRNAs are tiny non-coding RNA molecules that are vital for gene regulation at the post-transcriptional level. They achieve this by targeting the specific mRNAs for cleavage or translational repression. The research of miRNAs in *A. thaliana* has given us important knowledge about how they regulate various biological processes, for example, development, stress response, and disease resistance.

Computational Prediction and Experimental Validation: The discovery of miRNAs in *A. thaliana* has usually been done by a mixture of computer modeling and experimental validation methods. The first study in this area was carried out by Reinhart et al. [32]. They used a bioinformatics technique to find miRNA candidates in the *A. thaliana* genome based on some structural and sequence characteristics. The researchers demonstrated several miRNA candidates and experimentally verified their expression and target mRNA cleavage using Northern blot analysis and the 5' RACE (rapid amplification of cDNA ends) experiments.

Jones-Rhoades and Bartel [33] have the computational pipeline for miRNA prediction in *A. thaliana* that was also improved by Jones-Rhoades and Bartel. They looked for hairpin structures in the genome which could be the beginnings of miRNA, then they filtered the results based on the sequence conservation across plant species and the possible target sites in the mRNAs. As a result, the precise miRNA candidates were then experimentally confirmed by small RNA sequencing and target mRNA cleavage assays.

High-Throughput Sequencing and Advanced Bioinformatics: The emergence of high-throughput sequencing techniques has made the identification of miRNA in *A. thaliana* more complete and faster. Ng et al. [34] used the technique of small RNA sequencing and the most modern bioinformatics tools to discover new miRNAs in *A. thaliana* under different stress situations. They created a machine-learning technique to determine miRNA precursors by their structural and sequence features and then tested the expression of selected miRNAs using quantitative real-time PCR (qRT-PCR). The research they did showed that there are several miRNAs that are stress-responsive and have important functions in the regulation of the stress responses of plants.

Just like this, Karlova and his team [35] used high-throughput sequencing and computational analysis to find out the miRNAs that are involved in the regulation of the flowering time in *A. thaliana*. The scientists found out, among others, several miRNAs that affect the regulators of the flowering time, like FLOWERING LOCUS T (FT) and APETALA2 (AP2), and proved their functions through genetic and molecular studies. Computational predictions are a good thing for miRNA identification, but experimental validation is a must for confirming their expression and the biological functions they have. Different methods have been used for miRNA confirmation in *A. thaliana*, so far, the Northern blot analysis, quantitative RT-PCR (qRT-PCR), and in situ hybridization. The functional characterization of miRNAs generally refers to the analysis of target mRNA cleavage or translational repression using methods such as 5' RACE, RNA-induced silencing complex (RISC) immunoprecipitation, and proteomics techniques. Moreover, genetic studies that had as parts miRNA overexpression or knockdown lines have shed some light on the functions of certain miRNAs in the different biological processes in *A. thalian.*

Using bioinformatics techniques, researchers discovered 616 unique miRNAs in *Cajanus cajan,* of which 118 originate from different families. From these, the 578 were expressed in matters not previously reported in MirBase21.As a result, 1373 target genes with 180 miRNA sequences were discovered. Among these, 298 of the target genes were studied at the protein level. Furthermore, 3919 long non-coding RNAs (lncRNAs) eventually been identified, and 87 of these lncRNAs were the targets of 66 miRNAs that were predicted. These miRNAs and lncRNAs are the key performers in the regulation of growing yield, quality, and tolerance to stress in *C. cajan* [36].

Rice (*Oryza sativa*) is the mainstay food source and a crucial model for the research on monocot plant biology. Through the process of lncRNA identification and characterization in rice, researchers have been able to obtain knowledge about the probable roles they play in different biological processes, for example: growth, development, and stress responses. A groundbreaking survey of the lncRNAs in rice was made by Ding et al. [37]. They used a multi-step method involving high-throughput sequencing, computational analysis, and experimental confirmation to discover and describe lncRNAs in rice. Their study involved the following steps:

The authors of the study collected strand-specific RNA-seq data from different rice tissues, such as seedlings, roots, leaves, and panicles. The RNA-seq data were pieced together into transcripts, and the longer RNA gene candidates were identified by the exclusion of the known protein-coding transcripts, the small non-coding RNAs, and the other annotated RNA classes.

The coding capacity of the lncRNA candidates was evaluated using computational tools such as the Coding Potential Calculator (CPC) and the Coding-Non-Coding Index (CNCI) to detect them from the possible protein-coding transcripts. The expression levels and patterns of the lncRNAs that were identified were analyzed in different tissues and developmental stages, thus, the evolution of their possible functional roles was revealed.

The researchers studied the gene sequence conservation of the found lncRNAs in different rice subspecies and related plant species to get the data on their evolutionary conservation. Experimental validation: Some lncRNAs that were chosen were experimentally validated employing qRT-PCR and RNA-FISH. The former was used to confirm their expression and the latter to verify their localization. Ding et al. by using this complete method identified and characterized more than 2,000 lncRNAs in rice, many of which had the tissue-specific and the developmental stage-specific expression patterns, which can be the indicators of their involvement in the biological processes.

These studies prove that combining high-throughput sequencing technologies, computational analysis, and experimental validation is crucial for the precise identification and characterization of lncRNAs in crop plants like rice. The results from these research activities are of great value in the study of lncRNA-mediated gene regulation and lncRNA-mediated gene regulation in economically important crop species.

Ultimately with this knowledge, we can say that the studies should scrutinize the exact functions of the discovered miRNAs and lncRNAs within post-transcriptional gene regulation, which may appear to help improve the plant yield of the orchid species, *Vanilla planifolia.* It is noteworthy that additional research must be done to completely understand the regulatory pathways with the discovered miRNAs, lncRNAs, and their target genes. These networks provide a thorough highway into molecular mechanisms that control plants' multiple development and stress response pathways. Possibly in next stage of investigation is to focus on marker-assisted selection and its wide application to carve the future with breed improvements, for example, better flavor, increased yield, and resistance to both biotic and abiotic stresses [39].

The explanatory analysis between *Vanilla planifolia* with its relatives might let us figure out how orderly the system of the miRNA and lncRNA functioning in terms of the evolutionary principles and we also can get broad-based genomic insight into the special "DNA" of *Vanilla Plantifolia*. Combining miRNA and lncRNA analyses with the omics layers like proteomics, transcriptomics, and metabolomics will reveal non-coding RNA functional aspects and their regulatory structure within the *V. planifolia* orchid.

The emerging research areas mentioned give us a chance to expand our understanding of the noncoding RNAs and the Vanilla species in interest and then exploit this knowledge not only for developing sustainable crop improvement but also for improving agriculture as a whole. It illustrates how important it is to implement and develop more research and studies in these areas in order to get to know completely the role of these non-coding RNA in *Vanilla planifolia* and for the future of agriculture and sustainable crop production.

CHAPTER - 3

MATERIAL AND METHODOLOGY

3.1 Material and Methodology

This analysis requires a Linux operating system (with a 64-bit version and at least 4 GB RAM and sufficient storage) to ensure smooth working.

Raw data retrieval i.e. raw reads from an RNA-seq experiment in FASTQ format that includes different reads from an experiment on *Vanilla planifolia* were selected for demonstration. Three samples were selected from experiment on three transcriptomic sequencing analyses on vegetative bud, reproductive bud, and mix bud tissues of V. planifolia which aims to identify floral transition transcripts (https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=237672) also downloaded Vanilla planifolia genome from NCBI.

Here's the list of following tools used for miRNA and lncRNA identification: FastQC

- Bowtie (bowtie2-2.4.5-linux-x86_64)
- FASTX-Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_ toolkit/),
- Samtools (samtools-1.16.1)
- SRA Toolkit (sratoolkit.3.0.0-ubuntu64)
- Cufflinks (cufflinks-2.2.1.Linux_x86_64)
- ViennaRNA Package v2.4.15 (http://www.tbi.univie.ac.at/~ivo/RNA/)CPC2
- miRDeep-P packageRepbase database
- TransDecoder-TransDecoder-v5.7.1
- Trinity

Methodology

Firstly, the raw reads were obtained from NCBI (Bioproject) containing SRA data [Accession: PRJNA237672 and ID: 237672]. The data had three samples (SRR1171643, SRR1171644, SRR1171645) which were selected for the identification of miRNAs and lncRNAs, encompassing the vegetative bud, reproductive bud, and mix bud tissues of

V. planifolia with the goal of identifying floral transition transcripts. The size of the data retrieved was approximately between 3GB to 4GB.

Data preprocessing

Initially, at the beginning of the process, data was subjected to **Fastq-dump** using SRA Toolkit (https://github.com/ncbi/sra-tools), which made the data ready for reading and saving to the FASTQ format. Therefore, otherwise, most bioinformatics tools will be created to operate in FASTQ format files.

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) allows the quality control of the sequence, raw data obtained from the high-throughput sequencing pipelines. It represents an integrative framework of analytics executed quickly, allowing the discovery of data flaws that should be corrected before further analysis. It is used to undergo deepening checks on their raw sequence data before the continuation of downstream analysis and perform functions like sequence duplication, adapter contamination detection, overrepresentation of sequences, and base quality estimation.

Indexing and mapping

Bowtie2 (https://github.com/BenLangmead/bowtie2) is one of the fast as well as small heap structures for long reference sequences and alignment sequencing reads. features of this tool are an option for different alignment types - the local, the paired-end, and the gapped alignments.

Downloaded *Vanilla planifolia* genome in FASTA format from NCBI (GenBank assembled sequence). Bowtie had been exercised for genome indexing via the process that made more effective retrieval of particular genomic regions possible and alignment quicker. To make sure that there exists a close alignment rate between the two samples and the reference genome, we mapped our samples to the reference genome (*Vanilla planifolia*).

Cufflinks the cufflinks (https://github.com/cole-trapnell-lab/cufflinks) approach usually serves the tasks of quantification and transcript assembly [40]. It takes the RNA-sequence reads as an input, makes transcripts, and then estimates their abundance. Moreover, it examines the regulation and differential expression. The application of Cufflinks is to sizing and assembly, of the expression of miRNAs, and lncRNAs from RNA-sequence data. This protocol is used as a prerequisite to determining the amount of these ncRNA species and recognizing their expression. This program aggregates counts and rules out differential expression for each transcript in a sequence. We resorted to the Software Tool Cufflinks as a tool for assessing the potential number of fragments whose origin lay both in transcripts as well as in genes.

Cuffmerge, the script within the cufflinks package (cufflinks-2.2.1.Linux_x86_64), is employed for merging transcript assemblies obtained from cufflinks into a single set of transcripts. The ability to designate the combination of transcript assemblies to be used in building samples or condition brackets would allow cuffmerge to identify common and condition-specific lncRNAs and miRNAs. This is a critical stage of constructing a broad public database that can be further used for downstream analysis.

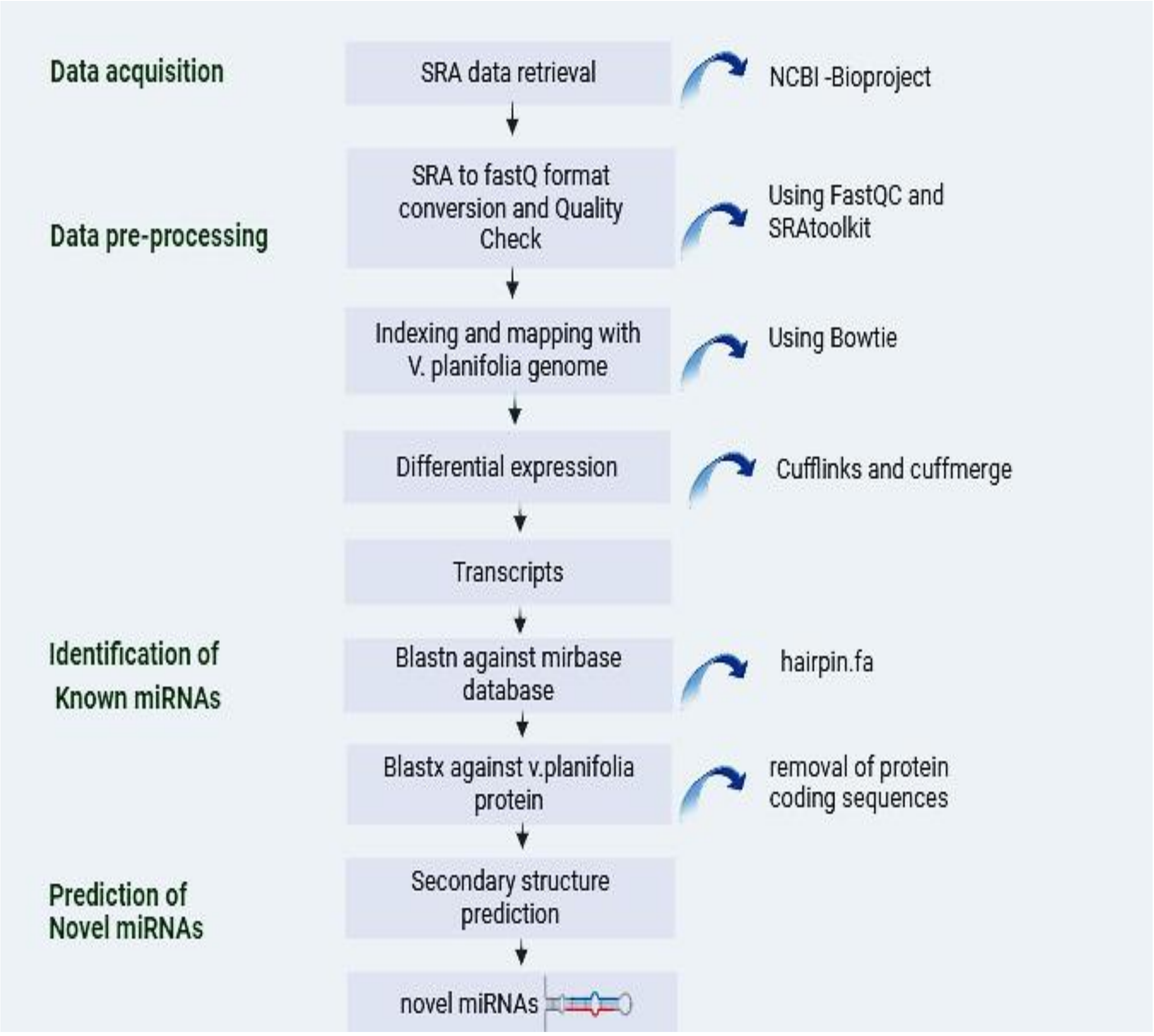

Data acquisition
SRA data retrieval
NCBI -Bioproject
Data pre-processing
SRA to fastQ format conversion and Quality Check
Using FastQC and SRAtoolkit
Indexing and mapping with V. planifolia genome
Using Bowtie
Differential expression
Cufflinks and cuffmerge
Transcripts
Identification of Known miRNAs
Blastn against mirbase database
hairpin.fa
Blastx against v.planifolia protein
removal of protein coding sequences
Prediction of Novel miRNAs
Secondary structure prediction
novel miRNAs

Fig 3. 1 Workflow for miRNA identification in *Vanilla planifolia.*

1) 3.2 We employed two distinct methods for the identification of miRNAs: First, miRNA prediction from transcript data

This first step of the process entailed initial data retrieval of all known miRNA and hairpin.fa sequences by using miRBase database which is an information hub with all miRNA sequences and their annotation up to date.

Transcript Selection: Through a stagewise analysis, only non-coding transcripts obtained from cufflinks were selected to apply the computational miRNA prediction. Entering onto the screen were these noncoding transcripts which most probably carried the miRNA sequence and which, therefore, kept on being selected for processing.

BLASTn Analysis: The hairpin.fa (Fasta format sequences of all miRNA hairpins) from the miRBase database along with the pre-miRNA sequence were compared against the pooled non-coding transcripts using BLASTn. BLASTn is a traditional tool that is extensively applied for comparing nucleotide sequences. Here we tune our parameters to work with an allowable with an e-value of 1000. It was expected that the in-depth bioinformatics analysis could point out the matching miRNA sequences with the known miRNAs or non-coding transcripts therefore, it would be obvious to discover the novel miRNA candidates.

Noncoding Confirmation: The sequences 'particularity' was examined to see if they were coding, this was done using Blastx analysis against the *V. planifolia* proteins database. Blastx is used for determining the identity of individual amino acids resulting from a nucleic query sequence translated in many reading frames also known as Orfs Among the sequences that have at least an 80% similarity to known protein-coding sequences those that were eliminated for further analysis. This step reduced the likelihood that the remaining reads would resemble coding sequences.

Coding Potential Assessment: The remaining sequences previously regarded as non-coding sequences had their code and then further underwent CPC analyses with the NR database. CPC is the auxiliary means that serves to define the ability to translate transcripts. From the sequences that were identified as the coding part, we excluded the ones that were not coding. This resulted in the consideration of potential non-coding candidates that have not been explored yet. Following this approach, we tabulated the sets of sequences which was obtained from the transcripts of the vegetative bud, reproductive, and mixed bud systematically and consequently, it was proven that those sequences were non-coding and had the functionality of micro followed by secondary structure prediction using RNAfold database. (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi)

Prediction of miRNA Targets

In contrast to animal miRNAs, plant miRNAs usually have perfect matching or they are neighborly with their targets. When the completeness of miRNA target prediction is concerned, the enormous spectrum of tools with varied specificity and sensitivity is used with the special feature. Among the tools that become very popular in s Target Prediction, include comTAR, TAPIR, psRNATarget, and psRobot. Similarly computational tools like miRNA specificity of the latest computational would often predict the miRNA targets using quite a sensitive and powerful technique called degradome sequencing. We used the **psRNATarget** web server for miRNA target prediction. which is user-friendly and has a graphical user interface (GUI) in place.

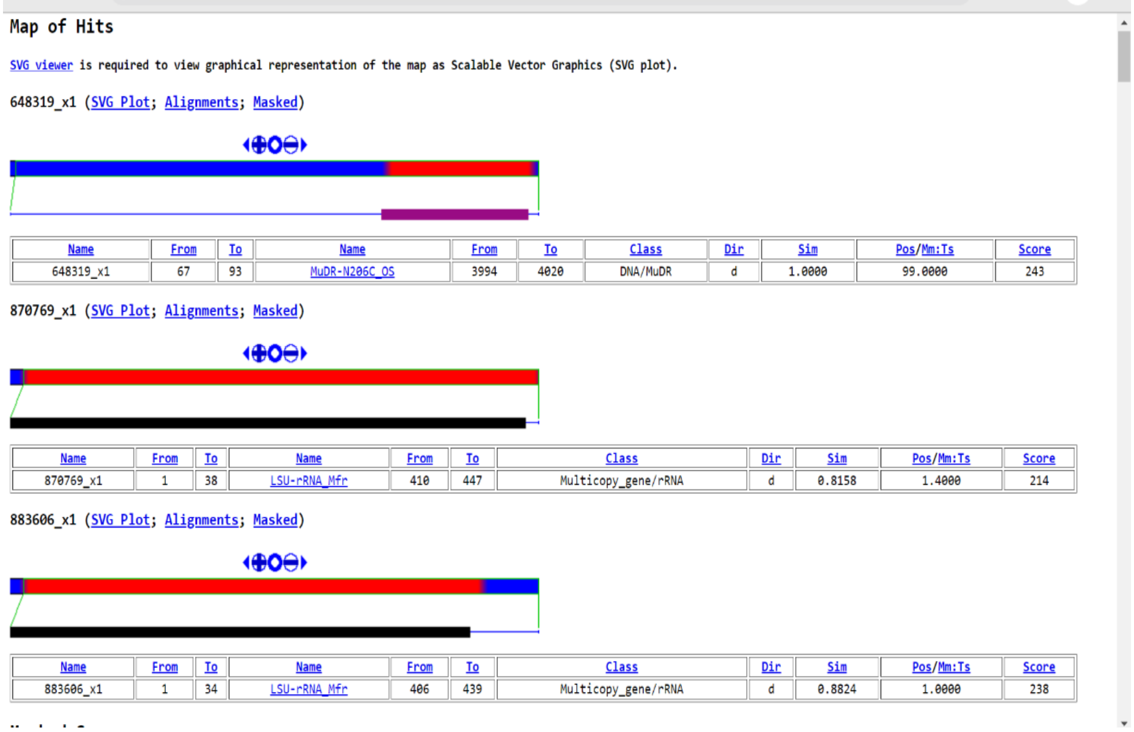
The key step of prediction uses a modified Smith-Waterman algorithm to locate the best microRNA–mRNA base matching position and to show whether the microRNA participates in the miRNA-mediated translational repression or mRNA cleavage. For prediction, we went to the online portal and chose a selection type based on the species. Next, entered the small RNA sequences followed by the species database. This happens when the given species is not available in the form of transcript sequences and the users can upload both the small RNA and the transcript sequences for the prediction.

3.3 <u>The second method for miRNA identification is as follows</u>

Data Processing:

- The filtering step (quality check) is applied for all three samples and then the cleaned reads from each sample are further merged into one fastq file by using the ‘cat’ command.
- The combined reads were converted to fasta format and then collapsed into unique tags. For this purpose, the fastx_collapser program available in FASTX-Toolkit is used to collapse the reads for downstream analysis. The resultant file (named unique tag) is then further formatted and processed to fit the format for miRNA prediction software employed in the next steps.

- The final process of data preparation includes read mapping to tRNA, rRNA, snRNA, snoRNA, and repeat sequences and the removal of these reads. For the removal of these RNAs (rRNA, tRNA, snRNA, and snoRNA) there exists the possibility of using Bowtie using a database of r-, t-, sn-, and sno-RNA sequences; this database can be obtained from the Rfam database. Similarly, to eliminate read mapping to repeat regions, the Repbase database (https://www.girinst.org/repbase/). It was firstly subjected to bowtie2 for indexing of the database followed by alignment using bowtie2-build.



Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

648319_x1 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
648319_x1	67	93	MuDR-N206C_OS	3994	4020	DNA/MuDR	d	1.0000	99.0000	243

870769_x1 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
870769_x1	1	38	LSU-rRNA_Mfr	410	447	Multicopy_gene/rRNA	d	0.8158	1.4000	214

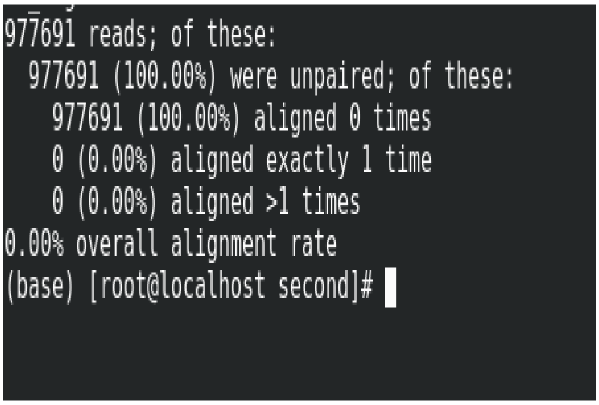
883606_x1 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
883606_x1	1	34	LSU-rRNA_Mfr	406	439	Multicopy_gene/rRNA	d	0.8824	1.0000	238

Fig 3. 2 Output repeats derived consistently from the Repbase database.

Identification of known miRNAs:

The filtered reads obtained from data processing were compared to the known plant miRNAs obtained from the online miRBase database for alignment [40], Bowtie is used to identify the known or conserved miRNAs by aligning the miRBase against the reads obtained. From this, we got a zero-alignment rate as shown in Fig 3.3 which specifies that the reads carried forward for novel miRNA prediction.



```
977691 reads; of these:
  977691 (100.00%) were unpaired; of these:
    977691 (100.00%) aligned 0 times
    0 (0.00%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
0.00% overall alignment rate
(base) [root@localhost second]# 
```

Fig 3.3 Alignment against miRBase showing zero-overall alignment rate.

Identification of novel miRNAs:

The non-annotated sequences are then used to predict the novel miRNA, based on alignment against the miRBase database. In this example, the prediction will be done with the help of the software miRDeep-P package. miRDeep-P is an user-friendly Perl script having nine Perl scripts that work step by step for predicting the miRNAs according to some plant species-specific filters available online and are free of cost [41]. The needs of different users such as biologists, molecular biologists, physical scientists, statisticians, and data analysts are taken into consideration when designing this package. The reads which do not match with the miRBase sequences are then aligned on the reference genome (*Vanilla planifolia*) using the Bowtie tool for novel miRNA identification.
Next, the alignments were in SAM format and were converted to fasta using miRDeep-P using the script "convert_SAM_to_blast.pl"

Subsequently, the alignments qualified by the following criteria were then allowed through the use of miRDeep-P script "filter_alignments.pl" It was then filtered to retain sequence identity 100%, alignment.

In the reads that are analysed, the reads that do not overlap with the annotated features such as exons, CDS, or any such features that have been assigned to the species under consideration were discarded for the purpose of further analysis. The corresponding annotations with which overlapping was performed were obtained from public databases like NCBI (https://www.ncbi.nlm.nih.gov/). This step is performed with the help of miRDeep-P "overlap. pl" script and "alignedselected." pl" scripts.

The following script copies the fasta sequences of the reads filtered in the previous tier.

Then, we excise the potential miRNAs from the reference genome by using the script "excise_candidate. pl". The script uses the reference genome, a fasta file format, and the filtered alignments as inputs. Similar to the authors of the miRDeep-P, the researchers suggest that it is best to take a 250-bp window for extracting the sequences for both the monocot and dicot plants.

Later the secondary structure of these sequences was predicted with the help of the software ViennaRNA Package v2.4.15 (http://www.tbi.univie.ac.at/~ivo/RNA/) with RNAfold utility and then removal of non-coding sequences using CPC2 web server.

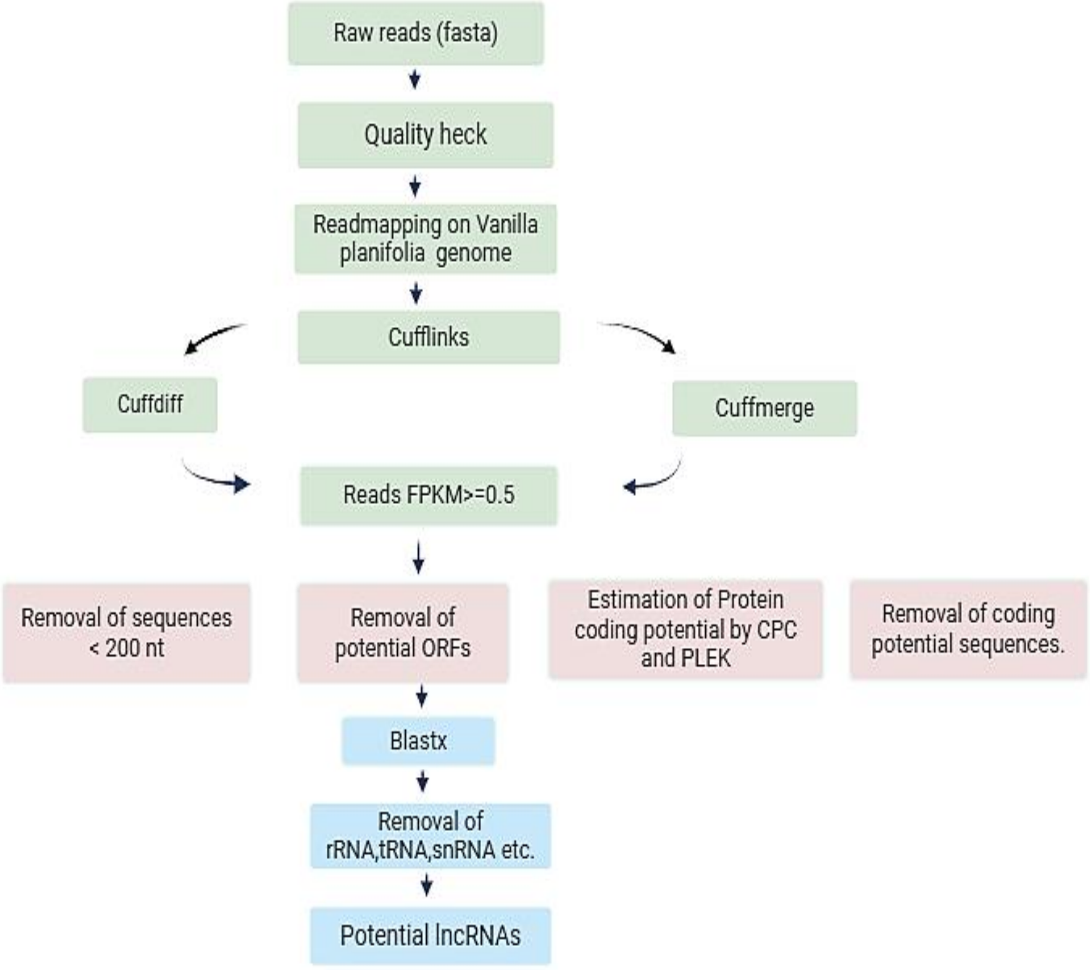
```
>CM028150.1_1 strand:+ excise_beg:95860 excise_end:96053

GUUUGAAGAGAUGGAAUUUUUGUCCGAUGGCUUUGUAACGUGAACAAU
AUAGAUUCGGUGAGAGCCCUCGUCAUGGCCGAUGGGAACGAAGGUCUUUC
UGUUGGUGGGAAAGGCAGUGAUAUuuaggaggacgagaaggaggaggaggagaacaguAAG
AAUGGAAGAAGCUUUCAUGUGAACCAGGUAAACG

((((............((((((((((...((((((...((.(((((......).)))))).))))))((((((......)))))).......((((((((.......)))))))).....
.........))))))))))...........)))).......(((((((...)))))))................ (-39.40)
```

The optimal secondary structure in dot-bracket notation with the minimum free energy of –39.10 kcal/mol, using RNAfold utility from ViennaRNA Package.

RNA filtration for prediction of long non-coding RNAs



Raw reads (fasta)
Quality heck
Readmapping on Vanilla planifolia genome
Cufflinks
Cuffdiff
Cuffmerge
Reads FPKM>=0.5
Removal of sequences < 200 nt
Removal of potential ORFs
Estimation of Protein coding potential by CPC and PLEK
Removal of coding potential sequences.
Blastx
Removal of rRNA,tRNA,snRNA etc.
Potential lncRNAs

Fig 3.4 Workflow for lncRNA identification in *Vanilla planifolia*.

Identification of LncRNAs 3.4

Greater than 200 nucleotides in length transcripts that were obtained after data pre-processing were selected and then filtered, and the transcripts <200 nucleotides of length were neglected. The 200nucleotide constraint proves to be the useful in terms of aiming for a quite large dataset but gives the data having more chances to bring out the essential features and roles of lncRNAs.

Based on the potential orf the **ORF finder**, this regard, minimum at or UCSC (https://eu.genome.jp/tools/orf/), a graphical and analytical application tool that locates all open reading frames in either sequence supplied by the user or a sequence already available in the database is worth exploring. The indication of the ORF of a sequence localizes that present sequence codes for some proteins [21]. This investigation occurred as some ORFs showed the ability to be broken down and were removed in the vice of ORF Finder and Transdecoder tool.

Sequence identification of transcripts causing determination

Biological transcripts were identified by matching with each other via BLAST search. The program computes the statistical significance by examining similarities of protein sequences and nucleotide sequences between the input sequences and known sequences in the protein databases. To disarm the transcripts with coding capability, BLASTx was performed on COG databases [8]. The COG database follows a systematic classification scheme of proteins that allows researchers to make functional inferences by comparing them to the known attributes of orthologous proteins in this way anyone can interpret the results generated by the genomic and proteomic experiments. On the other hand, uniport provides bioinformaticians, bioportal researchers, and others interested with access to a wide variety of sequences of proteins and their respective annotations. The UniProt data collection not only remains the most reliable but also is constantly updated, to facilitate various downstream applications like functional annotation, amino acid analysis, and protein-protein interactions.

Elimination of the sequences for the protein domain

In the initial stage, the acquired transcripts were then used in the filtering process, whereby the transcripts containing coding potential were cut off. The CPC2 tool follows through the processes of CNC1. CPC2 (http://cpc2. gao-lab. (http://cnc.ecuadlabs.org/) is an RNA-cleavage site online finder that uses sequence features to achieve rapid and accurate predictions.

Differential expression of novel lncRNA:

These novel lncRNAs were further characterized in terms of their expression differences which can give more details on their participation in certain biological processes, developmental stages, or specific reactions to the certain stimulations. Besides, it is useful for the perception of the regulatory functions of lncRNAs and their possible significance for higher plants, crop plants, and stress adaptation.

We used prepDE.py script to make a count matrix (gene matrix and transcript matrix from transcripts of three samples obtained from cufflinks) and edgeR differential expression P-Value Threshold: Based on the results obtained in this study, the P value represents the likelihood of obtaining such values if the null hypothesis is true, value less than > 0.05 is considered to be statistically significant and gives the overall result of this investigation of possible relationships between the genetic factors and the growth rate. For instance, 05 is used as a criterion identifying the level of statistical significance. When looking at the problem in the context of differential expression analysis, a P value <0. 05. This indicates that the observed differential expression is not by chance but likely regulated by gene products.

LogFC Threshold: is a relative measurement of the impact of a condition on gene expression, which is the logarithm of the fold change. The result with a gene having a directionally significant up-regulation or down-regulation in the examined tissues was identified when the value of a log FC absolute value was > +2 or < -2. A logFC value larger than 2 means higher expression and a small value of logFC less than -2 means lower expression has been achieved.

CHAPTER - 4

RESULTS

4.1 Results

Identification of miRNAs and lncRNAs in *V. planifolia* using publically available transcriptomic information

To identify the long non-coding RNAs present in the genome of *A. thaliana*, we used the transcriptomic data that is available on the NCBI database [Accession: PRJNA237672 and ID: 237672]. Our dataset yielded 3 replicates as shown in Table 4.1. Three transcriptomic sequencing analyses were performed on vegetative bud, reproductive bud, and mixed bud tissues of V. planifolia that aimed to identify floral transition transcripts.

Table 4.1 Summary of raw data collection from NCBI.

Conditions	ID's	NO. of Reads	No. of bases	Size
Vegetative bud	SRR1171643	27,636,224	5G	3.2Gb
Reproductive bud	SRR1171644	25,718,096	4.6G	3Gb
Mixed bud	SRR1171645	27,609,939	5G	3.3Gb

After the FASTQC-analysis, we obtained a clean dataset, which was used for transcriptome assembly as well as for the identification of miRNAs and long non-coding RNA. The raw data was already trimmed so we skipped the trimmomatic step that was applied for removing reads with low quality, overlapping sequences, adapter contamination, and low quality based on a set of thresholds.

Table 4.2 FASTQC output of raw data.

Sample Name	Base sequence Quality	Sequence Duplication level	Overrepresented Sequences	GC% content
SRR1171643	PASS	FAIL	PASS	49%
SRR1171644	PASS	FAIL	PASS	48%
SRR1171645	PASS	FAIL	PASS	48%

After the FASTQC-analysis, we obtained a clean dataset, which was used for transcriptome assembly as well as for the identification of miRNAs and long non-coding RNA. The raw data was already trimmed so we skipped the trimmomatic step that was applied for removing reads with low quality, overlapping sequences, adapter contamination, and low quality based on a set of thresholds.

Each sample with a sequence duplication level has a value of "FAIL" within this column, which suggests that there is duplication present.The column Overrepresented Sequences displays if one or some certain sequences occur too frequently here each set of columns has the value "PASS" values under the column, meaning that the samples have no over-represented sequences as shown in Table 4.2.In general, it seems that each row consists of a certain data sample or dataset and each column presents certain information on the quality, features, and content of the sequences in those samples.

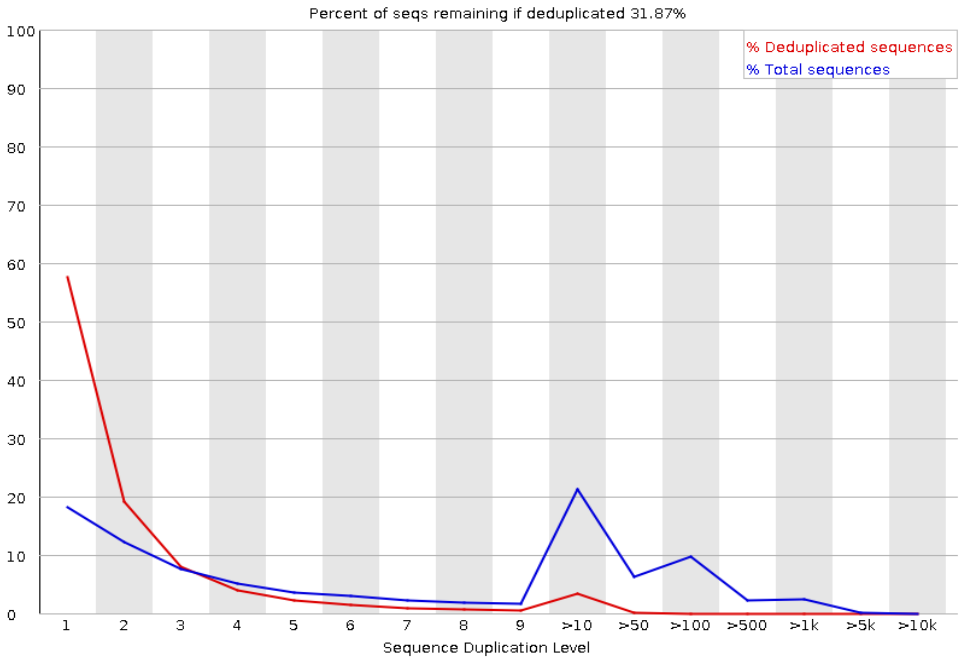


Fig 4.1 The assortment of duplications among sequences using FASTQC.

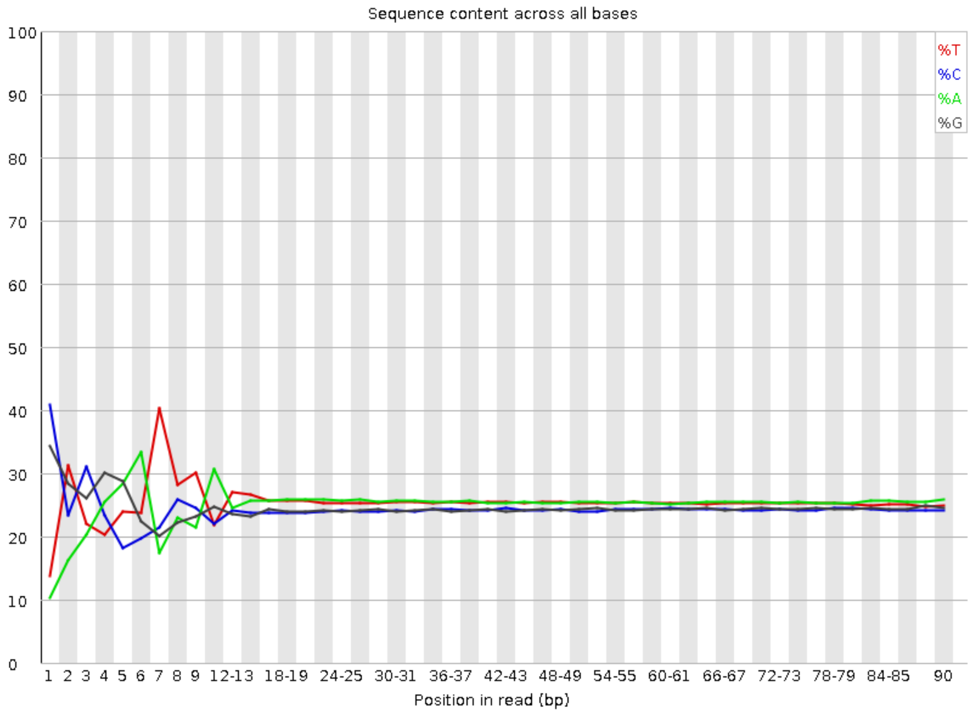


Fig 4.2 Quality check for the sequence through the FASTQC program with all the bases.

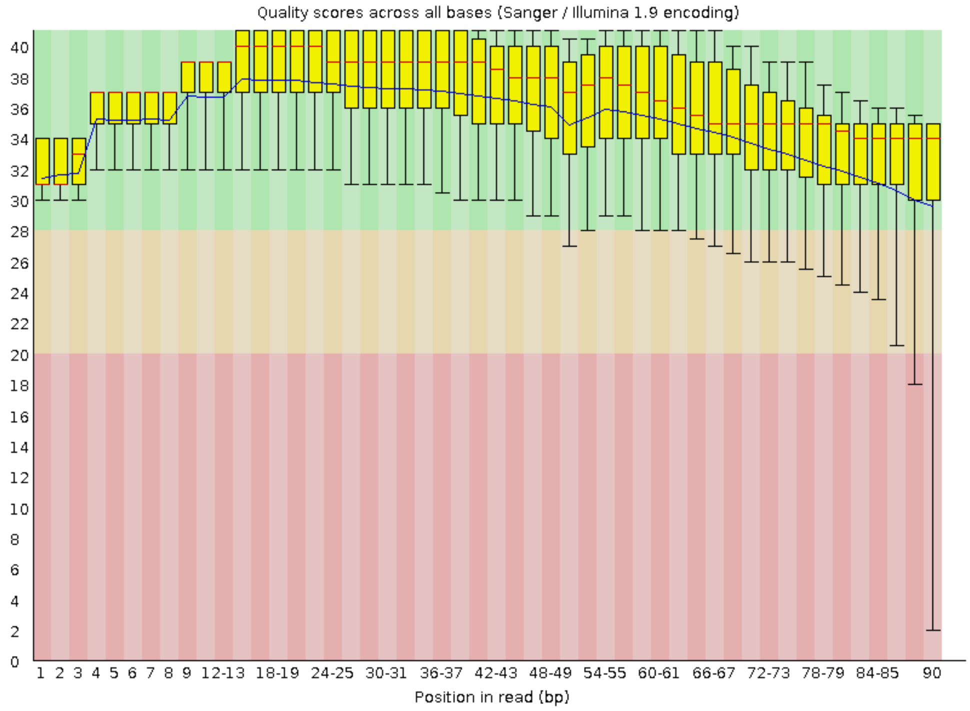
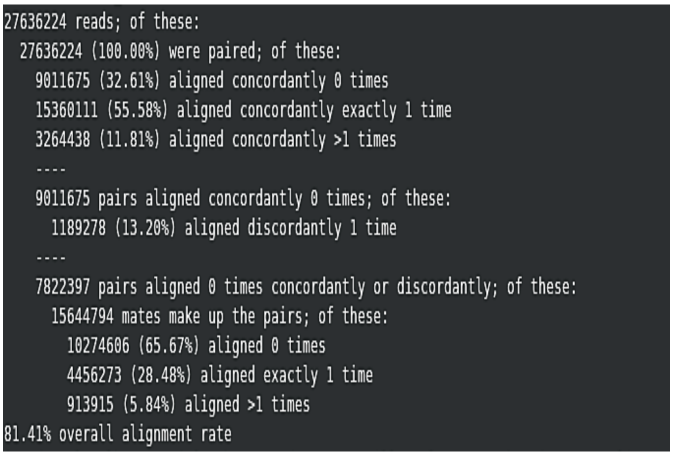


Fig 4.3 Using FASTQC quality check representing quality scores across sequences.

Table 4.3 Descriptive statistics to clean data.

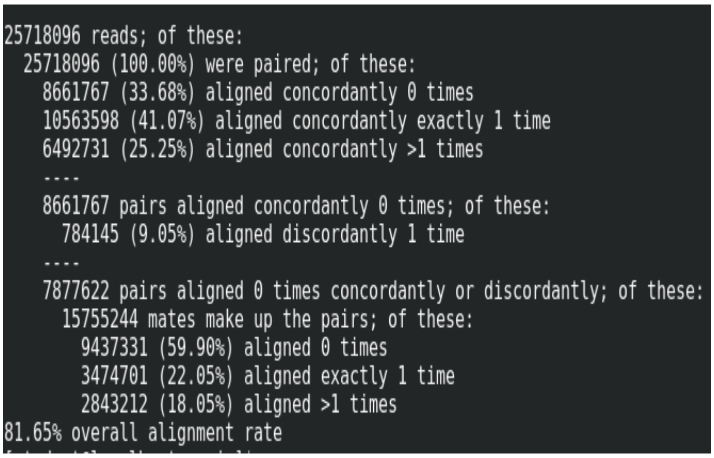
Samples	Total Reads
SRR1171643	27636224
SRR1171644	25718096
SRR1171645	27609939

The obtained clean data was then mapped on the reference genome i.e. *Vanilla planifolia* using the Bowtie2 (bowtie2-2.4.5-linux-x86_64) tool in order to check the overall alignment of our samples with the reference genome. The reads from the vegetative bud sample (81.44%) as shown in Fig 4.4. were successfully mapped with the reference genome. Similarly, the reads from the reproductive bud (80.11%) as shown in Fig 4.4, and the last mixed bud with (81.65%) as shown in Fig 4.5 were successfully mapped with the reference genome.



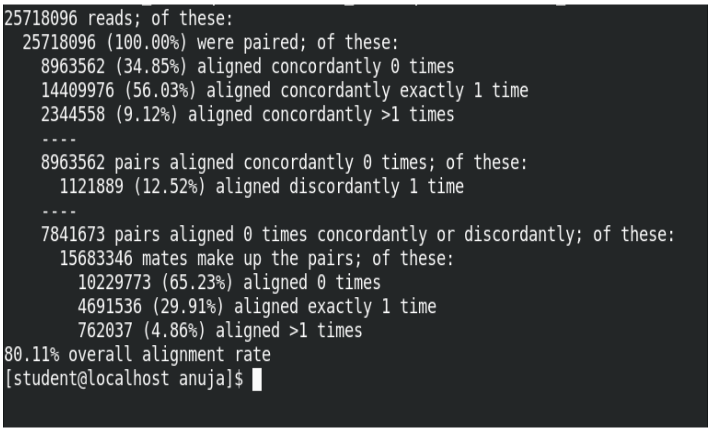
```
27636224 reads; of these:
  27636224 (100.00%) were paired; of these:
    9011675 (32.61%) aligned concordantly 0 times
    15360111 (55.58%) aligned concordantly exactly 1 time
    3264438 (11.81%) aligned concordantly >1 times
    ----
    9011675 pairs aligned concordantly 0 times; of these:
      1189278 (13.20%) aligned discordantly 1 time
    ----
    7822397 pairs aligned 0 times concordantly or discordantly; of these:
      15644794 mates make up the pairs; of these:
        10274606 (65.67%) aligned 0 times
        4456273 (28.48%) aligned exactly 1 time
        913915 (5.84%) aligned >1 times
81.41% overall alignment rate
```

Fig 4.4 Mapping results of a sample SRR1171643 with the reference genome



```
25718096 reads; of these:
  25718096 (100.00%) were paired; of these:
    8661767 (33.68%) aligned concordantly 0 times
    10563598 (41.07%) aligned concordantly exactly 1 time
    6492731 (25.25%) aligned concordantly >1 times
    ----
    8661767 pairs aligned concordantly 0 times; of these:
      784145 (9.05%) aligned discordantly 1 time
    ----
    7877622 pairs aligned 0 times concordantly or discordantly; of these:
      15755244 mates make up the pairs; of these:
        9437331 (59.90%) aligned 0 times
        3474701 (22.05%) aligned exactly 1 time
        2843212 (18.05%) aligned >1 times
81.65% overall alignment rate
```

Fig 4.5 Mapping results of a sample SRR1171644 with the reference genome



```
25718096 reads; of these:
  25718096 (100.00%) were paired; of these:
    8963562 (34.85%) aligned concordantly 0 times
    14409976 (56.03%) aligned concordantly exactly 1 time
    2344558 (9.12%) aligned concordantly >1 times
    ----
    8963562 pairs aligned concordantly 0 times; of these:
      1121889 (12.52%) aligned discordantly 1 time
    ----
    7841673 pairs aligned 0 times concordantly or discordantly; of these:
      15683346 mates make up the pairs; of these:
        10229773 (65.23%) aligned 0 times
        4691536 (29.91%) aligned exactly 1 time
        762037 (4.86%) aligned >1 times
80.11% overall alignment rate
[student@localhost anuja]$ █
```

Fig 4.6 Mapping results of a sample SRR1171645 with the reference genome**.**

miRNA identification

Data Collection and Preparation - miRNA and pre-miRNA Retrieval-All the sequences of the published miRNA and pre-miRNA are present in the miRBase database (a database that holds total miRNA sequences) specifically hairpin.fa was downloaded for the analyses.
Utilization of transcriptome data which are non-coding transcript. The non-coding transcript from transcriptome datasets was chosen to predict microRNA, demonstrating the dedication to achieving comprehensive results.

miRNA Prediction

BLASTn Analysis: The retrieved sequences from the miRBase database (https://www.mirbase.org/download/hairpin.fa) were screened against the pooled transcripts by performing BLASTn. Where query sequence was transcript data and the database mirbase data.

We obtained **505** novel miRNAs.

BLASTx analysis: Specific Process described: The Blastx program was used against a particular protein database (*Vanilla planifolia* protein from protein database) with a sequence identity limit of c. ≥80%. The ranges showing higher homology (of about ≥80% identity) to the coding sequences were the only ones found to be included in the subsequent analysis.

We obtained **273920** from Blastx which was then filtered with a cutoff of more than 80% and removal of duplicates and got 10613 which was further carried out removed duplicates and finally got 40.

Further Verification: At the end, the sequences were elucidated using two different methods, such as CPC (Coding Potential Calculator). Any segments used for the codon search were not classified for analysis. This fractionation determines whether the identified miRNA sequences are noncoding and those that are protein-coding sequences are filtered out from the list of possible miRNAs.

Out of 40, we removed the coding ones and got **27** (non-coding) novel miRNAs.

The secondary structure predictions were carried out with the help of the RNAfold web server.

CMGCUUUGUCACACUGGUGUAUUUUCCAUCGAAAGAAUGCUGAACAAGUU
GUUCAAACAUGGGAUAAGCAGUUUCACAGUUCCAAAAAGGAGCAAAAGAU
UCCUUUCCUAUAUCUUGCCAAUGACAUUCUACAGAAUAGUAGGCGUAAUG
GUAUGGAAUUUGUUGCUGAGUUUUGGAAGGUGCUUCCAUCUGCAGUUAAA
GAUGUCUCUGAAAAUGGUGAAGAACACGGAAAAAAUGUGGUGUCAAGACU
G
....((((.(((((..((((.(((.((((....(((.((((...((.(((.....))).)).....))))....(((..((((((((...(((((.(((((((...((.......((((
.((...((((....)))).)).))))....))...))))))).)))))...)))))))).)))....(((((........)))))..)))....)))).)))..)))).........)).
))).))))....

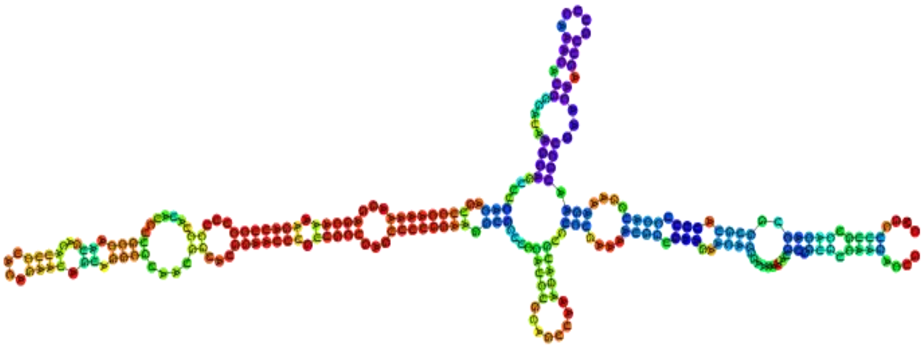
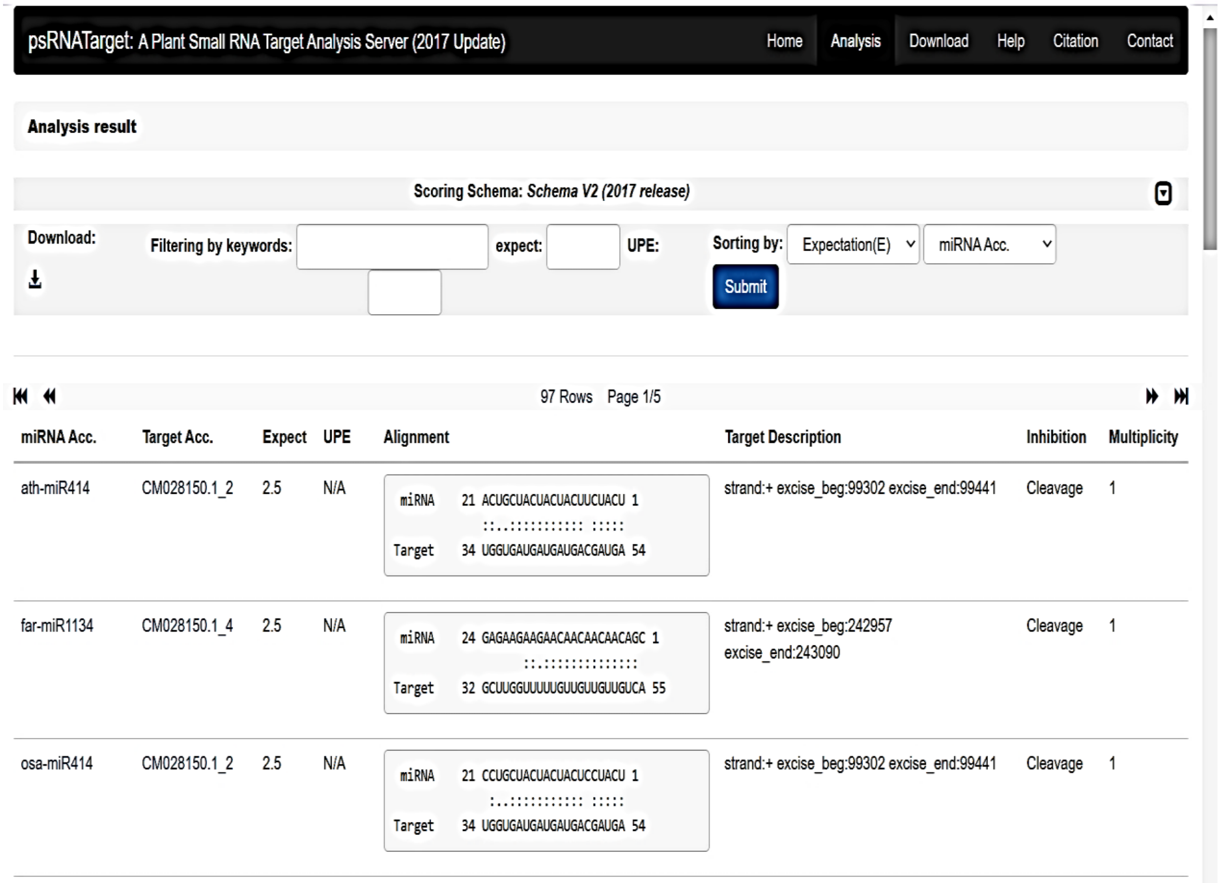


Fig 4.7 Secondary structure prediction of a microRNA generated by RNA secondary structure prediction tool RNAfold.
(http://rna.tbi.univie.ac.at/RNAfold/vTzyLGjmwX/sequence1_pp.pdf)



psRNATarget: A Plant Small RNA Target Analysis Server (2017 Update)
Home
Analysis
Download
Help
Citation
Contact

Analysis result

Scoring Schema: *Schema V2 (2017 release)*

Download: Filtering by keywords: expect: UPE: Sorting by: Expectation(E) miRNA Acc. Submit

97 Rows Page 1/5

miRNA Acc.	Target Acc.	Expect	UPE	Alignment	Target Description	Inhibition	Multiplicity
ath-miR414	CM028150.1_2	2.5	N/A	miRNA 21 ACUGCUACUACUACUUCUACU 1 ::..::::::::::: ::::: Target 34 UGGUGAUGAUGAUGACGAUGA 54	strand:+ excise_beg:99302 excise_end:99441	Cleavage	1
far-miR1134	CM028150.1_4	2.5	N/A	miRNA 24 GAGAAGAAGAACAACAACAACAGC 1 ::.::::::::::::::: Target 32 GCUUGGUUUUGUUGUUGUUGUCA 55	strand:+ excise_beg:242957 excise_end:243090	Cleavage	1
osa-miR414	CM028150.1_2	2.5	N/A	miRNA 21 CCUGCUACUACUACUCCUACU 1 :..::::::::::: ::::: Target 34 UGGUGAUGAUGAUGACGAUGA 54	strand:+ excise_beg:99302 excise_end:99441	Cleavage	1

Fig 4.8 psRNATarget result for miRNA target prediction.

- Prediction for miRNA-target interactions as shown in Fig 4.8 with miRNAs: ath-miR414, far-miR1134, and osa-miR414.
- The miRNAs are expected to be recognized and cleave target mRNA sequences. The value of the expectation score is 2. 5 reflects an average level of trust in these forecasts.
- The binding sites and the cleavage for the inhibition type are also outlined to suggest that these miRNAs are potent mRNA degradation when the miRNAs bind to their respective mRNA targets.

lncRNA identification

The Cufflinks platform was used to calculate FPKM values for each transcript we obtained **213833** which was then filtered with FPKM <0.5 and got **213833** transcripts with the removal of duplicates ultimately getting **500.**

This estimation tool (Cufflinks) is used to determine the amount of broken fragments extracted per transcript and gene samples. With the cuff-merge tool, the unique and non-overlapping set of transcripts (three transcripts) was identified. Also, we found some different designated genes in the sample we got. Perform subsequent Differential Analysis, the number of differentially expressed genes, several upregulated genes, and downregulated could be done using our sample.

The processed pool of reads achieved through a filtration pipeline sets the criteria that separate lncRNAs from other molecules based on the essential parameters. Those transcripts with a length of more than **200** nucleotides were selected for additional processing and those which were of a shorter length were automatically discarded from the analysis using Linux command. Retrieved sequences were then were a total of 500 transcripts and it was taken forward for ORF finding. The ones with an ORF (open reading frame) were excluded, leading to a total of **51** transcripts that stayed.

Then Blastx was performed with 3 protein databases like NCBI non-redundant protein database, Swiss-Prot database, and COGs databases with the evalue - 0.001.

Table 4. 4 Statistical analysis of Novel lncRNAs identification from various protein databases after duplicates and coding potential sequences removal.

Protein databases	Reads obtained after Cutoff <50%	After the removal of duplicate	Total novel lncRNAs obtained
COGs database	314	6	22 remains After removing the coding potential sequences
Swiss-Prot database	1319	24	
Refseq Database	3797	26	

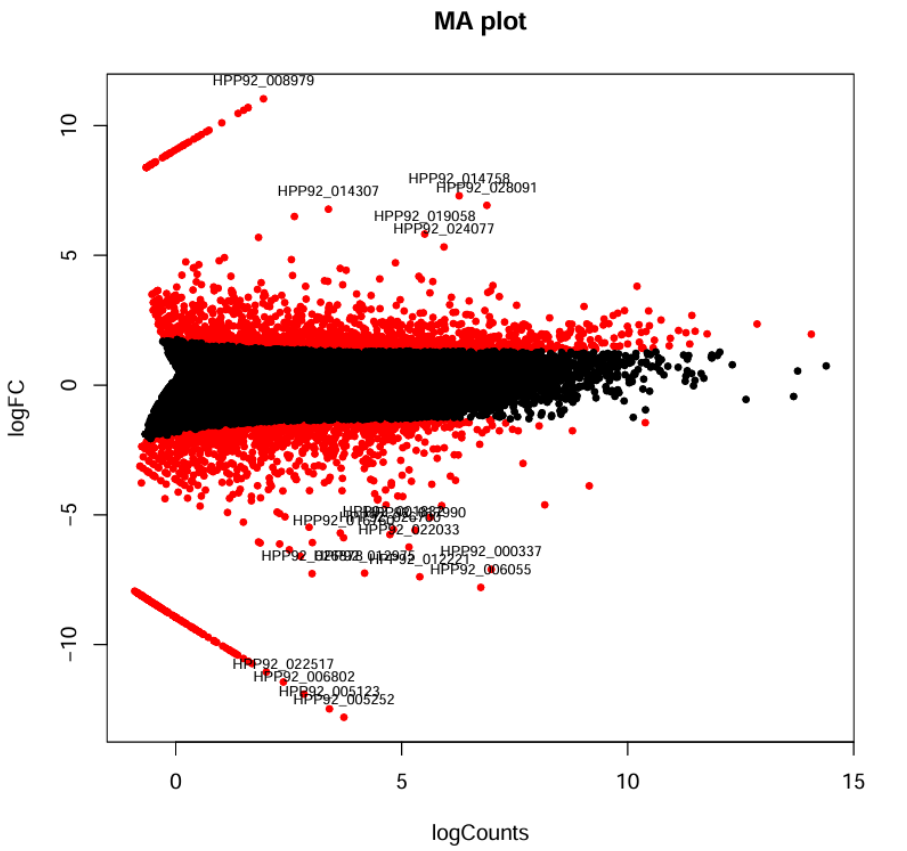


Fig 4.9 MA plot representing Differentially Expressed Genes.

MA plots are a form of a scatter plot and are most commonly used in search of gene expression patterns. They may be used: It is useful for visualizing differences between measurements taken in two conditions, the MA plot shown in Fig 4.9 can be explained as follows:

X-axis (log counts): reflects the mean value of the gene (or some other characteristics being investigated) in the two compared conditions, and it is frequently in log form. In this case, it is relatively probable to represent the mean of the concentrations of lncRNAs in the form of the log-transformed counts.

Y-axis (logFC): This axis shows the multiple of the fold change, referred to as the log fold change (logFC), of each gene between the two conditions. The value of logFC greater than 0 demonstrates that a specific gene is more expressed in one condition compared to the other condition, while the value of logFC less than 0 makes it conform that the specific gene is more expressed in other conditions as compared to other conditions.

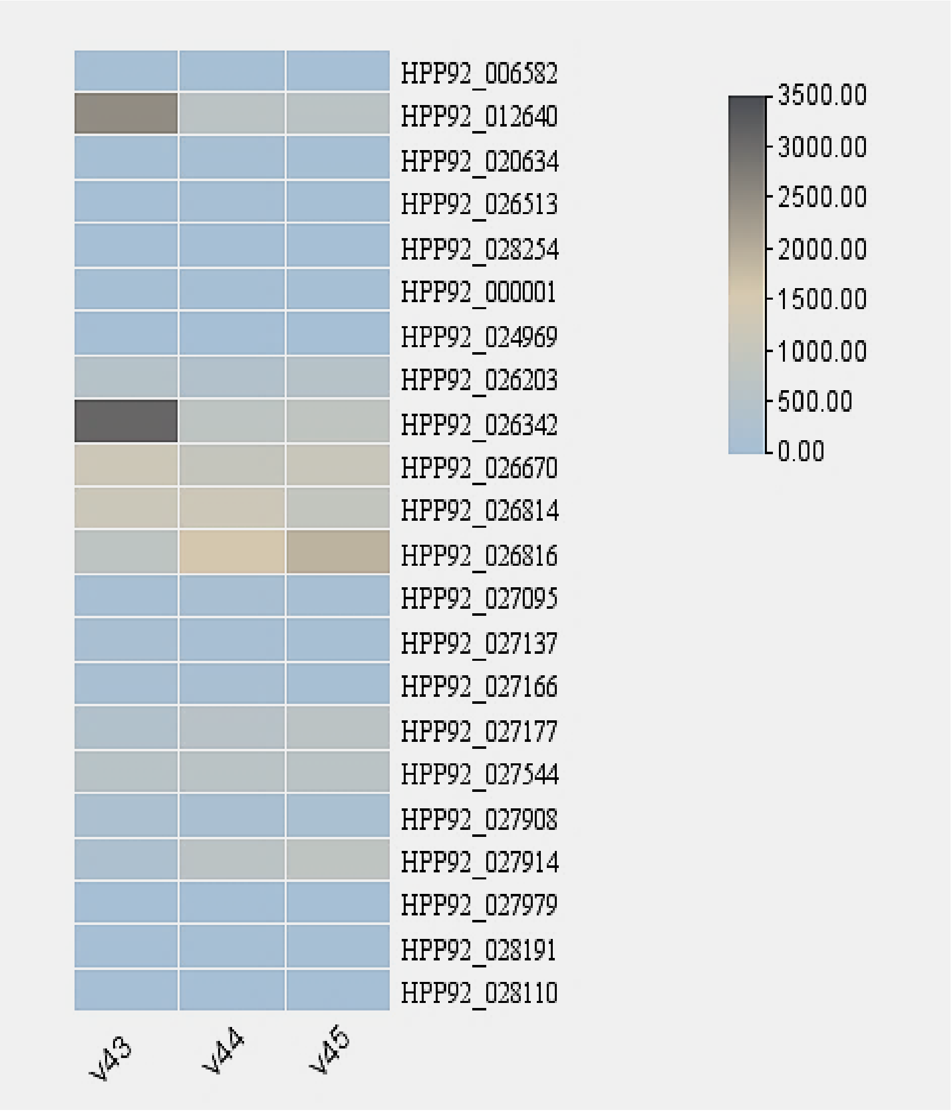


Fig 4.10 Heatmap of expression Levels of Genes (HPP92) Across Different samples.

There is a color scale bar in Fig 4.10 that goes from light blue to dark gray representing: Light Blue (0. 00 can be interpreted as the lesser range of values, and it would not be unreasonable to expect it to be right at zero or even below it. Darker Colors (up to 3500. 00) signify higher values and the black color indicates even higher values on this scale that range from 0 to 3500. 00.

Analysis

Patterns and Trends: The best approach is to look at the dissimilar areas and or regions where color patterns are similar. For instance, if there is a row where most of the cells for different columns are shaded, this means that one is likely to have high values for these variables for the particular identifier, for instance, HPP92_006582.To compare the hierarchy of color intensity, we can compare rows and columns of two, in which identifiers or variables could easily be identified as having high or low values.

All in all, this heatmap gives a tabular look of the data matrix and the degree of the color corresponding to each cell enables one to understand the high and low values or differences and get a preview of what to look into between different identifiers and the variable types.

CHAPTER - 5

CONCLUSION

5.1 CONCLUSION

MiRNAs and lncRNAs belonging to the *Vanilla planifolia* genome-wide miRNAs are pivotal to comprehend the post-transcriptional gene regulation (PTGR) and its significance to improvements in crop production and agriculture attributes. The non-coded RNAs particularly miRNAs and long non-coded RNAs are majorly involved in the regulation of the genes, especially in the plants of https://eataway.org/signature-dishes/vanilla-extract/ The discovered miRNAs and lncRNAs in *Vanilla planifolia* could be efficiently used for crop development projects and so the harvest will be high and nutritional value immense together with stress resistance [3]. Comprehending the microRNA-target relationship in *Vanilla planifolia* helps us illuminate the molecular reactions in stress areas in plants. This is conjointly useful in understanding the mechanisms of crop improvement.

In conclusion, this study aimed to identify miRNAs and lncRNAs through a procedure involving several steps like translational selection, BLASTn search, and sequence read against miRBase. Other than that, applying bioinformatics tools including; FastQC software, Bowtie2 for genome indexing, and Cufflinks for the transcript assembly helps better analyze raw sequence data.

Moreover, the use of concepts such as miRDeep-P and RNAfold has made it possible to predict secondary structures and targets interacting with miRNA, information. Furthermore, the recognition of lncRNAs with the help of Cufflinks and methodologies invoking FPKM also helped in comprehending the understanding of non-coding RNA specialization in *Vanilla planifolia*.

Overall, the study has given a robust systematic approach identification and analysis of miRNAs and lncRNAs of *vanilla planifolia* with the help of Next Generation Sequencing technology, hence forming a sturdy basis for additional research in plant genomes and molecular biology. The results outlined in this work are useful in enriching the current literature about non-coding RNA types and their roles in plant processes.

The genome-wide miRNAs and lncRNAs in *Vanilla planifolia* are a major step forward in the interpretation of post-transcriptional gene regulation and its potential in germplasm bases for crop improvement and agriculture traits. The study provides a precious insight into the regulatory functions of ncRNAs and the fact that they may be used in crop improvement programs, thus, adding to the successful development of new crop varieties with improved agronomic properties and stress tolerance.

In this study venture for the first time in plant species genome-wide the identification of miRNAs and lncRNAs in *Vanilla planifolia* has been accomplished. Unique: A total of 27 miRNAs and 22 lncRNAs have been identified through original and meticulous research. This unprecedented work not only contributes to the identification and characterization of the non-coding RNA regulation mechanism in *Vanilla planifolia* but also facilitates the identification of new regulatory miRNAs and lncRNAs in this important orchid species for both growth and development in addition to specialized metabolism. The list of miRNAs and lncRNAs provided in this study serves as a robust reference tool for future genetics and breeding studies that focus on improving the agronomic characteristics *of Vanilla planifolia.*

REFERENCES

1. J.G. Fouche and L. Jouve, "*Vanilla planifolia*: history, botany and culture in Reunion Island," Agronomie, vol. 19, pp. 689-703, 1999.
2. S. Bory, M. Grisoni, M.F. Duval, and P. Besse, "Biodiversity and preservation of vanilla: present state of knowledge," Genetic Research on Crop Evolution, vol. 55, pp. 551-571, 2008.
3. D. Correll, "Vanilla: its botany, history, cultivation and economic importance," Economic Botany, vol. 7, pp. 291-358, 1953.
4. A. Smith et al., "Comprehensive mapping of miRNA profiles in Arabidopsis thaliana," J. Plant Res., vol. 123, no. 1, pp. 69–78, Jan. 2010.
5. B. Jones and L. Chen, "Identification and analysis of miRNAs in plants," Genes & Dev., vol. 20, no. 13, pp. 1740–1749, Jul. 2006.
6. Wang, T. Z., et al., "The interplay between noncoding RNAs and insulin in diabetes," Cancer Letters, vol. 419, pp. 53-63, 2018.
7. Sunkar, R., et al., "Cloning and characterization of microRNAs from rice," Plant Cell, vol. 17, no. 5, pp. 1397-1411, 2005.
8. D. Patel et al., "Identification of miRNAs and their targets involved in drought stress response in pigeonpea," Plant Mol. Biol. Report., vol. 32, no. 4, pp. 1118–1132, Aug. 2014
9. Franco-Zorrilla, J. M., et al., "Target mimicry provides a new mechanism for regulation of microRNA activity," Nature Genetics, vol. 39, no. 8, pp. 1033-1037, 2007.
10. Cai, X., and Cullen, B. R., "The H19 lncRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r," Nature Cell Biology, vol. 13, no. 7, pp. 803-811, 2011.
11. LMI-DForest, "LMI-DForest: A deep forest model towards predicting lncRNA-miRNA interactions," Computational Biology and Chemistry, vol. 75, pp. 287-294, 2018.
12. F. Zhang et al., "Roles of long noncoding RNAs in plant development and stress responses," Trends in Plant Science, vol. 20, no. 10, pp. 637–646, Oct. 2015.
13. B.J. Reinhart et al., "MicroRNAs in plants," Genes & Development, vol. 16, no. 13, pp. 1616–1626, July 2002
14. X. Chen, "A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development," Science, vol. 303, no. 5666, pp. 2022–2025, Apr. 2004
15. Y. Sunkar and J.-K. Zhu, "Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis," Plant Cell, vol. 16, no. 8, pp. 2001–2019, Aug. 2004

16. V. Ambros, "The functions of animal microRNAs," Nature, vol. 431, no. 7006, pp. 350–355, Sep. 2004.
17. X. Chen, "A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development," Science, vol. 303, no. 5666, pp. 2022–2025, Apr. 2004.
18. Y. Chekanova, "Long non-coding RNAs and their functions in plants," in Current Opinion in Plant Biology, vol. 27, pp. 207–216, 2015. Available: DOI: 10.1016/j.pbi.2015.08.003
19. R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," Cell, vol. 75, no. 5, pp. 843–854, Dec. 1993.
20. B. J. Reinhart et al., "MicroRNAs in plants," Genes & Development, vol. 16, no. 13, pp. 1616–1626, Jul. 200
21. L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," in Nature Reviews Genetics, vol. 5, no. 7, pp. 522-531, 2004.
22. E. M. Thomson et al., "Extensive post-transcriptional regulation of microRNAs and its implications for cancer," in Genes & Development, vol. 22, no. 16, pp. 2202-2207, 2008.
23. A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," in Nucleic Acids Research, vol. 42, D68-D73, 2014.
24. V. Ambros, "The functions of animal microRNAs," in Nature, vol. 431, no. 7006, pp. 350-355, 2004.
25. M. Schmittgen and K. J. Livak, "Analyzing real-time PCR data by the comparative C(T) method," in Nature Protocols, vol. 3, no. 6, pp. 1101-1108, 2008. [Online]. Available: DOI:10.1038/nprot.2008.73
26. F. Liu, C. Marquardt, T. Lister, S. Swiezewski, and C. Dean, "Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing," in Science, vol. 327, no. 5961, pp. 94–97, 2010.
27. M. Heo and V. N. Kim, "Regulating the regulators: posttranslational modifications of RNA silencing factors," in Cell, vol. 139, no. 1, pp. 28–31, 2009.
28. M. E. Dinger et al., "Insights into the role of long non-coding RNAs in disease mechanisms," in Nature Reviews Genetics, vol. 19, no. 1, pp. 20-39, 2018.
29. J. Kapranov et al., "RNA maps reveal new RNA classes and a possible function for pervasive transcription," in Science, vol. 316, no. 5830, pp. 1484-1488, 2007. [Online]. Available: DOI:10.1126/science.1138341
30. P. Carninci et al., "The transcriptional landscape of the mammalian genome," in Science, vol. 309, no. 5740, pp. 1559-1563, 2005.

31. S. A. Teichmann and J. C. Marioni, "Single-cell genomics: From understanding biology to deriving clinical applications," in Cell Stem Cell, vol. 23, no. 6, pp. 786-799, 2018

32. B. J. Reinhart, E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel, "MicroRNAs in plants," Genes Dev., vol. 16, no. 13, pp. 1616–1626, Jul. 2002, doi: 10.1101/gad.1004402.

33. M. W. Jones-Rhoades and D. P. Bartel, "Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA," Mol. Cell, vol. 14, no. 6, pp. 787–799, Jun. 2004, doi: 10.1016/j.molcel.2004.05.027.

34. D. W. Ng et al., "Stress-Induced MicroRNA Transcriptomes in Arabidopsis thaliana," Sci. Rep., vol. 9, no. 1, p. 7550, Dec. 2019, doi: 10.1038/s41598-019-43907-2.

35. R. Karlova et al., "Identification of microRNA targets in plants using parallel analysis of RNA ends," Plant Physiol., vol. 171, no. 4, pp. 2211–2221, Apr. 2016, doi: 10.1104/pp.16.00042.

36. K. G., et al. "Identification and Characterization of Long Noncoding RNAs and miRNAs Involved in Seed and Pod Development in C. cajan." Scientific Reports, vol. 9, no. 1, pp. 18191, 2019. DOI: 10.1038/s41598-019-54340-6.

37. J. Ding et al., "A genetic atlas of maize biology," Plant J., vol. 92, no. 3, pp. 477–492, Nov. 2017, doi: 10.1111/tpj.13676.

38. Y. Wang et al., "Genome-wide identification and expression analysis of the lncRNA-associated regulatory network in response to drought stress in rice," Rice, vol. 13, no. 1, p. 47, Dec. 2020, doi: 10.1186/s12284-020-00407-6.

39. Jia, H. et al. Genome-wide computational identification and manual annotation of human long noncoding RNA genes genome-wide computational identification and manual annotation of human long noncoding RNA genes. Bioinformatics 1478–1487 (2010).

40. A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," Nucleic Acids Res., vol. 47, pp. D155–D162, 2019.

41. X. Yang and L. Li, "miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants," Bioinformatics, vol. 27, pp. 2614–2615, 2011.

42. Trapnell, C. et al. Differential gene and transcript expression analysis of RNAseq experiments with TopHat and Cufflinks. Nature Protocols 7, 562–578 (2013).

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 30/5/24

Type of Document (Tick): □ PhD Thesis ☑ M.Tech/M.Sc. Dissertation □ B.Tech./B.Sc./BBA/Other

Name: Anuja Bhowmik Department: MSc Biotechnology Enrolment No 225111005

Contact No. 7876807010 E-mail. 225111005@juitsolan.in

Name of the Supervisor: Dr. Shikha Mittal

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): GENOME-WIDE IDENTIFICATION OF MIRNAS AND LNCRNAS OF VANILLA PLANIFOLIA

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

- Total No. of Pages = 55
- Total No. of Preliminary pages = 7
- Total No. of pages accommodate bibliography/references = 7

Anuja 30/5/24
(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 7 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Shikha Mittal 30/05/2024.
(Signature of Guide/Supervisor)

30/05/2024
Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
30th/05/2024	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String	05%	Word Counts	10,431
			Character Counts	59,678
Report Generated on		Submission ID	Page counts	47
30th/05/2024		2391565276	File Size	1.23M

30/5/24.
Checked by
Name & Signature

30/05/24
Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com